

**Using Cochran's Z Statistic to Test the Kernel-Smoothed IRF Differences
between Focal and Reference Group**

Yinggan Zheng

Mark J. Gierl

Centre for Research in Applied Measurement and Evaluation

University of Alberta

6-110 Education North

Edmonton, AB, Canada T6G2G5

Using Cochran's Z Statistic to Test the Kernel-Smoothed IRF

Differences between Focal and Reference Group

Introduction

Differential item functioning (DIF) is of great interest to researchers and educators given that DIF poses a potential threat to test fairness. DIF occurs when examinees at the same ability level but from different groups have a different probability of answering an item correctly. A variety of parametric and nonparametric procedures have been proposed to detect DIF occurrences and quantify the magnitude of DIF, such as the item response theory (IRT) methods (Lord, 1980; Thissen, Steinberg, & Wainer, 1993), the Mantel-Haenszel statistic (MH; Holland & Thayer, 1988), the Simultaneous Item Bias Test (SIBTEST) (Shealy & Stout, 1993a, 1993b), and Logistic Regression (LR; Swaminathan & Rogers, 1990). A common feature among these DIF procedures is that they assess DIF across the entire ability range of examinees (Global DIF). However, recent studies have shown that DIF is sometimes present only in a specific range of ability or the direction of DIF changes across ability levels (Local DIF). For example, Gierl and Bolt (2001) provided a sample math item in an English-French translation test for which DIF was detected only at localized places along the ability scale, as opposed to DIF that is uniform across ability levels.

Recently, graphical inspection of non-parametrically estimated item response functions (IRFs) has become a useful way of studying DIF, particularly local DIF (e.g., Maydeu-Olivares, Morera, & D'Zurilla, 1999; Scrams and McLeod, 2000; Ramsay, 1991, 2000; Douglas, Stout & DiBello, 1996). IRF defines the probability of a correct response to an item as a function of student's latent ability (θ) measured by the test. If $P_R(\theta)$ and $P_F(\theta)$ denote the IRFs for the

reference and focal group, respectively, then DIF occurs whenever $P_R(\theta) \neq P_F(\theta)$ at some (θ) .

One approach to estimating $P_R(\theta)$ and $P_F(\theta)$ is to use the kernel smoothing procedure where the functional relationship between a student's latent ability (θ) and the probability of answering the item correctly can be estimated (Ramsay, 1991). The benefit of using this nonparametric technique is that the IRFs can take any functional form that is free of the systematic bias potentially suffered by parametric procedures when the presumed parametric model may not reflect reality. TESTGRAF (Ramsay, 2000) is a procedure that can be used to graphically compare the kernel smoothed focal and reference group IRFs to identify DIF items. However, the graphical DIF analysis does not provide a hypothesis testing statistic that can be used objectively to determine the occurrence of DIF. Therefore, the first purpose of this study is to apply the kernel smoothing procedure to a nonparametric DIF statistic, Cochran's Z, to statistically test the significance of focal and reference group IRF differences. Marascuilo and Slaughter (1981) used Cochran's Z after students were coarsely classified into a small number of ability groups (e.g., high, medium, low) based on their test scores. The use of test scores as a matching criterion has the potential to introduce bias into IRF comparisons when groups have different latent ability distributions. More recently, Bolt and Gierl (2006) applied the regression correction procedure, currently used in SIBTEST, to Cochran's Z with an attempt to adjust matching error in the matching criterion. Their finding suggested that the statistical performance of this DIF statistic had been improved to some degree when the regression correction procedure was applied. Hence, the second purpose of the study is to conduct a simulation study to investigate whether the kernel smoothing procedure can further improve the performance of Cochran's Z in terms of Type I error and power in DIF detection.

Method

Kernel-smoothed Cochran's Z

Cochran's Z is a nonparametric statistic designed to test the null hypothesis $P_R(\theta) - P_F(\theta) = 0$ against an alternative hypothesis $P_R(\theta) - P_F(\theta) = C$, where $P_R(\theta)$ denotes reference group IRF, $P_F(\theta)$ denotes focal group IRF, and C is any non-zero constant. Kernel smoothing is a nonparametric regression technique (Ramsay, 1991, 2000).

The calculation of the kernel-smoothed Cochran's Z involves four steps. In step 1, the estimates of latent ability variable for each examinee in each group, $\hat{\theta}_i, i = 1, 2, \dots, n$, are obtained by using the kernel smoothing procedure. In step 2, matching subtest true scores are calculated and the frequencies of matching subtest true scores in each group are determined. In step 3, the kernel-smoothed estimates of the studied items IRF corresponding to each subtest scores are obtained. In step 4, the three statistics are calculated based on the kernel-smoothed studied item IRF and the frequencies of matching subtest true scores. Steps 1 to 3 are adapted from Douglas, Stout, and DiBello's (1996) study, where the kernel smoothing procedure was used to improve the performance of SIBTEST.

