University of Alberta

Using Three Nonparametric Statistics to Test the Kernel-smoothed IRF Differences between Reference and Focal Groups

by

Yinggan Zheng

A thesis submitted to the Faculty of Graduate Studies and Research
in partial fulfillment of the requirements for the degree of

Master of Education
in
Measurement, Evaluation, and Cognition

Department of Educational Psychology

Edmonton, Alberta
Spring 2008

Your file Votre référence
ISBN: 978-0-494-45759-7
Our file Notre référence
ISBN: 978-0-494-45759-7

University of Alberta

Library Release Form

Name of Author: Yinggan Zheng

Title of Thesis: Using Three Nonparametric Statistics to Test the Kernel-smoothed IRF Differences between Reference and Focal Groups

Degree: Master of Education

Year this Degree Granted: 2008

Signature

Abstract

The present study combined the kernel smoothing procedure and three nonparametric DIF statistics—Cochran's *Z*, Fisher's χ^2 , and Goodman's *U*— to statistically test the difference between the kernel-smoothed IRF for reference group and the IRF for focal group. Simulation studies were conducted to investigate the Type I error and power of the proposed kernel-smoothed (KS) statistics. For the purpose of comparison, the Type I error and power rates with no correction (NC) and with regression correction (RC) were also include in the simulation. The results suggest that the kernel-smoothed Cochran's Z can be the statistic to test the difference between the kernel-smoothed IRFs when the sample size was small. When the sample size was moderate and large, the kernel-smoothed Cochran's Z and Fisher's χ^2 could be the candidates. However, we have to be aware of the fact that the Type I errors for both of them tend to be liberal.

Table of Contents

List of Tables

Chapter I: Introduction

Overview

Differential item functioning (DIF) is of great interest to researchers and educators given that DIF poses a potential threat to test fairness. As stated in the *Standards for Educational and Psychological Testing* (*Standards*) (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999),

> The test is not biased or offensive with regard to race, sex, native language, ethnic origin, geographic region, or other factors. Test developers are expected to exhibit sensitivity to the demographic characteristics of test takers. Steps can be taken during test development, validation, standardization, and documentation to minimize the influence of cultural dependency, using statistics to identify differential item difficulty, and examining the comparative accuracy of predictions for different groups. (http://www.unl.edu/buros/bimm/html/lesson03.html)

Canadians have used large-scale standardized testing at many levels within the education system since the 1980s (Rogers & Klinger, 2007). Whether at the provincial, national, or international level, DIF in standardized testing is a constant concern.

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 statistical procedures have been proposed to detect the occurrence of DIF and to quantify the magnitude and the direction 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).

The IRT methods (Lord, 1980; Thissen, Steinberg, & Wainer, 1993) can be considered as one of the pioneer procedures to detect DIF in the testing process. These methods, based on item response theory (IRT), provide a useful theoretical framework for DIF analysis because they explicitly use between-group differences in the item parameters. The general framework of the IRT methods involves estimating item parameters separately for the reference and focal groups. After placing the groups on the same scale, differences between the item parameters for the two groups can then be compared. When the parameters are identical for the two groups, the item does not display DIF. Otherwise, the item displays DIF.

Mantel and Haenszel (1959) introduced a procedure for the study of item evaluation. This procedure then was adapted by Holland and Thayer (1988) for use in assessing differential item functioning. It is based on the analysis of contingency tables. This method differs from the IRT approaches in that examinees are typically matched on an observed variable (such as total test score), and then proportions of examinees in the focal and reference groups who answer the studied item correct or incorrect are compared. The method has been shown to be effective with reasonably small examinee samples (e.g., 200 examinees per group) (Holland & Thayer, 1988).

The SIBTEST procedure (Roussos & Stout, 1996; Shealy & Stout, 1993a) is a relatively recent addition to the list of DIF statistics. Based on the ratio of the weighted difference in proportion-correct scores (for reference and focal group

members) to its standard error, it includes a test of significance. It also includes several conceptual innovations. The first of these is that the matching criterion is a latent score rather than the observed score. Estimation of this matching score includes a linear regression correction that has been shown to be useful in controlling Type I errors (Shealy & Stout, 1993b). Additionally, SIBTEST allows for an evaluation of DIF amplification or cancellation across items within a testlet or bundle (Douglas, Roussos, & Stout, 1996).

Logistic regression (Swaminathan & Rogers, 1990) may be conceptualized as a link between the contingency table methods (e.g., Mantel-Haenszel, SIBTEST) and the IRT methods. The contingency table methods form groups based on discrete score categories. By contrast, logistic regression treats the total score as a continuous variable and predicts an examinee's performance on the studied item based on the examinee's total score and group membership. The most notable feature of the logistic regression procedure is that it is designed to identify both uniform (unidirectional) DIF, which occurs when an item favors one group over another throughout the ability continuum, and non-uniform (crossing) DIF, which occurs when there is an ability-by-group membership interaction. Simulation studies in which the MH and LR have been compared have been conducted by Rogers and Swaminathan (1993). The study demonstrated that the Logistic regression procedure is as powerful as the MH procedure in detecting uniform DIF, and more powerful than the MH procedure in detecting non-uniform DIF.

A common feature among these DIF procedures is that they assess DIF across the entire ability range of examinees (called 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.

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

Purpose of the Study

Therefore, the first purpose of this study is to apply the kernel smoothing procedure to three nonparametric DIF statistics—Cochran's *Z*, Fisher's χ^2, and Goodman's *U*—to statistically test the significance of focal and reference group IRF differences. The three statistics were used to test different performances among ethnic groups by Marascuilo and Slaughter (1981). To use these statistics, examinees were classified into a small number of ability groups (e.g., high, medium, low) based on their internal test scores. The use of internal 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 the three statistics in an attempt to adjust the potential bias in the matching criterion. Their findings suggested that the statistical performance of these DIF statistics was improved to some degree when the regression correction procedure was applied. Hence, the second purpose of the present study is to conduct a simulation study to investigate whether the kernel smoothing procedure can further improve the performance of the three modified statistics in terms of Type I error and power in DIF detection.

Organization of the Thesis

The thesis is organized in five chapters. Chapter II provides an overview of the kernel smoothed IRF estimation technique and reviews the three nonparametric

DIF statistics considered in this research. Chapter III presents the detailed steps for applying the kernel smoothing technique with the three statistics to test the significance of IRF differences. The procedure and manipulated factors for the simulation studies designed to assess Type I error and power are also described in this Chapter. Chapter IV presents the results of the simulation studies. Chapter V discusses the major findings and the implications of this study for DIF analysis. Limitations of this study and directions for future research are also presented.

Chapter II: Literature Review

Kernel-Smoothed IRF Estimation

In the Item Response Theory framework, item response functions (IRFs) that model the functional relationship between an examinee's latent ability (θ) and the probability of answering the item correctly are usually specified in the form of parametric models such as the Rasch, 2PL, and 3PL models. However, parametric models cannot always estimate the item characteristics curve in an accurate and efficient way. For example, the most widely used model, the 3PL model, is problematic when an item is extremely easy. In this situation, there are virtually no data available for estimating the guessing parameter *c*. As a result, large changes in the parameter *c* are compensated for by the corresponding changes in the discrimination parameter *a*, which causes poor estimation of parameter *a* (Ramsay, 2000). Even when the item is of moderate difficulty, the covariances between the 3PL parameter estimates are high, and large amounts of data are required to estimate the item parameters precisely (Thissen & Wainer, 1982). This outcome has led to research for estimating IRFs without the restriction of a parameterized functional form (Altman, 1992; Douglas, 1996, 1997; Douglas & Cohen, 2001; Ramsay, 1991, 2000).

Kernel smoothing is a nonparametric regression technique that has been introduced in measurement practice (Ramsay, 1991). Nonparametric regression is a set of techniques for estimating a regression curve without making strong

assumptions about the shape of the true regression function. Ramsay (1991) discussed the use of kernel smoothing to estimate IRFs. The benefit of using kernel smoothing is that the IRFs can take any functional form that is free of the systematic bias potentially suffered by parametric procedures when the presumed model does not fit the data perfectly. Using the kernel smoothing technique, Ramsay (2000) developed a program for the graphical analysis of multiple choice and questionnaire data: TESTGRAF. TESTGRAF can graphically present the kernel smoothed IRFs and help the user visually compare the focal and reference group IRFs in order to identify DIF items. For more detailed information about TESTGRAF, the reader is referred to Ramsay (2000).