Step 1: Estimate the latent ability. Suppose there are m matching subtest items and $2n$ examinees in a test. To simplify, the number of examinees in each group is equal (i.e., n examinees in reference group and n examinees in focal group). Consider item $j, j = 1, 2, \dots, m$ of the matching subtest items in one group, for instance, the reference group. Rank the number-correct scores of the matching subtest items among the n examinees for this group with item j excluded. The rank for each examinee is divided by n to put the score on the $[0, 1]$ scale. The

obtained rank for examinee i is denoted by $\hat{\theta}_i^{(j)}$. For each item j , kernel smoothing estimation is completed using the formula:

$$\hat{P}_j(\theta_q) = \frac{\sum_{i=1}^N K\left(\frac{\hat{\theta}_i^{(j)} - \theta_q}{h}\right) Y_{ij}}{\sum_{i=1}^N K\left(\frac{\hat{\theta}_i^{(j)} - \theta_q}{h}\right)},$$

where Y_{ij} is the score (1 or 0) of the i th examinee's on item j of the matching subtest, $K(u)$ is the kernel smoothing function, θ_q is the target ability point, and h is the bandwidth parameter.

In Douglas, Stout, and DiBello's (1996) study, the quadratic kernel smoothing function (i.e.,

$K(u) = 1 - u^2, |u| \leq 1$) and the bandwidth parameter $h = 0.7N^{-0.2}$ was used. The same kernel

smoothing function and bandwidth parameter are adopted in this study. The estimates of latent

ability θ_i are obtained by summing $\hat{P}_j(\hat{\theta}_i^{(j)})$ for $j = 1, 2, \dots, m$; that is

$$\hat{\theta}_i = \sum_{j=1}^m \hat{P}_j(\hat{\theta}_i^{(j)}).$$

The estimates of latent ability, $\hat{\theta}_i, i = 1, 2, \dots, n$, ranges from 0 to m because each

$\hat{P}_j(\hat{\theta}_i^{(j)}), (i = 1, 2, \dots, n; j = 1, 2, \dots, m)$ ranges from 0 to 1. However, the estimates of latent

ability, $\hat{\theta}_i, i = 1, 2, \dots, n$, are calculated separately for the reference and focal groups. Different

ability distributions for the reference and focal groups will affect the estimation of abilities.

Therefore, the estimates of latent abilities from two groups are pooled and converted to

percentile estimates on the uniform ability scale ranging from 0 to 1. Then, the estimates of

pooled latent ability, denoted by $\hat{\theta}_i', i = 1, 2, \dots, 2n$, are obtained based on the percentile rank of

$\hat{\theta}_i, i = 1, 2, \dots, n$ after reference and focal groups are combined.

Step 2: Calculate the frequency of matching subtest true scores. To calculate the frequency of matching subtest true scores, the $\hat{\theta}_i^t, i = 1, 2, \dots, 2n$ obtained from step 1 are aligned to the matching subtest true score $k, k = 0, 1, 2, \dots, m$. In this study, the centre value is used to categorize the estimates of latent ability. For example, if the estimate of latent ability $\hat{\theta}_i^t$ for an examinee is 1.1, which fall in the interval of $[0.5, 1.5)$, then the matching subtest true score of 1 is assigned to this examinee. In doing so, the estimates of latent ability can be aligned to matching subtest true scores, and the result is denoted by $\hat{\theta}_k, k = 0, 1, 2, \dots, m$. Consequently, the frequency of each matching subtest true score for examinees can be calculated.

Step 3: Estimate kernel-smoothed IRF of studied item. To estimate the kernel-smoothed IRF of studied item, the kernel smoothing procedure is used according to

$$\hat{P}(\theta_q) = \frac{\sum_{i=1}^N K\left(\frac{\hat{\theta}_i^t - \theta_q}{h}\right) Y_i}{\sum_{i=1}^N K\left(\frac{\hat{\theta}_i^t - \theta_q}{h}\right)},$$

where Y_i is the response to the studied item of the i th examinee, $\hat{\theta}_i^t$ is the pooled estimate of latent ability which is obtained from step 1, and θ_q is the target ability point. In this step, target ability point is set as $(m + 1)$ points between 0 and 1 (i.e., $\frac{0}{m}, \frac{1}{m}, \frac{2}{m}, \dots, \frac{m-1}{m}, \frac{m}{m}$). Again,

$K\left(\frac{\hat{\theta}_i^t - \theta_q}{h}\right)$ is the quadratic kernel smoothing function and $h = 0.7N^{-0.2}$ is the bandwidth

parameter.

Then, the estimated probability difference for each studied item under the condition of ability level $\hat{\theta}_k$ between reference and focal groups is calculated using

$$\hat{\Delta}_{\hat{\theta}_k} = \hat{P}_{R\hat{\theta}_k} - \hat{P}_{F\hat{\theta}_k}$$

Step 4: Calculate the three Kernel-smoothed Cochran's Z. The final step is to calculate the kernel-smoothed test statistic. The formula for kernel-smoothed Cochran's Z is:

$$Z = \frac{\hat{\Delta}_0}{SE_{\hat{\Delta}_0}},$$

where $\hat{\Delta}_0 = \frac{\sum_{k=0}^m W_{\hat{\theta}_k} \hat{\Delta}_{\theta_k}}{\sum_{k=0}^m W_{\hat{\theta}_k}}$ is the weighted average difference between IRFs across all valid subtest

scores with $\hat{\Delta}_{\theta_k}$ which is defined in Step 3. The standard error of $\hat{\Delta}_0$ is given by:

$$SE_{\hat{\Delta}_0}^2 = \frac{1}{\left(\sum_{k=0}^m W_{\hat{\theta}_k}\right)^2} \sum_{k=0}^m W_{\hat{\theta}_k} \frac{N_{R\hat{\theta}_k}^2 \hat{p}_{R\hat{\theta}_k} (1 - \hat{p}_{R\hat{\theta}_k}) + N_{F\hat{\theta}_k}^2 \hat{p}_{F\hat{\theta}_k} (1 - \hat{p}_{F\hat{\theta}_k})}{(N_{R\hat{\theta}_k} + N_{F\hat{\theta}_k})^2},$$

where $W_{\hat{\theta}_k} = \frac{N_{R\hat{\theta}_k} N_{F\hat{\theta}_k}}{N_{R\hat{\theta}_k} + N_{F\hat{\theta}_k}}$, and $N_{R\hat{\theta}_k}$ and $N_{F\hat{\theta}_k}$ are defined above. $\hat{p}_{R\hat{\theta}_k}$ and $\hat{p}_{F\hat{\theta}_k}$ are the estimated

probabilities for each studied item for ability level $\hat{\theta}_k$. Under the null hypothesis, Z has an approximate standard normal distribution.

Type I Error Study

In order to investigate the Type I error rates and power of the kernel-smoothed Cochran's Z statistic in DIF detection, simulation studies were conducted. Item response data were generated from a three-parameter logistic (3PL) item response model. Each generated test consisted of 26 items—25 matching subtest items and a studied item. In order to compare the Type I error rates using the kernel smoothing procedure with those using regression correction, the same 25 non-DIF matching items used in Bolt and Gierl's (2006) Type I error study were used in this study. Table 1 contains the summary information for these items. Three factors expected to affect the Type I error rates were considered: sample size, ability distribution

difference, and item parameters of the studied item. Table 2 shows the summary information for these factors. In total, 3 (sample size) X 4 (ability distribution) X 4 (studied item) =48 conditions were manipulated to investigate the Type I error rates for the proposed kernel-smoothed Cochran's Z statistic. For each condition, 100 replications were performed. Two-sided hypothesis tests were used with a significant level of 0.05.

Power Study

The power study was designed to investigate the performance of the kernel-smoothed Cochran's Z in detecting DIF items. Three DIF items were studied, ranging from easy to difficult, with varying amounts of discrimination. The item parameters for each studied item are given in Table 3. Each simulated test consisted of 25 matching items and one studied item. The power study employed the same twenty-five matching subtest items used in the Type I error study. Sample size and ability distribution were also manipulated with the same levels as those in the Type I error study. A total of 3 (sample size) X 4 (ability distribution) X 3 (studied item) =36 conditions were evaluated. For each condition, 100 data sets were generated. Two-side hypothesis tests were used to examine the occurrence of DIF for the studied item using an alpha level of 0.05. For the purpose of comparison, the Type I error and power rates with no correction or regression correction were also calculated under each simulation condition.

Results

The results of Type I errors and power for Cochran's Z under the different simulation conditions are presented in Tables 4 and 5, respectively. Each table compares the Type I error rates for the corresponding test when no correction (NC), regression correction (RC), and kernel smoothing (KS) under each simulation condition. For each condition, the impact of the

manipulated factors in this study—ability distribution, sample size, and item parameter—on the Type I error rate is also summarized.

Type I errors

Given that only 100 replications were conducted for each condition, the standard error was relatively large for this simulated Type I error study (0.02). Therefore, the lower and upper limits of 95 percent confidence interval for the nominal Type I error at 0.05 level were 0.01 and 0.09. Use of this interval would mean that values less than 0.01 would imply a conservative test while values greater than 0.09 would imply a liberal test. This did not seem reasonable.