Kernel smoothing estimation is based on local averaging. Suppose one has a set of independent variable values $x_i, i = 1,...,n$, and a corresponding set of dependent variable values $y_i, i = 1,...,n$. The objective is to estimate a smooth curve defined by function P with value $P(x)$. For example, one might want to compute the value $P(x_q)$ at an independent variable value x_q, which is called the targeted point. The targeted point may or may not coincide with any of the data values $x_i, i = 1,...,n$. An intuitive way of estimating $P(x_q)$ is to compute the average of those values $y_i, i = 1,...,n_k$ associated with values $x_i, i = 1,...,n_k$ that are close to the targeted point x_q. This technique is called local averaging. Usually, one could let $P(x_q)$ be the arithmetic mean of the y_is corresponding to the k x_is closest to x_q, or, more commonly, let $P(x_q)$ be the arithmetic mean of the y_is corresponding to the x_is

which are not more than h units from x_q. Then the kernel smoothing regression function can be written as:

$$\hat{P}(x_q) = \frac{\sum_{i=1}^{N} K(\frac{x_i - x_q}{h}) y_i}{\sum_{i=1}^{N} K(\frac{x_i - x_q}{h})},$$

where $K(\frac{x_i - x_q}{h})$ is the kernel smoothing function and $\frac{x_i - x_q}{h}$ is the argument of the function $K(u)$ and h is called the smoothing parameter or the bandwidth parameter. The value of h controls the size of the difference between data value $x_i, i = 1,...,n$ and target point x_q (see Härdle (1990) and Ramsay (2000) for more details). In TESTGRAF, h is set as a function of sample size N: $h = 1.1N^{-0.2}$.

Three commonly used kernel smoothing functions can be applied to define local averaging: uniform, quadratic, and Gaussian (Ramsay, 1991). The three kernel smoothing function are presented as follows:

Uniform: $K(u) = 0.5, |u| \le 1$, and 0 otherwise,

Quadratic: $K(u) = 1 - u^2, |u| \le 1$, and 0 otherwise,

Gaussian: $K(u) = \exp(\frac{-u^2}{2})$.

No matter which of the three functional forms is taken, $K(u)$ is always zero or positive for all values of the argument u, $K(0)$ is always the maximum value taken by $K(u)$, and $K(u)$ always goes to zero as u deviates more and more in either direction from 0. In TESTGRAF, the Gaussian kernel smoothing function was adopted.

When kernel smoothing is applied to IRF estimation, the independent variable is $\theta_i, i = 1,...,n$, the latent ability variable. The dependent variable is the probabilities that examinees answer an item correctly, $p_i, i = 1,...,n$. Therefore, the kernel smoothing regression function applied in IRF estimation can be written as:

$$\hat{P}(\theta_q) = \frac{\sum_{i=1}^{N} K(\frac{\theta_i - \theta_q}{h}) p_i}{\sum_{i=1}^{N} K(\frac{\theta_i - \theta_q}{h})},$$

where $K(\frac{\theta_i - \theta_q}{h})$ is the kernel smoothing function, θ_i is ability estimate for examinee i, θ_q is the target ability point, p_i is the probability of answering the item correctly by examinee i, and h is the bandwidth parameter.

Although it is impossible to know the true value of the latent ability for each examinee, the kernel smoothing procedure can be operationalized by using the estimated latent ability for each examinee. In this sense, the kernel smoothing regression equation above is written as follows:

$$\hat{P}(\theta_q) = \frac{\sum_{i=1}^{N} K(\frac{\hat{\theta}_i - \theta_q}{h}) p_i}{\sum_{i=1}^{N} K(\frac{\hat{\theta}_i - \theta_q}{h})}.$$

The details about the procedure used to estimate the latent ability variable, $\hat{\theta}_i$, are presented in the chapter III of method.

Three Regression-corrected Nonparametric Statistics

Marascuilo and Slaughter (1981) proposed six statistical procedures for identification of potentially biased test items. Three of these —Cochran's *Z*, Fisher's χ^2, and Goodman's *U* —were adapted by Bolt and Gierl (2006). They applied a

regression correction procedure, which is currently implemented in SIBTEST, to reduce measurement error in the matching criterion. These three statistics with the regression correction are reviewed in this section.

Cochran's Z test.

Cochran's Z test evaluates the null hypothesis

$$H_0 : \Delta_1 = \Delta_2 = \cdots = \Delta_k = 0$$

against

$$H_1 : \Delta_1 = \Delta_2 = \cdots = \Delta_k = \Delta_0,$$

where $\Delta_k = p^*_{Rk} - p^*_{Fk}$, in which p^*_{Rk} and p^*_{Fk} denote the probabilities of success for examinees in reference and focal groups, respectively, for the valid subtest score level *k* after using regression correction and Δ_0 is a specified value. If the null hypothesis is rejected, one would conclude that a constant difference in the probabilities of success exists across all of the ability levels. The estimate of Δ_0, $\hat{\Delta}_0$, can be computed as:

$$\hat{\Delta}_0 = \frac{\sum_k^N W_k \hat{\Delta}_k}{\sum_k^N W_k},$$

where

$$W_k = \frac{N_{Rk} N_{Fk}}{N_{Rk} + N_{Fk}},$$

is the weight associated with each valid subtest score. The final test statistic is given by

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

where

$$SE^2_{\hat{\Delta}_0} = \frac{1}{(\sum_k^N W_k)^2} \sum_k^N W_k \frac{N^2_{Rk} p^*_{Rk}(1-p^*_{Rk}) + N^2_{Fk} p^*_{Fk}(1-p^*_{Fk})}{(N_{Rk}+N_{Fk})^2}.$$

Under the null hypothesis, Z has a approximate standard normal distribution. Therefore, the value of Z can be tested in terms of whether its absolute value is greater than 1.96 at the 0.05 significance level.

Fisher's χ^2 test.

Fisher's χ^2 test evaluates the statistical dependence between group membership (reference/focal) and item response (correct/incorrect) conditioned on valid subtest score level k (Bolt & Gierl, 2006). Fisher's χ^2 provides an omnibus test of:

$$H_0 : \Delta_1 = \Delta_2 = \cdots = \Delta_k = 0$$

against

$$H_1 : \text{at least one } \Delta_k \neq 0 .$$

where $\Delta_k = p^*_{Rk} - p^*_{Fk}$, in which p^*_{Rk} and p^*_{Fk} denote the probabilities of success for examinees in reference and focal groups, respectively, for the valid subtest score level k after using regression correction.

The formula for the test statistic with regression correction applied is:

$$\chi^2_k = \frac{(N_{Rk}+N_{Fk}-1)(n^*_{Rk0}n^*_{Fk1} - n^*_{Rk1}n^*_{Fk0})^2}{(n^*_{Rk0}+n^*_{Rk1})(n^*_{Rk0}+n^*_{Fk0})(n^*_{Rk1}+n^*_{Fk1})(n^*_{Fk0}+n^*_{Fk1})},$$

where N_{Rk} and N_{Fk} denote the total number of examinees having valid subtest score k in the reference and focal group, respectively; n^*_{Rk1} and n^*_{Fk1} denote the regression-corrected number of examinees (frequencies) in the reference and focal groups who obtained valid subtest score k and answered the item correct; and n^*_{Rk0} and n^*_{Fk0} denote the regression-corrected number of examinees in each group who answered

the item incorrectly. The regression-corrected frequencies can be calculated using the total number of examinees multiplied by the adjusted conditional proportion correct or incorrect scores obtained from the SIBTEST extended output. The test statistic is approximately distributed as a χ^2 with one degree of freedom when $N_{Rk} + N_{Fk} \geq 20$ and each of n^*_{Rk1}, n^*_{Fk1}, n^*_{Rk0}, and n^*_{Fk0} is at least 3 (Kanji, 1993).

The χ^2_k statistic at each matching score level can also be summed across a range of matching score levels to produce an omnibus test of DIF:

$$\chi^2_T = \sum_{k \in K} \chi^2_k$$

where K consists of all matching score levels of interest that satisfy the necessary cell size criteria for a suitable χ^2 approximation. The test statistic χ^2_T is distributed as a χ^2 with degrees of freedom equal to the number of matching score levels included in K.

Goodman's U test.

Goodman's U test evaluates whether the amount of DIF in an item varies across ability levels. In statistical terms, Goodman's U is used to test

$$H_0 : \Delta_1 = \Delta_2 = \cdots = \Delta_k = \Delta_0,$$

against

$$H_1 : H_0 \text{ is false.}$$

The formula for computing the test statistics is:

$$U = \sum_k^N W_k (\hat{\Delta}_k - \hat{\Delta}_0)^2 ,$$

where $\hat{\Delta}_0 = \frac{\sum_k^N W_k \hat{\Delta}_k}{\sum_k^N W_k}$ is the average difference between IRFs across all valid subtest scores with Δ_k defined the same way as for the Cochran's Z test, $W_k = \frac{1}{SE^2_{\hat{\Delta}_k}}$ is the weight applied to the displacement quantity $(\hat{\Delta}_k - \hat{\Delta}_0)$, and $SE^2_{\hat{\Delta}_k} = \frac{p^*_{Rk}(1-p^*_{Rk})}{N_{Rk}} + \frac{p^*_{Fk}(1-p^*_{Fk})}{N_{Fk}}$ is the error variance.

The same requirements as for the Fisher's χ^2 test needs to be met for Goodman's U test (i.e., $N_{Rk} + N_{Fk} \geq 20$, and each of n^*_{Rk1}, n^*_{Fk1}, n^*_{Rk0}, and n^*_{Fk0} is at least 3). The test statistic U is approximately distributed as a χ^2 distribution under H_0 with degrees of freedom equal to the number of valid subtest score categories used in the computation of U.

Bolt and Gierl (2006) conducted both a simulation study and a real data study to assess each of the three statistics. Two factors, sample size and ability distribution difference, were manipulated in the simulation study. Simulation study for Type I error was conducted for both not corrected and corrected statistics, but the power study was only conducted for the three corrected statistics. The results from the Type I error study demonstrated the effectiveness of the regression correction procedure in improving the performance of the three DIF statistics in some conditions. However, the Type I error rates were still high after regression correction, ranging from 0.12 to 0.20, when there was an ability distribution difference and sample size was large. Moreover, the Type I error rates using regression correction for highly discriminating and easy items were high across the three statistics. For example, under one simulation condition, three items had Type I error rates of .48, .78 and .43,

respectively, for the regression-corrected Fisher's χ^2, .54, .92, and .55 for the regression-corrected Cochran's *Z*, and .26, .62, and .26 for the regression-corrected Goodman's *U*. In the real data study, data from six high school certification examinations were used to study global and local DIF across English- and Chinese-speaking student groups. Based on the degree of agreement among the three statistic test, the items may be candidates for local DIF analysis were selected. By testing for DIF at any location along the ability scale for candidate items, Fisher's χ^2 and Goodman's U, but not Cochran's *Z*, appeared to be useful in identifying items that may display DIF at some ability levels but not others.

Chapter III: Method

Three Kernel-Smoothed Statistics

The calculation of the kernel-smoothed statistics for Fisher's χ^2, Cochran's Z, and Goodman's U involved four steps. In step 1, the estimates of latent ability variable for each examinee in each group, $\hat{\theta}_i, i = 1,2,...,n$, were obtained by using the kernel smoothing procedure. In step 2, matching subtest true scores were calculated and the frequencies of matching subtest true scores in each group were determined. In step 3, the kernel-smoothed estimates of the studied items IRF corresponding to each subtest scores were obtained. In step 4, the three statistics were calculated based on the kernel-smoothed studied item IRF and the frequencies of matching subtest true scores. Steps 1 to 3 were 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 $j, j = 1,2,...,m$ 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 was completed using the formula:

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

where Y_{ij} is the score (1 or 0) of the ith 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 were adopted in this study. The estimates of latent ability θ_i were obtained by summing $\hat{P}_j(\hat{\theta}_i^{(j)})$ for $j = 1,2,...,m$; that is

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

The estimates of latent ability, $\hat{\theta}_i, i = 1,2,...,n$, ranged from 0 to m because each $\hat{P}_j(\hat{\theta}_i^{(j)}), (i = 1,2,...,n; j = 1,2,...,m)$ ranges from 0 to 1. However, the estimates of latent ability, $\hat{\theta}_i, i = 1,2,...,n$, are separately calculated for 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 were pooled and converted to percentile estimates on the uniform ability scale that ranged from 0 to 1. Then, the estimates of pooled latent ability, denoted by $\hat{\theta}_i', i = 1,2,...,2n$, were obtained based on the percentile rank of $\hat{\theta}_i, i = 1,2,...,n$ after reference and focal groups were combined.

Step 2: Calculate the frequency of matching subtest true scores. To calculate the frequency of matching subtest true scores, the $\hat{\theta}_i', i = 1,2,...,2n$ obtained from step

1 were aligned to the matching subtest true score $k, k = 0,1,2,...,m$. In this study, the centre value was used to categorize the estimates of latent ability. For example, if the estimate of latent ability $\hat{\theta}_i'$ for an examinee was 1.1, which fell in the interval of [0.5, 1.5), then the matching subtest true score of 1 was assigned to this examinee. In doing so, the estimates of latent ability can be aligned to matching subtest true and the result is denoted by $\hat{\theta}_k, k = 0,1,2,...,m$. Consequently, the frequency of each matching subtest true score for examinees could 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

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

where Y_i is the response to the studied item of the ith examinee, $\hat{\theta}_i'$ 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., $0/m, 1/m, 2/m, ..., m-1/m, m/m$). Again, $K(\frac{\hat{\theta}_i' - \theta_q}{h})$ 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 was 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 statistic: Cochran's Z, Fisher's χ^2*, and Goodman's U.* The final step is to calculate the three kernel-smoothed test statistics.

For the kernel-smoothed Cochran's *Z*, the null hypothesis is the same as the one described in Chapter II. 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:

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

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 as the same way as for kernel-smoothed Fisher's χ^2 described 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 a standard normal distribution.

For the kernel-smoothed Fisher's χ^2, the null hypothesis is the same as the one described in Chapter II. However, the formula for the test statistic with kernel smoothing applied is different from the one with correction regression applied:

$$\chi^2_{\hat{\theta}_k} = \frac{(N_{R\hat{\theta}_k} + N_{F\hat{\theta}_k} - 1)(n_{R\hat{\theta}_{k0}} n_{F\hat{\theta}_{k1}} - n_{R\hat{\theta}_{k1}} n_{F\hat{\theta}_{k0}})^2}{(n_{R\hat{\theta}_{k0}} + n_{R\hat{\theta}_{k1}})(n_{R\hat{\theta}_{k0}} + n_{F\hat{\theta}_{k0}})(n_{R\hat{\theta}_{k1}} + n_{F\hat{\theta}_{k1}})(n_{F\hat{\theta}_{k0}} + n_{F\hat{\theta}_{k1}})},$$

where $N_{R\hat{\theta}_k}$ and $N_{F\hat{\theta}_k}$ denote the total number of examinees having ability level $\hat{\theta}_k$ in the reference and focal group respectively, $n_{R\hat{\theta}_{k1}}$ and $n_{F\hat{\theta}_{k1}}$ denotes the kernel-smoothed number of examinees (frequencies obtained from step 2) in the reference and focal groups who obtained valid subtest score $\hat{\theta}_k$, and $n_{R\hat{\theta}_{k0}}$ and $n_{F\hat{\theta}_{k0}}$ denotes the kernel-smoothed number of examinees in each group who obtained valid subtest score $\hat{\theta}_k$. The test statistics can only be calculated when $N_{R\hat{\theta}_k} + N_{F\hat{\theta}_k} \geq 20$ and each of $n_{R\hat{\theta}_{k1}}$, $n_{F\hat{\theta}_{k1}}$, $n_{R\hat{\theta}_{k0}}$, and $n_{F\hat{\theta}_{k0}}$ is at least 3 for $\chi^2_{\hat{\theta}_k}$ to approximate the chi-square distribution with 1 degree of freedom.

The $\chi^2_{\hat{\theta}_k}$ statistic at each matching score level can also be summed across a range of matching score levels to produce an omnibus test of DIF:

$$\chi^2_{\hat{\theta}_T} = \sum_{k \in K} \chi^2_{\hat{\theta}_k}$$

where K consists of all matching score levels of interest that satisfy the necessary cell size criteria for a suitable χ^2 approximation. The test statistic $\chi^2_{\hat{\theta}_T}$ is distributed as a χ^2 with degrees of freedom equal to the number of matching score levels included in K.

For the Goodman's *U*, the kernel-smoothed formula is:

$$U = \sum_{k=0}^{m} \frac{(\hat{\Delta}_{\theta_k} - \hat{\Delta}_0)^2}{SE^2_{\hat{\Delta}_{\theta_k}}},$$

where $\hat{\Delta}_0 = \dfrac{\sum_{k=0}^{m} W_{\hat{\theta}_k} \hat{\Delta}_{\theta_k}}{\sum_{k=0}^{m} W_{\hat{\theta}_k}}$ is defined the same way as for the kernel-smoothed Cochran's *Z* test and $SE^2_{\hat{\Delta}_k} = \dfrac{\hat{P}_{R\hat{\theta}_k}(1-\hat{P}_{R\hat{\theta}_k})}{N_{R\hat{\theta}_k}} + \dfrac{\hat{P}_{F\hat{\theta}_k}(1-\hat{P}_{F\hat{\theta}_k})}{N_{F\hat{\theta}_k}}$ is the error variance.

Under the null hypothesis, U approximates the χ^2 distribution with degrees of freedom equal to the number of valid subtest scores providing that $N_{R\hat{\theta}_k} + N_{F\hat{\theta}_k} \geq 20$, and each of $n_{R\hat{\theta}_{k1}}$, $n_{F\hat{\theta}_{k1}}$, $n_{R\hat{\theta}_{k0}}$, and $n_{F\hat{\theta}_{k0}}$ is at least 3.

Type I Error Study

A simulation study was conducted to investigate the Type I error rates of the three proposed kernel-smoothed statistics in DIF detection. Three factors expected to affect the probability of a Type I error were considered: sample size, ability distribution difference, and item parameters of the studied item. 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. To compare the results of Type I error rates using regression correction with those from kernel-smoothed correction, the same 25 no-DIF matching items used in Bolt and Gierl's (2006) Type I error study were used in this simulation study. Table 1 contains the item parameters for the 25 items. Table 2 shows the summary information for these factors. In total, 3 (sample size) × 4 (ability distribution) × 4 (studied item) = 48 tests were generated to investigate the Type I error rates for the three proposed kernel-smoothed statistics. For each test, 100 replications were performed. Two-sided hypothesis tests were used with a significance level of 0.05. For comparison purpose, the Type I error rates with no correction and regression correction were also calculated under each simulation condition.