Therefore, the lower and upper limits were modified as follows: the empirical Type I error rate was considered conservative if it was less than 0.02, reasonable if it was greater than or equal to 0.02 and less than or equal to 0.08, and liberal if it was greater than 0.08.

Table 4 presents the empirical Type I error rates of the Cochran's Z test across different simulation conditions. For the $N_R(0, 1)$ and $N_F(0, 1)$ (i.e., no ability distribution differences between reference and focal groups), the empirical Type I error rates using NC, RC, and KS were all in the inclusive range of 0.02 to 0.08 when the sample size was small (500/500). This result indicates that the three procedures produce reasonable Type I error rates across the studied items when there was no ability difference and the sample size was small. Under moderate sample size (1000/1000), the empirical Type I error was conservative for item 2 (0.01) using NC and for item 1 (0.01) using RC. The remaining empirical Type I error rates were reasonable. Under the large sample size (2500/2500) condition, the empirical Type I error rates using NC were conservative (0.01) for items 1 and 4. Likewise, the empirical Type I error rates using RC were conservative (0.01) for items 1 and 4. Using KS, the Type I errors were liberal for item 3 (0.09) and item 4 (0.12). Therefore, as the sample size increased, the Type I error rates using the

NC and RC procedures tended to be conservative while the Type I error rates using the KS procedure tended to be liberal when there was no ability difference between reference and focal groups. There was no noticeable influence of item parameters on Type I error under this condition.

For the $N_R(0, 1)$ and $N_F(0, 2)$ (i.e., no difference for ability mean, one standard deviation for reference group, and two standard deviation for focal group) condition, the empirical Type I error rates using NC, RC, and KS were all reasonable under the 500/500 sample size condition. For the 1000/1000 sample size, the empirical Type I error rates using NC were liberal for items 1 and 3, while the values for items 2 and 4 were reasonable. The Type I error rates using RC were reasonable across the four studied items. In contrast, the empirical Type I error rates using KS were inflated for three items: the error rates for items 2, 3, 4 were liberal, ranging from 0.09 to 0.13. When sample size was increased to 2500/2500, the Type I error rates using NC were liberal for items 3 and 4 – 0.12 and 0.14, respectively. The Type I error using RC was conservative for item 2 (0.00), while the Type I errors using KS were liberal for items 1 and 4 (0.09 and 0.13, respectively). The remaining Type I error rates were reasonable. Therefore, as the sample size increased, the Type I errors using NC and KS tended to be liberal while the Type I error for RC tended to be conservative when there was no difference between mean ability and the standard deviations differed by one. There was no systematic influence on the Type I error rates due to item parameters.

For the $N_R(0.5, 1)$ and $N_F(-0.5, 1)$ (i.e., no difference for standard deviation, the ability mean was 0.5 for reference group and -0.5 for focal group) condition, when the sample size was small, one of the four empirical Type I error rates using NC was liberal (0.10 for item 3), two of the four Type I error rates using RC were conservative (0.00 for item 1 and 0.01 for item 3), and

the remaining items were reasonable. For the moderate sample size (1000/1000) condition, the empirical Type I error rates using NC were liberal for all four items, ranging from 0.09 to 0.16. On the other hand, all four Type I errors using RC were conservative: 0.00 for items 2 and 4 and 0.01 for items 1 and 3. Using KS, three of the Type I error rates were liberal, ranging from 0.09 to 0.14. Lastly, for the largest sample size (2500/2500), the empirical Type I error rates using NC were inflated for all four items, ranging from 0.36 to 0.62. The Type I error rates using RC were conservative (0.00) for items 2 and 3 and reasonable for items 1 and 4 (0.03 and 0.02, respectively). The Type I error rates using KS were reasonable with one exception, 0.09 for item 1. Therefore, when there was mean ability distribution difference, NC produced liberal Type I error when sample size was moderate and large. The Type I error rates using RC tended to be conservative across small, moderate, and large sample sizes. Lastly, the Type I error rates using KS tended to be reasonable when sample size was small, liberal when sample size was moderate, and reasonable when sample size was large. No systematic influence was noted for item parameters under this condition.

For the N_R (0.5, 1) and N_F (-0.5, 2) (i.e., the ability mean was 0.5 for reference group and -0.5 for focal group, the standard deviation was 1 for reference group and 2 for focal group,) condition, when the sample size was small, one Type I error rate was liberal (0.11 for item 3) using NC, one was conservative (0.00 for item 1) using RC, and one was liberal (0.09 for item 3) using KS. The remaining rates empirical Type I error rates were reasonable. When the sample size was moderate, two Type I errors using NC were liberal (items 1 and 3) and three Type I errors using KS were liberal (items 1, 3, and 4). The remaining Type I error were reasonable. When the sample size was large, the same Type I error pattern observed for the moderate sample size condition was observed for NC and KS. Using RC, however, two Type I errors were