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

Matching Item	Item Parameter		
	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. Manipulated Factors in Type I Error Study

Sample Size		Ability Distribution		Item Parameter		
Ref.	Foc.	Ref.	Foc.	*a*	*B*	*c*
500	500	$N_R(0, 1)$	$N_F(0, 1)$	1.00	-0.75	0.20
1000	1000	$N_R(0, 1)$	$N_F(0, 2)$	1.00	0.75	0.20
2500	2500	$N_R(0.5, 1)$	$N_F(-0.5, 1)$	1.50	-0.75	0.20
		$N_R(0.5, 1)$	$N_F(-0.5, 2)$	1.50	0.75	0.20

Sample Size

The first factor manipulated was sample size. The results of many studies have revealed that sample sizes can influence the Type I error and power rates of DIF procedures (e.g., Bolt & Gierl, 2006; Douglas, Stout & DiBello, 1996; Jodoin & Gierl, 2001; Narayanan & Swaminathan, 1994; Roussos & Stout, 1996b). For

example, in Bolt and Gierl's study (2006), both the Type I error and power rates became inflated as the sample size increased for the three statistics using the regression correction. Three levels of sample size were considered in this simulation study. Let n_R and n_F denote the reference and focal group sample sizes, respectively. Values of n_R, n_F were set at (500, 500), (1000, 1000), and (2500, 2500), representing a small, moderate, and large sample sizes. Given the reference and focal groups have the same number of examinees, sample size was balanced across all conditions.

Ability Distribution

The second manipulated factor was the difference between the ability distributions of the reference and focal groups. Group ability mean was commonly used to reflect group differences in the previous DIF simulation studies (e.g., Douglas, Stout & DiBello, 1996; Fidalgo, Ferreres, & Muniz, 2004; Gierl, Jodoin, & Ackerman, 2000; Narayanan & Swaminathan, 1994; Roussos & Stout, 1996b). However, Bolt and Gierl (2006) suggested that group ability standard deviation differences can also cause substantial variation in Type I error and power rates of DIF procedures. Therefore, differences in both group ability mean and standard deviation were considered in the present study. Normal distributions for both groups were assumed. Four situations were considered: N_R (0, 1) versus N_F (0, 1), N_R (0, 1) versus N_F (0, 2), N_R (0.5, 1) versus N_F (-0.5, 1), and N_R (0.5, 1) versus N_F (-0.5, 2).

Item Parameters

Although item parameters were not a manipulated factor in Bolt and Gierl (2006)'s study, they pointed out that Type I error rates of the three statistics were inflated under certain items, especially for easy or difficult items. In addition, their study showed that the power rates of the three statistics varied when items differed in discrimination and/or difficulty. Consequently, the item parameters of the studied item were manipulated in the Type I error study.

A Type I error is found when a non-DIF item is detected as showing DIF. Therefore, to study the Type I error rates for the three statistics using kernel-smoothing, the studied item was simulated as a non-DIF item. The item parameters were the same for both the focal and the reference groups. The discrimination parameter was set at 1 or 1.5, the difficulty parameter was set at -0.75 or 0.75, and the guessing parameter was set at 0.2. In total, four items were used in the Type I error study. Their item parameters were (1, -0.75, 0.2), (1, 0.75, 0.2), (1.5, -0.75, 0.2), and (1.5, 0.75, 0.2), respectively. Therefore, each simulated test consisted of 25 matching items and one of these four items. The item parameters of the 26 items are the same for the reference and focal groups.

Power Study

The power study was designed to investigate the performance of the three kernel-smoothed statistics 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 of the studied items. The power study employed the same 25 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) × 4 (ability distribution) × 3 (studied item) = 36 tests were generated. For each test, 100 data sets were simulated. Two-sided 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 power rates with no correction and with regression correction were also calculated under each simulation condition.

Table 3. Parameters for Studied Items in Power Study

Studied Item	Group	Item parameter		
		a	*b*	*c*
1	Ref.	1.50	0.00	0.20
	Foc.	0.50	0.00	0.20
2	Ref.	2.00	1.00	0.20
	Foc.	0.40	0.00	0.20
3	Ref.	1.80	0.00	0.20
	Foc.	0.40	0.50	0.20

Chapter IV: Results

Type I Error Study

Type I Error results for Cochran's *Z*, Fisher's χ^2, and Goodman's *U* tests under the different simulation conditions described in the previous chapter are presented in Tables 4 to 6, respectively. Each table compares the Type I error rates for the corresponding test when no correction (NC), regression correction (RC), and kernel smoothing (KS) were used under each simulation condition. For each test, 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.

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.

Cochran's Z

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

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, α=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;

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 indicated that the three procedures produced 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 to 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 system 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 item1 and 0.01 for item 3), and the remaining 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 found 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 above, 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.

Fisher's χ^2

Table 5 displays the Type I error rates of the Fisher's χ^2 test across different conditions. As shown in Table 5, for the N_R (0, 1) and N_F (0, 1) condition, the empirical Type I error rates using NC were reasonable except for item 4 (0.01) when the sample size was 500/500. The Type I error rates using RC were reasonable across all the studied items. In contrast, the empirical Type I error rates using KS were conservative except for item 4 (0.03). When the sample size was 1000/1000, the empirical Type I error rates using NC and RC were reasonable across all the studied items. Alternatively, the Type I error rates using KS were conservative except for item 4 (0.02). When the sample size was large (2500/2500), the empirical Type I error using NC and RC were reasonable, ranging from 0.05 to 0.07. The empirical Type I error rates using KS were conservative except for item 4 (0.03). Therefore, the empirical Type I error rates using the three procedures were stable as sample size increased. The rates using NC and RC tended to be reasonable while the rates using KS tended to be conservative when there was no ability difference. Influence of item parameters was found under this condition. The Type I error rates using KS for item 4 (a=1.5, b=0.75, and c=0.20) were reasonable while the rates for the remaining items (items 1, 2, and 3) were conservative as sample size varied.

For the N_R (0, 1) and N_F (0, 2) condition, when sample size was small, the Type I error rates using NC were reasonable except for item 1 (0.11). The Type I error rates

Table 5. Type I Error for Fisher's χ^2 with No Correction, Regression Correction, and Kernel Smoothing

(Proportion of rejections out of 100 replications for each condition, α=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	0.02	0.01	0.05	0.05	0.00	0.06	0.05	0.00
	Item 2	0.02	0.03	0.00	0.03	0.03	0.00	0.07	0.06	0.01
	Item 3	0.05	0.05	0.00	0.04	0.05	0.00	0.06	0.06	0.01
	Item 4	0.01	0.04	0.03	0.05	0.05	0.02	0.07	0.07	0.03
N_R (0,1), N_F (0,2)	*Item 1*	0.11	0.05	0.00	0.15	0.12	0.00	0.26	0.07	0.00
	Item 2	0.05	0.04	0.00	0.17	0.12	0.00	0.29	0.09	0.00
	Item 3	0.07	0.03	0.00	0.14	0.05	0.00	0.22	0.08	0.02
	Item 4	0.05	0.03	0.00	0.14	0.07	0.01	0.44	0.12	0.02
N_R (0.5,1), N_F (-0.5,1)	*Item 1*	0.06	0.05	0.01	0.29	0.09	0.00	0.57	0.07	0.06
	Item 2	0.25	0.16	0.01	0.25	0.12	0.03	0.58	0.12	0.01
	Item 3	0.08	0.02	0.00	0.39	0.12	0.04	0.81	0.09	0.03
	Item 4	0.25	0.08	0.00	0.34	0.07	0.00	0.75	0.07	0.03
N_R (0.5,1), N_F (-0.5,2)	*Item 1*	0.08	0.04	0.00	0.39	0.19	0.02	0.48	0.09	0.00
	Item 2	0.05	0.07	0.00	0.08	0.07	0.01	0.08	0.06	0.00
	Item 3	0.07	0.00	0.00	0.42	0.13	0.00	0.84	0.10	0.05
	Item 4	0.07	0.08	0.00	0.08	0.11	0.00	0.07	0.11	0.01

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

using RC were reasonable while the rates using KS were conservative across all the studied items. When the sample size was moderate, the Type I error rates using NC increased noticeably: the Type I errors were liberal for all four items, ranging from 0.14 to 0.17. Using RC, the empirical Type I error rates for items 1 and 2 were liberal (both were 0.12) while the rates for items 3 and 4 were reasonable (0.05 and 0.07, respectively). The Type I error rates using KS were conservative across the studied items (0.00 for items 1, 2 and 3, 0.01 for item 4). When sample size was large, the empirical Type I error rates using NC were again liberal, ranging from 0.22 to 0.44. In addition, they were approximately two times higher than the value for the moderate sample size condition with one exception (approximately three times for item 4). The empirical Type I error rates using RC under this condition were reasonable for items 1 and 3 (0.07 and 0.08, respectively) and liberal for items 2 and 4 (0.09 and 0.12, respectively). Using KS, the empirical Type I error rates were conservative for items 1 and 2 (both were 0.00) and reasonable for items 3 and 4 (both were 0.02). As sample size increased, the Type I error rates were inflated noticeably using NC, tended to be liberal using RC, and tended to be conservative consistently using KS. No systematic influence of item parameters was found for the three procedures in this condition.

For the N_R (0.5, 1) and N_F (-0.5, 1) condition, when sample size was small, the empirical Type I error rates using NC for item 1 and 3 were reasonable (0.06 and 0.08, respectively). On the other hand, the rates for item 2 and 4 were liberal (both were 0.25). Using RC, the empirical Type I error rates were reasonable with one

exception (0.16 for item 2). The empirical Type I error rates using KS were conservative across all four studied items. When the sample size was moderate, the empirical Type I error rates using NC were liberal, ranging from 0.25 to 0.39. Using RC, the Type I error rates were also liberal with one exception (0.07 for item 4). In contrast, the rates using KS were conservative for items 1 and 4 (both were 0.00) and reasonable for items 2 and 3 (0.03 and 0.04, respectively). When the sample size was large, the empirical Type I error rates using NC were liberal and larger than that for the moderate sample size, ranging from 0.57 to 0.81. The Type I error rates using RC were reasonable for items 1 and 4 (both were 0.07) and liberal for items 2 and 3 (0.12 and 0.09, respectively). Using KS, however, the empirical Type I error rates were reasonable with one exception (0.01 for item 2). When there was mean ability distribution difference, influence of item parameters was found for NC. The Type I errors using NC for difficult items (items 2 and 4) were liberal regardless of sample size. In contrast, the Type I error rates for easy items (items 1 and 3) were reasonable when the sample size was small and then became liberal as the sample size increased.

For the N_R (0.5, 1) and N_F (-0.5, 2) condition, when sample size was small, the empirical Type I error using NC and RC were reasonable with one exception (0.00 for item 3 using RC). However, the Type I error rates using KS were conservative for all studied items. When the sample size was moderate, the Type I error rates using NC were liberal for items 1 and 3 (0.39 and 0.42, respectively), while the rates for items 2 and 4 were reasonable (both were 0.08). The empirical Type I errors using RC were liberal, ranging from 0.11 to 0.19, with one exception (0.07 for item 2). But the rates

were lower than the rates using NC. The Type I error rates using KS were conservative except for item 1 (0.02). When the sample size was large, the empirical Type I error rates using NC for items 1 and 3 were liberal and larger than for the moderate size condition (0.48 vs. 0.39 for item 1, 0.84 vs. 0.42 for item 3). On the other hand, the rates for item 2 and 4 were still reasonable (0.08 and 0.07, respectively). Using RC, the Type I error rates were liberal with one exception (0.06 for item 2). Using KS, the empirical Type I error rates were conservative except for item 3 (0.05). Like the pattern of the results for Cochran's Z test, the b-parameter, influenced the Type I error rates using NC when there were differences between the mean abilities and the standard deviations. As sample size increased, the empirical Type I error rates using NC for easy items (b-parameters for items 1 and 3 were -0.75) increased. For example, the Type I error rates for item 3 increased from 0.07, to 0.42, and then to 0.84 when sample size increased from small, to moderate, and to large. On the other hand, the empirical Type I error rates using NC were reasonable across different sample size conditions for difficult items (b-parameters for items 2 and 4 were 0.75). For example, for item 4, the Type I error rates were 0.07, 0.08, and 0.07 when the sample sizes were 500/500, 1000/1000, and 2500/2500. In contrast, item parameter did not influence the Type I error when RC and KS were used.

To summarize, for Fisher's χ^2, the empirical Type I error rates using NC were affected by the manipulated factors, including sample size, ability distribution, and item parameters. When the sample size was small, the Type I error rates using NC were reasonable with few exceptions (0.25 for item 2 and 0.25 for item 4) under the

N_R (0.5, 1) and N_F (-0.5, 1) condition. However, the Type I error rates using NC increased as sample size increased and ability mean and standard deviation each differed by one. In addition, under the N_R (0.5, 1) and N_F (-0.5, 2) condition, Type I error rates using NC were influenced by item difficulty. Using RC, the empirical Type I error rates increased as sample size increased for items 1 and 3. The rates also increased when there were differences between the ability distributions for items 1 and 3. However, the degree of increase for RC was smaller than for NC. Using KS, the empirical Type I error rates were conservative or reasonable across simulation conditions.

Goodman's U

The results for Goodman's *U* test using NC, RC and KS are summarized in Table 6. For the N_R (0, 1) and N_F (0, 1) condition, when the sample size was small, the empirical Type I error rates using NC and RS were reasonable, with one exception (0.01 for item 4 using NC). The Type I error rates using KS were conservative across all four studied items. When the sample size was moderate and large, the Type I error rates using NC and RC were reasonable while the rates using KS were conservative across all studied items. The sample size and item parameter did not show any impact on the Type I error rates using the three procedures when there was no difference between ability means and standard deviations.

Table 6. Type I Error for Goodman's U with No Correction, Regression Correction, and Kernel Smoothing

(Proportion of rejections out of 100 replications for each condition, α=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	0.02	0.00	0.04	0.04	0.00	0.06	0.05	0.00
	Item 2	0.02	0.02	0.00	0.03	0.04	0.00	0.07	0.06	0.00
	Item 3	0.05	0.04	0.00	0.04	0.06	0.00	0.05	0.06	0.00
	Item 4	0.01	0.02	0.00	0.03	0.04	0.00	0.05	0.06	0.00
N_R (0,1), N_F (0,2)	*Item 1*	0.01	0.02	0.00	0.14	0.09	0.00	0.18	0.02	0.00
	Item 2	0.06	0.04	0.00	0.17	0.11	0.00	0.15	0.09	0.00
	Item 3	0.05	0.03	0.04	0.09	0.03	0.00	0.16	0.06	0.00
	Item 4	0.02	0.03	0.00	0.11	0.06	0.00	0.25	0.11	0.00
N_R (0.5,1), N_F (-0.5,1)	*Item 1*	0.01	0.04	0.00	0.07	0.07	0.00	0.03	0.02	0.00
	Item 2	0.04	0.07	0.00	0.07	0.08	0.00	0.05	0.08	0.00
	Item 3	0.04	0.02	0.00	0.14	0.12	0.00	0.14	0.07	0.00
	Item 4	0.02	0.04	0.00	0.04	0.04	0.00	0.12	0.04	0.01
N_R (0.5,1), N_F (-0.5,2)	*Item 1*	0.04	0.03	0.00	0.11	0.12	0.00	0.27	0.05	0.00
	Item 2	0.04	0.05	0.00	0.07	0.09	0.00	0.05	0.07	0.00
	Item 3	0.00	0.00	0.00	0.14	0.11	0.00	0.37	0.07	0.00
	Item 4	0.06	0.07	0.00	0.05	0.07	0.00	0.06	0.08	0.00

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

For the N_R (0, 1) and N_F (0, 2) condition, when the sample size was small, the empirical Type I error rates using NC were reasonable except for item 1 (0.01), while the rates using RC were reasonable for all studied items. Using KS, the Type I error rates were conservative except for item 3 (0.04). When the sample size was moderate, all four Type I error rates using NC were liberal, ranging from 0.09 to 0.17. Two rates using RC were also liberal, 0.09 for item 1 and 0.11 for item 2. All the Type I error rates using KS were conservative. When the sample size was large, the Type I error rates using NC were liberal for all studied items, ranging from 0.15 to 0.25. The error rates using RC were liberal for items 2 (0.09) and 4 (0.11), and reasonable for items 1 (0.02) and 3 (0.06). Using KS, the error rates were again conservative (0.00 for all items). When standard deviation differences presented, the empirical Type I error rates using NC and RC became inflated as sample size increased. While the Type I error rates using KS were consistently conservative across the different sample size conditions. In contrast, item parameter did not influence the Type I error when NC, RC, and KS were used.

For the N_R (0.5, 1) and N_F (-0.5, 1) condition, when the sample size was small, the empirical Type I error rates using NC and RC were reasonable with one exception (0.01 for item 1 using RC). The error rates using KS were conservative for all items. When the sample size was moderate, the rates using NC and RC were reasonable except for item 3 (0.14 using NC and 0.12 using RC). The rates using KS were still conservative for all items. When the sample size was large, two of the Type I error rates using NC (0.03 and 0.05 for items 1 and 2) were reasonable and two of the rates

(0.14 and 0.12 for items 3 and 4) were liberal. All the Type I error rates using RC were reasonable, ranging from 0.02 to 0.08. Using KS, the empirical Type I error rates were conservative for all the items. When there was difference between the mean abilities, the empirical Type I error rates using NC were inflated for a few items as sample size increased. While the Type I error rates using RC were reasonable as sample size increased with one exception (item 3 under moderate sample size). Again, the rates using KS were consistently conservative across the different sample size conditions. The item discrimination parameter (a parameter) influenced the Type I errors when NC was used. Using NC, the Type I error rates for items 1 and 2 (a=1.00) were reasonable for all three sample sizes while the rates for items 3 and 4 (a=1.50) were reasonable when the sample size was small and liberal when the sample sizes were moderate and large (with one exception of 0.04 for item 4 when the sample size was moderate). In contrast, item discrimination did not influence the Type I error when RC and KS were used.