conservative (0.01 for items 2 and 4). Different from the three ability distribution conditions described previously, the results for the N_R (0.5, 1) and N_F (-0.5, 2) condition showed strong impact of item difficulty (b -parameter). The empirical Type I error rate using NC for items 1 and 3 (b -parameters for items 1 and 3 were -0.75) increased noticeably as the sample size increased. For item 1, for example, the Type I error rates increased from 0.05, to 0.11, and then to 0.28 when the sample size increased from small to moderate to large. In contrast, the empirical Type I error rates using NC were reasonable across the different sample size conditions for items 2 and 4 (b -parameters for items 2 and 4 were 0.75). For example, for item 2, the Type I error rates were 0.03, 0.02, and 0.02 when the sample sizes were 500/500, 1000/1000, and 2500/2500, respectively. Using RC and KS, however, the impact of item difficulty was not found.

To summarize, for the Cochran's Z test, the empirical Type I error rates using NC were affected by sample size, ability distribution differences, and item parameter values. When there was no mean ability difference between the reference and focal groups, the empirical Type I error rates using NC were reasonable when the sample size was small and conservative or reasonable when the sample size was large or moderate. However, when there were differences between mean abilities or between the standard deviations of the ability distributions of the reference and focal groups, the Type I error rates using NC increased as sample size increased. When the reference and focal groups differed in both mean ability and standard deviation (N_R (0.5, 1) and N_F (-0.5, 2)), Type I error rates were inflated for the easy items (items 1 and 3), but not for the difficult items (items 2 and 4), as sample size increased. But, the results suggested that the item parameters did not strongly affect the Type I error rates using RC and KS when the sample size was small and there was no mean ability difference. However, when there was a

difference between the mean abilities, the empirical Type I error rates using RC tended to be conservative while the KS Type I error rates tended to be liberal as sample size increased.

Power

In order to interpret the power results, outcomes were categorized as low, moderate, and high according to Cohen's (1962, 1992) criteria. He found that the mean power rate to detect medium effect sizes was 0.48 at the two-tailed 0.05 level of significance (1962). Also, he argued that a procedure could be considered as having excellent power if its power rates were above 0.80 (1992). Therefore, in the present study, power was considered low if the rate was less than 0.48, moderate if the rate was in the closed interval 0.48 and 0.80, and high if the rate exceeded 0.80.

Table 5 displays the results using NC, RC, and KS for the Cochran's Z test. For the N_R (0, 1) and N_F (0, 1) condition, when the sample size was small, the power rates using NC, RC, and KS varied across the three studied items. For item 1, the power rates using NC and RC were low, at 0.04 for both. The power rate using KS was higher, but still low, at 0.29. For item 2, the power rates were 1.00 when NC, RC, and KS were used, indicating that all three procedures correctly identified the occurrence of DIF across the 100 generated data sets. For item 3, the power rates using NC and RC were low, at 0.34 and 0.33, respectively, while the rate using KS was high at 0.84. When the sample size was moderate, the power rates using NC and RC were low for item 1 (0.02 for NC and RC), while the power rate using KS was low, at 0.43. For item 2, the power rates were again high at 1.00 across three procedures. For item 3, the rates were both moderate (0.64) for NC and RC, while the power rate using KS was high (0.95). When sample size was large, the power rates using NC, RC, and KS were high across the three studied items, ranging from 0.82 to 1.00, with two exceptions (0.01 for NC and 0.02 for RC for item 1).

For the $N_R(0, 1)$ and $N_F(0, 2)$ condition, when the sample size was small, the power rates using NC, RC, and KS produced a similar outcome with the $N_R(0, 1)$ and $N_F(0, 1)$ condition. That is, for item 1, the power rates were low using NC and RC (0.04 and 0.09, respectively), while the power using KS was higher, but still low, at 0.33. For item 2, the power rates were 1.00 when NC, RC, and KS were used. For item 3, the power rates using NC and RC were low, at 0.11 and 0.13, respectively, while the rate using KS was moderate at 0.75. When the sample size was moderate, the power rates using NC and RC were low for item 1 (0.01 for NC and 0.00 for RC), while the power rate using KS was moderate, at 0.52. For item 2, the power rates were again high at 1.00 across three procedures. For item 3, the rates were both low for NC (0.31) and RC (0.30), while the power rate using KS was high (0.98). When sample size was large, the power rates using NC, RC, and KS were high across all studied items, ranging from 0.88 to 1.00, with two exceptions (0.03 for NC and 0.04 for RC for item 1).

For the $N_R(0.5, 1)$ and $N_F(-0.5, 1)$ condition, when sample size was small, the power rates using NC, RC, and KS for item 1 were low (0.06, 0.04 and 0.29, respectively). For item 2, the power rates using NC, RC, and KS were all high and close to 1.00, ranging from 0.98 to 0.99. For item 3, the power rates were low for NC (0.33), low for RC (0.10), and high for KS (0.85). When the sample size was moderate, the power rates using NC and RC for item 1 remained low, 0.12 and 0.02, respectively, and the rate using KS was moderate (0.55). For item 2, the power rates using NC, RC, and KS were high (1.00, 1.00, and 0.98). For item 3, the power rates were moderate for NC (0.74), low for RC (0.24), and high for KS (0.94). When sample size was large, the power rates were high across procedures and items, except for NC (0.51) and RC (0.00) for item 1.

For the $N_R (0.5, 1)$ and $N_F (-0.5, 2)$ condition, when the sample size was small, the power rates using NC and RC for item 1 were low, 0.09 and 0.06 respectively. In contrast, the power rate using KS for item 1 was moderate (0.78). For item 2, the power rates using NC and RC were both moderate (0.73 and 0.79). However, the power rate using KS was still high for item 2 (0.97). For item 3, the power rates using NC and RC were low (0.30 and 0.25), while the power rates using KS was high (0.99). When the sample size was moderate, the power rates using NC and RC for item 1 were low (0.16 and 0.04), while the power using KS was high (0.98). For item 2, the power rates were high for all three procedures, 0.99, 1.00, and 1.00, respectively. The power rates using NC and RC were low (0.42 and 0.31, respectively) for item 3, while the rate using KS was high (1.00). When the sample size was large, the power rates were high across the procedures and items with two exceptions: the power rates for item 1 using NC (0.45) and using RC (0.19).

To summarize, for Cochran's Z , NC and RC yielded low power for item 1 (ranging from 0.00 to 0.45) consistently with one exception (0.51 for NC at large sample size when $N_R (0.5, 1)$ and $N_F (-0.5, 2)$), but moderate to high power for item 2 (ranging from 0.73 to 1.00) across different sample sizes and ability distributions. For item 3, NC and RC yielded comparable power rates when there was no ability mean difference. However, NC produced better power rates than RC when there was ability mean difference between groups. KS produced the highest power rates across different conditions for Cochran's Z . These results showed the strength in using KS procedure for Cochran's Z to detect DIF.

Conclusions and Future Research

Recently, practitioners and researchers have become interested in the graphical comparison of non-parametrically estimated IRFs for DIF analysis. This interest occurs because the graphical comparison of non-parametrically estimated reference and focal group IRFs has the potential to detect both non-uniform DIF and local DIF without making strict assumptions about the student ability distribution and the functional forms of IRFs. However, in order to objectively determine the occurrence of IRF differences, DIF hypothesis testing statistics are needed. Ramsay (1991) introduced kernel smoothing, a general technique for nonparametric estimation, to measurement practice. However, the procedure Ramsay proposed does not provide a hypothesis testing statistic that can be used objectively to determine the occurrence of DIF.

The present study combined the kernel smoothing procedure and the nonparametric DIF statistic Cochran's Z to statistically test the difference between the kernel-smoothed IRF for reference group and the IRF for focal group. To calculate the kernel-smoothed statistics, examinees' latent abilities were estimated using the kernel smoothing technique and these estimates served as the matching criterion for DIF detection. Using the latent ability estimates rather than subtest observed scores as the matching criterion can be considered as a latent-variable-matched DIF procedure (Douglas, Stout & DiBello, 1996). This procedure avoids the potential problems introduced by bias in DIF detection when groups have different latent ability distributions. After latent ability estimation, the true score for each examinee and the frequency of the true scores were calculated. The kernel smoothing technique was also applied in the calculation of the probabilities of answering the studied item correctly for the examinees in reference and focal groups at certain ability level.

Simulation studies were conducted to investigate the Type I error and power of the proposed kernel-smoothed (KS) Cochran's Z . For the Type I error study, three factors expected

to affect the results were manipulated: sample size, ability distribution difference, and item parameters of the studied item. There were three levels in the factor of sample size, four levels in the factor of ability distribution, and four levels in the factor of item parameters. In total, 3 (sample size) \times 4 (ability distribution) \times 4 (studied item) = 48 conditions were. One hundred replications were performed for each test. For power study, the same three factors were manipulated. There were three levels in the factor of sample size, four levels in the factor of ability distribution, and three levels in the factor of item parameters. In total, 3 (sample size) \times 4 (ability distribution) \times 3 (studied item) = 36 conditions were generated to investigate the power performance for the three kernel-smoothed statistics. Two-sided hypothesis tests were used with a significance level of 0.05 for both Type I error and power studies. The Type I error and power rates of Kernel Smoothed (KS) statistic were compared to those with No Correction (NC) and Regression Correction (RC) to evaluate the performance of the new statistic introduced in this study. The results of this study suggest the Type I error and power performance of Cochran's Z improved when kernel smoothing was applied. In particular, the power rates of KS were increased significantly compared to the rates from NC and RC. The study provided empirical evidence to support that the kernel smoothing procedure has potential to improve the performance of the currently existing DIF statistics.