For the N_R (0.5, 1) and N_F (-0.5, 2) condition, when the sample size was small, the empirical Type I errors using NC and RC were reasonable except for item 3 (0.00 for both NC and RC). The Type I error rates using KS were conservative for all studied items. When the sample size was moderate, the Type I error rates using NC for items 1 and 3 were liberal (0.11 and 0.14, respectively), while the error rates for items 2 and 4 were reasonable (0.07 and 0.05, respectively). The empirical Type I error using RC were liberal, ranging from 0.11 to 0.19, with one exception (0.07 for item 4). The Type I error rates using KS were conservative for all studied items.

When the sample size was large, the empirical Type I error rates using NC for item 1 and 3 were liberal (0.27 and 0.37, respectively). On the other hand, the rates for item 2 and 4 were reasonable (0.05 and 0.06, respectively). Using RC, the Type I error rates were liberal for all four studied items. Using KS, the empirical Type I error rates were conservative across all items. When both ability means and standard deviations differed between the reference and focal groups, the Type I error rates using NC were influenced by item difficulty (b-parameter) as in Cochran's Z and Fisher's χ^2 tests. As sample size increased, the empirical Type I error rates using NC for difficult items (b-parameters for items 1 and 3 were 0.75) increased. For example, the Type I error rates for item 1 increased from 0.04 to 0.11 and then to 0.27 when sample size increased from small to moderate and to large. On the other hand, the empirical Type I error rates using NC were reasonable across different sample size conditions for easy items (b-parameters for items 2 and 4 were -0.75). For example, for item 4, the Type I error rates were 0.06, 0.05, and 0.06 when the sample sizes were 500/500, 1000/1000, and 2500/2500. In contrast, item difficulty did not influence the Type I error when RC and KS were used.

To summarize, for Goodman's U statistic, the manipulated factors affected the empirical Type I error rates when NC was applied. When there were no ability differences between two groups, the Type I error rates using NC were reasonable regardless of sample size. Conversely, when the ability means differed, the Type I error rates turned to liberal across all items when the sample sizes were moderate and large. Item parameters influenced the Type I error rates using NC when ability means

and standard deviations differed between two groups (i.e. the N_R (0.5, 1) and N_F (-0.5, 2) condition). The Type error rates tended to be inflated for easy items (b = -0.75) than the rates for the difficult items (b = 0.75). Using RC, the empirical Type I error rates was influenced by sample size and ability distribution. When there was no ability difference, the Type I error rates using RC were reasonable as sample size increased. When there was ability difference, the Type I error rates tended to be liberal for items 1 and 3, especially under moderate sample size. When KS was used for Goodman's U, the empirical Type I error rates were consistently conservative level (most of them were equal to zeros).

Power Study

The results of power study for Cochran's Z, Fisher's χ^2, and Goodman's U tests under the different simulation conditions considered in this study are presented in Tables 7 to 9. In the power study, three distinct DIF items were studied, ranging from easy to difficult with varying amounts of discrimination. The selection of these items was based on the Douglas, Stout and DiBello's study (1996). The item parameters for each of these items were provided in Table 4. The remaining factors and their levels, sample size, and ability distribution were the same as those considered in the Type I error study.

In order to interpret the power results, power rates 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.

Cochran's Z

Table 7 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. In contrast, 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).

Table 7. Power Rates for Cochran's Z with No Correction, Regression Correction, and Kernel Smoothing

(Proportion of rejections out of 100 replications for each condition, α=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;

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 shared the similar trend 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 higher power rates across different conditions for Cochran's *Z*. The results showed the superiority using KS procedure for Cochran's *Z* to detect DIF.

Fisher's χ^2

The power results using NC, RC, and KS for Fisher's χ^2 are summarized in Table 8. 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 for item 1 were low (0.41, 0.38, and 0.19, respectively). For item 2, the power rates using all three procedures were high (1.00 for all of them). For item 3, the power rates using NC, RC, and KS were moderate (0.73, 0.76, and 0.79, respectively). When the sample size was moderate, the power rates were high across the three procedures and studied items with one exception (0.47 for item 1 using KS), ranging from 0.92 to 1.00. When sample size was large, all the power rates were high for all the procedures and studied items, ranging from 0.96 to 1.00.

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 for item 1 were low (0.27, 0.33, and 0.07, respectively). For item 2, the power rates using NC, RC, and KS were high (1.00, 1.00, and 0.99, respectively). For item 3, the power rates using NC were low (0.45), while the rates using RC and KS were moderate (0.63, and 0.53). When sample size was moderate, the power rates for item 1 were moderate (0.70) using NC, high (0.80) using RC, and low (0.35) using KS. For items 2 and 3, the power rates using NC, RC,

Table 8. Power Rates for Fisher's χ^2 with No Correction, Regression Correction, and Kernel Smoothing

(Proportion of rejections out of 100 replications for each condition, α=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*	0.41	0.38	0.19	0.92	0.92	0.47	1.00	1.00	0.96
	item 2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	item 3	0.73	0.76	0.79	0.99	0.99	0.98	1.00	1.00	1.00
$N_R(0,1), N_F(0,2)$	*item 1*	0.27	0.33	0.07	0.70	0.80	0.35	1.00	1.00	1.00
	item 2	1.00	1.00	0.99	1.00	1.00	1.00	1.00	1.00	1.00
	item 3	0.45	0.63	0.53	0.99	0.98	0.99	1.00	1.00	1.00
$N_R(0.5,1), N_F(-0.5,1)$	*item 1*	0.62	0.49	0.18	0.97	0.91	0.60	1.00	1.00	0.93
	item 2	1.00	1.00	0.99	1.00	1.00	1.00	1.00	1.00	1.00
	item 3	0.95	0.82	0.86	1.00	0.99	1.00	1.00	1.00	1.00
$N_R(0.5,1), N_F(-0.5,2)$	*item 1*	0.34	0.40	0.44	0.91	0.91	0.92	1.00	1.00	1.00
	item 2	0.99	1.00	0.79	1.00	1.00	1.00	1.00	1.00	1.00
	item 3	0.80	0.80	0.95	1.00	0.99	1.00	1.00	1.00	1.00

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

and KS were high, ranging from 0.98 to 1.00. When sample size was large, the power rates were high (1.00) across all procedures and items.

For the N_R (0.5, 1) and N_F (-0.5, 1) condition, when the sample size was small, the power rates using NC and RC for item 1 were moderate (0.62 and 0.49, respectively), while the power rate using KS was low (0.18). For items 2 and 3, the power rates were high across NC, RC, and KS, ranging from 0.82 to 0.95. When the sample sizes were moderate and large, the power rates were high across all procedures and studied items with one exception (0.60 for item 1 using KS when the sample size was moderate), ranging from 0.91 to 1.00.

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, RC, and KS were low for item 1 (0.34, 0.40, and 0.44, respectively). For items 2 and 3, the power rates using NC, RC, and KS were high with one exception (0.79 for item 2 using KS), ranging from 0.80 to 1.00. When the sample sizes were moderate and large, the power rates were high across procedures and items, ranging from 0.91 to 1.00.

To summarize, for item 1, all procedures yielded low power with two exceptions (0.62 for NC and 0.49 for RC at N_R (0.5, 1) and N_F (-0.5, 1)) when the sample size was small. NC and RC produced high power with one exception (0.70 for NC at N_R (0, 1) and N_F (0, 2) when the sample size was moderate), while KS produced moderate to high power with two exceptions (0.47 and 0.35 when there was no mean difference and sample size was moderate) when the sample sizes were moderate and large. For item 2, all procedures produced high power rates with one

exception (0.79 for KS at N_R (0.5, 1) and N_F (-0.5, 2) when the sample size was small). For item 3, all procedures yielded moderate to high power rates across different sample sizes and ability distributions with one exception (0.45 for NC when the sample size was small at N_R (0, 1) and N_F (0, 2)).

Goodman's U

Table 9 shows the power study results using NC, RC, and KS for Goodman's *U* test. For the N_R (0, 1) and N_F (0, 1) condition, when the sample size was small, the power rates using NC and RC were low across studied items, ranging from 0.30 to 0.35, with the exception of item 3 using RC (0.70). In contrast, the power rates using KS were 0.00 for all items. When the sample size was moderate, the power rates using NC and RC were high across all studied items, ranging from 0.89 to 1.00. However, the power rate using KS was low for the three studied items (0.08 for item 1, 0.01 for item 2, and 0.15 for item 3). When the sample size was large, the power rates using NC and RC were all high (1.00) across items. In contrast, the power rates using KS were low for items 1 and 2 (0.40 and 0.29, respectively) and high for item 3 (0.86).

For the N_R (0, 1) and N_F (0, 2) condition, when the sample size was small, the power rates using NC and RC were, with one exception, low across the three studied items, ranging from 0.26 to 0.42. The power rates for item 3 using RC was moderate (0.51). In contrast, the power rates using KS were 0.00 for all items. When the sample size was moderate, the power rates using NC and RC were moderate for item 1 (0.64 and 0.77, respectively). However, the rates using NC and RC were high for items 2

Table 9. Power Rates for Goodman's U with No Correction, Regression Correction, and Kernel Smoothing

(Proportion of rejections out of 100 replications for each condition, α=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*	0.35	0.33	0.00	0.93	0.92	0.08	1.00	1.00	0.40
	item 2	0.30	0.32	0.00	0.89	0.89	0.01	1.00	1.00	0.29
	item 3	0.30	0.70	0.00	0.99	1.00	0.15	1.00	1.00	0.86
$N_R(0,1), N_F(0,2)$	*item 1*	0.26	0.33	0.00	0.64	0.77	0.00	0.98	1.00	0.31
	item 2	0.33	0.38	0.00	0.83	0.90	0.05	1.00	1.00	0.73
	item 3	0.42	0.51	0.00	0.96	0.95	0.05	1.00	1.00	0.83
$N_R(0.5,1), N_F(-0.5,1)$	*item 1*	0.35	0.36	0.00	0.81	0.87	0.00	1.00	1.00	0.04
	item 2	0.31	0.37	0.00	0.86	0.82	0.00	1.00	1.00	0.06
	item 3	0.54	0.59	0.00	0.98	0.98	0.01	1.00	1.00	0.19
$N_R(0.5,1), N_F(-0.5,2)$	*item 1*	0.14	0.28	0.00	0.46	0.70	0.00	0.98	1.00	0.03
	item 2	0.43	0.48	0.00	0.93	0.96	0.27	1.00	1.00	1.00
	item 3	0.37	0.53	0.00	0.86	0.92	0.00	1.00	1.00	0.08

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

and 3. Using KS, the power rates were low across studied items (0.00 for item 1, 0.05 for item 2, and 0.05 for item 3).When the sample size was large, the power rates using NC and RC were high across studied items, ranging from 0.98 to 1.00. The power rates using KS were low for item 1 (0.31), moderate for item 2 (0.73), and high for item 3 (0.83).

For the N_R (0.5, 1) and N_F (-0.5, 1) condition, when the sample size was small, the power rates using NC and RC were low for items 1 and 2, ranging from 0.31 to 0.37, while the rates using NC and RC were moderate for item 3 (0.54 and 0.59, respectively). In contrast, the powers using KS were 0.00 for all items. When the sample size was moderate, the power rates using NC and RC were high across all studied items, ranging from 0.81 to 0.98. However, the power rate using KS was low for the three studied items (0.00 for item 1, 0.00 for item 2, and 0.01 for item 3). When the sample size was large, the power rates using NC and RC were all high (1.00) across items. Meanwhile, the power rates using KS were low for the three items (0.04, 0.06, and 0.19, respectively).

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 were low for item 1 (0.14 and 0.28, respectively). For items 2 and 3, the rates were low using NC (0.43 and 0.37, respectively) and moderate using RC (0.48 and 0.53, respectively). Again, the power rates using KS were 0.00 for all items. When the sample size was moderate, for item 1, the power rates were low using NC (0.46) and moderate using RC (0.70). For items 2 and 3, the rates were high when NC and RC were used, ranging from 0.86 to 0.96. However, the power rate using KS was low for the three studied items (0.00 for item 1, 0.27 for item 2, and 0.00 for item 3). When the sample size was large, the power rates using NC and RC were all high across items, ranging from 0.98 to 1.00. Meanwhile, the

power rates using KS were low for items 1 and 3 (0.03 and 0.08, respectively) and high for item 2 (1.00).

To summarize, for item 1, NC and RC yielded low power when the sample size was small. When the sample sizes were moderate, NC and RC produced moderate to high power rates with one exception (0.46 using NC at N_R (0.5, 1), N_F (-0.5, 2) conditions). When the sample size was large, NC and RC yielded high power. However, KS produced low power for item 1 regardless sample size and ability distributions. For item 2, NC and RC yielded low power when the sample size was small with one exception (0.48 using RC at N_R (0.5, 1) and N_F (-0.5, 2)). When the sample size was moderate and large, NC and RC produced high power rates. However, KS produced low power rates regardless sample size and ability distribution with two exceptions (0.73 and 1.00 at N_R (0, 1), N_F (0, 2) and N_R (0.5, 1), N_F (-0.5, 2) conditions when the sample size was large). For item 3, NC yielded low power rates while RC yielded moderate power rates when the sample size was small. When the sample sizes were moderate and large, NC and RC produced high power rates for item 3 as the ability distribution varied. However, KS yielded low power rates for item 3 regardless sample size and ability distribution with two exceptions (0.86 and 0.83 at N_R (0, 1), N_F (0, 1) and N_R (0, 1), N_F (0, 2) conditions when the sample size was large).

Chapter V: Discussion and Future Directions

Summary of Purpose and Method

Recently, practitioners and researchers have become interested in the graphical comparison of non-parametrically estimated IRFs for DIF analysis. This is 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 three nonparametric DIF statistics—Cochran's *Z*, Fisher's χ^2, and Goodman's *U*— 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) statistics. For Type I error study, three factors expected to affect the probability of a Type I error were considered: 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) × 4 (ability distribution) × 4 (studied item) = 48 tests were generated to investigate the Type I error rates for the three proposed kernel-smoothed statistics. 100 replications were performed for each test. For power study, the three factors as the same as in Type I error study 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) × 4 (ability distribution) × 3 (studied item) = 36 tests 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) statistics were compared to those with No Correction (NC) and Regression Correction (RC) to evaluate the performance of the new statistics introduced in this study.

Summary of Main Findings and Conclusion

Type I Errors

The summary findings are presented in terms of the percentages of Type I errors that were classified as conservative, moderate, and liberal. As shown in Table 10, the percentages are classified by sample size (small (500,500), moderate (1000, 1000),

and large (2500, 2500) given the results revealed that Type I error was influenced by sample size. The percentages are further classified by statistics used (Cochran's Z, Fisher's χ^2, and Goodman' U) and procedures (no correction (NC), regression correction (RC), and kernel smoothing (KS). The discussion is organized in terms of sample size.

Table 10. Percentages of Type I Errors Classified as Conservative, Moderate, and Liberal according to Statistics and Procedures

	Sample size	Small			Moderate			Large		
		Z	χ^2	U	Z	χ^2	U	Z	χ^2	U
Conservative	NC	0.0	6.3	25.0	6.3	0.0	0.0	12.5	0.0	0.0
	RC	18.8	6.3	6.3	31.3	0.0	0.0	43.8	0.0	0.0
	KS	0.0	93.8	93.8	0.0	75.0	100.0	0.0	56.3	100.0
Reasonable	NC	87.5	75.0	75.0	43.8	37.5	56.3	37.5	37.5	50.0
	RC	81.3	87.5	93.8	68.8	50.0	62.5	56.3	56.3	87.5
	KS	93.8	6.3	6.3	43.8	25.0	0.0	56.3	43.8	0.0
Liberal	NC	12.5	18.8	0.0	50.0	62.5	43.8	50.0	62.5	50.0
	RC	0.0	6.3	0.0	0.0	50.0	37.5	0.0	43.8	12.5
	KS	6.3	0.0	0.0	56.3	0.0	0.0	43.8	0.0	0.0

Note: Z for Cochran's Z, χ^2 for Fisher's χ^2, and U for Goodman's U;
NC for No Correction, RC for Regression Correction, and KS for Kernel Smoothing.

Small sample size. For Cochran's *Z*, the majority of Type I errors were reasonable using NC (87.5%), using RC (81.3%), and using KS (93.8%). There appears to be superior using KS compared with using NC or RC. For Fisher's χ^2, the majority of Type I error rates were also reasonable using NC (75.0%) and RC (87.5%). In contrast, the majority of Type I error rates were conservative using KS (93.8%). For Goodman's *U*, the similar pattern as for Fisher's χ^2 was found. That is, the majority of Type I error rates were reasonable using NC (75.0%) and RC (93.8%), while the majority of Type I error were conservative using KS (93.8%).

Moderate sample size. Comparison of the pattern of percentages for the moderate size sample size condition with the pattern noted above for the small sample size condition reveals that there was an interaction between sample size, test statistic, and procedures. For example, while the majority of Type I errors for Cochran's Z were reasonable using RC (68.8%), the majority was less than that observed when the sample size was small (68.8% vs. 87.5%). Further, 31.3% of Type I errors using RC were conservative when the sample size was moderate, but 18.8% when the sample size was small. In contrast, the majority of Type I errors using NC were liberal (50.0%), while only 12.5% of Type I errors were liberal when the sample size was small. Similarly, 56.3% of Type I errors were liberal using KS when the sample size was moderate, but only 6.3% Type I errors was liberal when the sample size was small.

In contrast, the incidence of Type I errors for Fisher's χ^2 and Goodman's U was more evenly divided between reasonable and liberal when NC and RC were used. Take Fisher's χ^2 as an example, 37.5% and 50.0% were reasonable using NC and RC, while 62.5% and 50.0% were liberal. While all of the conservative Type I errors occurred for KS when the sample size was moderate, no reasonable and liberal Type I errors occurred for KS. In contrast, the liberal Type I errors for Fisher's χ^2 and Goodman's U are essentially evenly divided between NC and RC.

Large sample size. The patterns of results for the large and moderate sample sizes are more comparable to each other than to the pattern for the small sample size, thereby clarifying the nature of the contribution of sample size to the interaction noted above. For example, while a large majority of Type I errors were reasonable for Cochran's Z (87.5%), when the sample size was small, the Type I errors were more

evenly divided between the three error size intervals for both Fisher's χ^2 and Goodman's *U* for both moderate and large sample sizes.