The results of the present study have practical implications. The performance of the Cochran's Z statistics improved significantly when kernel smoothing was applied compared to the power and Type I error rates from NC and RC. This result is of particular value because it provides researchers and practitioners with a new method for statistically confirming their findings from non-parametrically graphical DIF analysis. In turn, this result also suggested that the kernel smoothing procedure has the potential to improve the performance of nonparametric

DIF statistics because it can reduce the local error variance and instability often associated with nonparametric IRFs.

The most important limitation in this study is that the three kernel-smoothed statistics were not applied to real data situation. Only simulation studies were conducted. Although using 3PL or 2PL item response model to generate simulated data is a common method in the literature, it is not clear whether it is appropriate to use parametric methods to generate data and then analyze the generated data using non-parametric estimation procedure. Therefore, applying these procedures to real data situation is important and it will be the priority in future research.

References

- Bolt, D.M. & Gierl, M.J.(2006). Testing features of graphical DIF: application of a regression correction to three nonparametric statistical tests. *Journal of Educational Measurement*, 43, 313-333
- Cohen, J. (1992). A power primer. *Psychological bulletin*, 112, 155-159.
- Douglas, J.A., Stout, W., & DiBello, L.V. (1996). A kernel smoothed version of SIBTEST with applications to local DIF inference and function estimation. *Journal of Educational and Behavioral Statistics*, 21, 333-363.
- Gierl, M.J. & Bolt, D.M. (2001). Illustrating the use of nonparametric regression to assess differential item and bundle functioning across multiple groups. *International Journal of Testing*, 3&4, 249-270.
- Holland, P. W., & Thayer, D. T. (1988). Differential item functioning and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 129-145). Hillsdale, NJ: Erlbaum.
- Marascuilo, L.A. & Slaughter, R.E. (1981). Statistical procedures for identifying possible sources of item bias based on \hat{A}^2 statistics. *Journal of Educational Measurement*, 18, 229-248.
- Maydeu-Olivares, A., Morera, O. F., & D'Zurilla, T. J. (1999). Using graphical methods in assessing measurement invariance in inventory data. *Multivariate Behavioral Research*, 34, 397-420.
- Ramsay, J.O. (1991). Kernel smoothing approaches to nonparametric item characteristic curve estimation. *Psychometrika*, 56, 611-630.
- Ramsay, J.O. (2000). TESTGRAF manual. McGill University: Montreal, Quebec, Canada.

- Scrams, D.J & McLeod, L.D. (2000). An expected response function approach to graphical differential item functioning. *Journal of Educational Measurement*, 37, 263-280.
- Shealy, R. T., & Stout, W. F. (1993a). An item response theory model for test bias and differential test functioning. In P. Holland & H. Wainer (Eds.) *Differential Item Functioning* (pp. 197-239). Hillsdale, NJ.
- Shealy, R. T., & Stout, W. F. (1993b). A model-based standardization approach that separates true-bias/DIF from group ability differences and detects test bias/DIF as well as item bias/DIF. *Psychometrika*, 54, 159-194.
- Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement*, 27, 361-370.
- Thissen, D., Steinberg, L. & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P.W. Holland, & H. Wainer (Eds.), *Differential item functioning*. (pp. 67-113). Hillsdale, NJ: Lawrence Erlbaum.

Table 1. Item Parameters for Matching Items in Type I Error and Power Studies

Item	Item Parameters		
	a	b	c
1	1.53	1.21	0.20
2	0.89	-0.65	0.20
3	1.46	-0.09	0.20
4	0.73	0.65	0.20
5	1.39	1.99	0.20
6	0.49	0.22	0.20
7	0.52	-0.67	0.20
8	0.97	-0.38	0.20
9	1.10	1.78	0.20
10	0.81	-0.37	0.20
11	1.09	-0.75	0.20
12	0.64	-0.25	0.20
13	1.39	-1.07	0.20
14	1.23	2.78	0.20
15	0.42	0.72	0.20
16	1.46	-1.59	0.20
17	1.45	-2.00	0.20
18	1.18	-1.00	0.20
19	0.41	-0.49	0.20
20	1.00	-0.68	0.20
21	1.07	1.23	0.20
22	0.63	0.82	0.20
23	1.58	-1.66	0.20
24	1.00	-0.56	0.20
25	0.87	1.73	0.20

Table 2. Information for Manipulated Factors

Sample Size		Ability Distribution		Item Parameters		
Ref.	Foc.	Ref.	Foc.	a	b	c
500	500	$N(0, 1)$	$N(0, 1)$	1.00	-0.75	0.20
1000	1000	$N(0, 1)$	$N(0, 2)$	1.00	0.75	0.20
2500	2500	$N(0.5, 1)$	$N(-0.5, 1)$	1.50	-0.75	0.20
		$N(0.5, 1)$	$N(-0.5, 2)$	1.50	0.75	0.20

Table 3. Item Parameters for Studied Items in Power Study.