In summary, the attention was only paid to reasonable Type I errors for KS procedure since the main purpose of present study is to find a suitable statistic to detect the occurrence of DIF when kernel smoothing procedure was applied. Cochran's Z produced better Type I errors than Fisher's χ^2 and Goodman's U when kernel smoothing was used across the three sample sizes (93.8% vs. 6.3% and 6.3% for small sample size, 43.8% vs. 25.0% and 0.0% for moderate sample size, and 56.3% vs. 43.8% and 0.0% for large sample size). However, the percentages of reasonable Type I errors for the kernel-smoothed Cochran's Z were not the highest one when the sample sizes were moderate and large. The Regression-corrected Cochran's Z produced the best Type I errors (68.8%) under moderate sample size and the regression-corrected Goodman's U had the best Type I errors (87.5%) under large sample size.

Power

The summary findings are presented in terms of the percentages of power that were classified as low, moderate, and high. As shown in Table 11, the percentages are classified by sample size (small (500,500), moderate (1000, 1000), and large (2500, 2500). Like the structure of Type I errors, the percentages are further classified by statistics used (Cochran's Z, Fisher's χ^2, and Goodman' U) and procedures (no correction (NC), regression correction (RC), and kernel smoothing (KS). The discussion is also organized in terms of sample size.

Small sample size. For Cochran's *Z*, the majority of power rates were low using NC (66.7%) and using RC (66.7%). In contrast, the majority of power rates were high using KS (58.3%). Therefore, there appears to be superior using KS compared with

using NC or RC. For Fisher's χ^2, half of power rates using NC were low while half of power rates using RC were high. The power rates using KS were approximately evenly divided among low (33.3%), moderate (25.0%), and high (41.7%). This finding indicates that RC and KS improved the power performance of Fisher's χ^2 and RC produced even better results than KS (50.0% vs. 41.7%). For Goodman's *U*, the majority of power rates were low using NC (91.7%) and KS (100.0%). In contrast, the power rates using RC were evenly divided between low and moderate. This suggests that none of the three procedures had adequate power for Goodman's U when the sample size was small.

Table 11. Percentages of Power Classified as Low, Moderate, and High according to Statistics and Procedures

	Sample size	Small			Moderate			Large		
		Z	χ^2	U	Z	χ^2	U	Z	χ^2	U
Low	NC	66.7	50.0	91.7	50.0	0.0	8.3	25.0	0.0	0.0
	RC	66.7	25.0	58.3	58.3	0.0	0.0	33.3	0.0	0.0
	KS	25.0	33.3	100.0	8.3	16.7	100.0	0.0	0.0	66.7
Moderate	NC	8.3	16.7	8.3	16.7	8.3	8.3	8.3	0.0	0.0
	RC	8.3	25.0	41.7	8.3	0.0	16.7	0.0	0.0	0.0
	KS	16.7	25.0	0.0	16.7	8.3	0.0	0.0	0.0	8.3
High	NC	25.0	33.3	0.0	33.3	91.7	83.3	66.7	100.0	100.0
	RC	25.0	50.0	0.0	33.3	100.0	83.3	66.7	100.0	100.0
	KS	58.3	41.7	0.0	75.0	75.0	0.0	100.0	100.0	25.0

Note: Z for Cochran's Z, χ^2 for Fisher's χ^2, and U for Goodman's U;
NC for No Correction, RC for Regression Correction, and KS for Kernel Smoothing.

Moderate sample size. For Cochran's Z, the patterns of results for the moderate sample size are comparable with the results for the small sample sizes. The majority of power rates were also low using NC (50.0%), using RC (58.3%), but the majority was less than that observed when the sample size was small (50.0% vs.66.7% for NC, 58.3% vs. 66.7% for RC). In contrast, the majority of power rates were high using KS (75.0%) and the majority was more than that observed when the sample size was

small (75.0% vs. 58.3%). For Fisher's χ^2, the majority of power rates were high regardless using NC (91.7%), using RC (100.0%), or using KS (75.0%). In contrast, while the majority of power rates for Goodman's U were high using NC (83.3%) and RC (83.3%), all of power rates were low using KS.

Large sample size. The patterns of power rates for the large sample size are more convergent than the small and moderate sample sizes. For example, for Cochran's Z, two of third power rates were high using NC and RC. The remaining rates were low. For Fisher's χ^2, all the power rates were high across three procedures. In contrast, while all the power rates for Goodman's U were high using NC and RC, the majority of power rates were low using KS (66.7%), which shared the similar pattern with the moderate sample size.

In summary, the attention was paid to high power level for KS procedure this time. Cochran's Z produced better power than Fisher's χ^2 and Goodman's U when kernel smoothing was used under small sample size (58.3% vs. 41.7% and 0.0%). In contrast, under the condition of moderate and large sample sizes, Cochran's Z and Fisher's χ^2 produced the same better power than Goodman's U when kernel smoothing was applied. Again, the best power rate under moderate sample size was regression-corrected Fisher's χ^2 (100.0%) instead of kernel-smoothed Cochran's Z (75.0%).

Conclusion

In this study, the kernel smoothing procedure was applied to three nonparametric DIF statistics—Cochran's *Z*, Fisher's χ^2, and Goodman's *U*—to statistically test the difference between the kernel-smoothed IRF for reference and focal groups. Simulation studies were conducted to investigate the Type I errors and

power performance for these statistics. Results indicated that, among the three statistics, Cochran's Z showed the best performance in detecting the kernel-smoothed IRF differences for reference and focal groups under small sample size conditions. In comparison, when the sample size was moderate or large, both Cochran's Z and Fisher's χ^2 produced relatively high power rates in detecting kernel-smoothed IRF differences. However, the Type I errors of kernel-smoothed Cochran's Z tend to be liberal while the Type I errors of kernel-smoothed Fisher's χ^2 tended to be conservative under moderate and large sample size conditions.

Results showed that Goodman's U performed poorly when used with kernel-smoothed non-parametrically graphical DIF procedures: The Type I error rates were conservative for most simulation conditions; the power rates were low across items and ability distribution conditions when sample size was small. Even when the sample size was large, the power rates for Goodman's U using KS were still much lower than those using NC and RC. One possible reason is that Goodman's U used error variance instead of weighted error variance, as used in Cochran's Z and Fisher's χ^2 . these result suggested that some DIF statistics are not suitable to be used with non-parametrically graphical DIF procedure.

Implications for Practice

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.

Limitations of this Study and Directions for Future Research

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.

Among the three non-parametric statistics considered in this study, the performance of only one statistic, the Cochran's Z, was significantly improved by the kernel smoothed procedure in testing non-parametrically graphical DIF analysis. The second direction for future research therefore is related to the modification of Fisher's χ^2 and Goodman's U statistics to improve their performance in testing non-parametrically graphical DIF. For example, the performance of Goodman's U may be improved if the weighted error variance instead of error variance is used.

References

Altman, N. (1992). An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician, 46,* 175-185.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for Educational and Psychological Testing.* Washington, DC: American Educational Research Association.

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. (1996). Theory and applications of nonparametric regression in item response theory. *Dissertation Abstracts International: Section B: The Sciences and Engineering. 57(4-B),* pp.2652.

Douglas, J. A. (1997). Joint consistency of nonparametric item characteristic curve and ability estimation. *Psychometrika, 62,* 7-28.

Douglas, J. A., & Cohen, A. (2001). Nonparametric item response function estimation for assessing parametric model fit. *Applied Psychological Measurement, 25,* 234-243.

Douglas, J. A., Roussos, L. A., & Stout, W. (1996). Item-bundle DIF hypothesis testing: Identifying suspect bundles and assessing their differential functioning. *Journal of Educational Measurement, 33(4),* 465-484.

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.

Härtle, W. (1990). *Applied nonparametric regression.* Cambridge: Cambridge University Press.

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.

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

Kanji, G. K. (1993). *100 statistical tests.* Thousand Oaks, CA: Sage.

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

Marascuilo, L.A., & Slaughter, R. E. (1981). Statistical procedures for identifying possible sources of item bias based on Â2 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.

Meijer, R. R. & Baneke, J. J. (2004). Analyzing psychopathology items: A case for nonparametric item response theory modeling. *Psychological Methods, 9,* 354-368.

Narayanan, P., & Swaminathan, H. (1994). Performance of Mantel-Haenszel and simultaneous item bias procedure for detecting differential item functioning. *Applied Psychological Measurement, 18,* 315-328.

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.

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

Rogers, W. T., & Klinger, D. A. (2007). Purposes of an issues with the provincial testing programs in Alberta. Paper presented at the annual meeting of the National Council of Measurement in Education, Chicago, IL.

Roussos, L. A., & Stout, W. F. (1996). Simulation studies of the effects of small sample size and studied item parameters on SIBTEST and Mantel-Haenszel type I error performance. *Journal of Educational Measurement, 33,* 215-230.

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.

Thissen, D. & Wainer, H. (1982) Some standard errors in item response theory. *Psychometrika, 47,* 397-412