Studied Item	Group	Item parameters		
		a	b	c
1	Ref.	1.50	0.00	0.2
	Foc.	0.50	0.00	0.2
2	Ref.	2.00	1.00	0.2
	Foc.	0.40	0.00	0.2
3	Ref.	1.80	0.00	0.2
	Foc.	0.40	0.50	0.2

Table 4. Type I Error for Cochran's Z with No Correction, Regression Correction, and Kernel Smoothing

(Proportion of rejections out of 100 replications for each condition, $\alpha=0.05$)

Ability Distribution	Studied Item	Sample Size								
		500/500			1000/1000			2500/2500		
		NC	RC	KS	NC	RC	KS	NC	RC	KS
$N_R(0,1), N_F(0,1)$	Item 1	0.03	0.03	0.03	0.02	0.01	0.07	0.01	0.01	0.05
	Item 2	0.03	0.02	0.07	0.01	0.02	0.07	0.05	0.05	0.08
	Item 3	0.07	0.06	0.07	0.03	0.03	0.06	0.02	0.03	0.09
	Item 4	0.02	0.02	0.05	0.03	0.03	0.07	0.01	0.01	0.12
$N_R(0,1), N_F(0,2)$	Item 1	0.06	0.04	0.02	0.10	0.06	0.06	0.07	0.05	0.09
	Item 2	0.04	0.02	0.08	0.03	0.03	0.10	0.05	0.00	0.06
	Item 3	0.05	0.04	0.02	0.17	0.02	0.09	0.12	0.02	0.07
	Item 4	0.06	0.06	0.02	0.05	0.02	0.13	0.14	0.02	0.13
$N_R(0.5,1), N_F(-0.5,1)$	Item 1	0.04	0.00	0.07	0.09	0.01	0.09	0.36	0.03	0.09
	Item 2	0.06	0.02	0.05	0.14	0.00	0.12	0.46	0.00	0.04
	Item 3	0.10	0.01	0.02	0.16	0.01	0.14	0.43	0.00	0.07
	Item 4	0.07	0.02	0.04	0.16	0.00	0.03	0.62	0.02	0.07
$N_R(0.5,1), N_F(-0.5,2)$	Item 1	0.05	0.00	0.06	0.11	0.04	0.13	0.28	0.03	0.10
	Item 2	0.03	0.03	0.06	0.02	0.03	0.07	0.02	0.01	0.05
	Item 3	0.11	0.04	0.09	0.17	0.04	0.10	0.38	0.02	0.15
	Item 4	0.03	0.03	0.07	0.04	0.02	0.14	0.03	0.01	0.10

Note: NC for No Correction, RC for Regression Correction, KS for Kernel Smoothing;

Table 5. Power Rates for Cochran's Z with No Correction, Regression Correction, and Kernel Smoothing

(Proportion of rejections out of 100 replications for each condition, $\alpha=0.05$)

Ability Distribution	Studied Item	Sample Size								
		500/500			1000/1000			2500/2500		
		NC	RC	KS	NC	RC	KS	NC	RC	KS
$N_R(0,1), N_F(0,1)$	item 1	0.04	0.04	0.29	0.02	0.02	0.43	0.01	0.02	0.82
	item 2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	item 3	0.34	0.33	0.84	0.64	0.64	0.95	0.97	0.97	1.00
$N_R(0,1), N_F(0,2)$	item 1	0.04	0.09	0.33	0.01	0.00	0.52	0.03	0.04	0.88
	item 2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	item 3	0.11	0.13	0.75	0.31	0.30	0.98	0.88	0.93	0.98
$N_R(0.5,1), N_F(-0.5,1)$	item 1	0.06	0.04	0.29	0.12	0.02	0.55	0.51	0.00	0.86
	item 2	0.98	0.99	0.98	1.00	1.00	0.98	1.00	1.00	1.00
	item 3	0.33	0.10	0.85	0.74	0.24	0.94	1.00	0.90	0.99
$N_R(0.5,1), N_F(-0.5,2)$	item 1	0.09	0.06	0.78	0.16	0.04	0.98	0.45	0.19	1.00
	item 2	0.73	0.79	0.97	0.99	1.00	1.00	1.00	1.00	1.00
	item 3	0.30	0.25	0.99	0.42	0.31	1.00	1.00	0.96	1.00

Note: NC for No Correction, RC for Regression Correction, KS for Kernel Smoothing;