

INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

**ProQuest Information and Learning
300 North Zeeb Road, Ann Arbor, MI 48106-1346 USA
800-521-0600**

UMI[®]

University of Alberta

**The Effect of Large Ability Differences on Type I Error and Power Rates using
SIBTEST and TESTGRAF DIF Detection Procedures**

by

Andrea Julie Gotzmann



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

Department of Educational Psychology

Edmonton, Alberta

Fall 2001



**National Library
of Canada**

**Acquisitions and
Bibliographic Services**

**395 Wellington Street
Ottawa ON K1A 0N4
Canada**

**Bibliothèque nationale
du Canada**

**Acquisitions et
services bibliographiques**

**395, rue Wellington
Ottawa ON K1A 0N4
Canada**

Your file Votre référence

Our file Notre référence

The author has granted a non-exclusive licence allowing the National Library of Canada to reproduce, loan, distribute or sell copies of this thesis in microform, paper or electronic formats.

The author retains ownership of the copyright in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author's permission.

L'auteur a accordé une licence non exclusive permettant à la Bibliothèque nationale du Canada de reproduire, prêter, distribuer ou vendre des copies de cette thèse sous la forme de microfiche/film, de reproduction sur papier ou sur format électronique.

L'auteur conserve la propriété du droit d'auteur qui protège cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

0-612-69454-2

Canada

University of Alberta

Library Release Form

Name of Author: Andrea Julie Gotzmann

Title of Thesis: The Effect of Large Ability Differences on Type I Error and Power Rates using SIBTEST and TESTGRAF DIF Detection Procedures

Degree: Master of Education

Year this Degree Granted: 2001

Permission is hereby granted to the University of Alberta Library to reproduce single copies of this thesis and to lend or sell such copies for private, scholarly or scientific research purposes only.

The author reserves all other publication and other rights in association with the copyright in the thesis, and except as herein provided, neither the thesis nor any substantial portion thereof may be printed or otherwise reproduced in any material form whatever without the author's prior written permission.



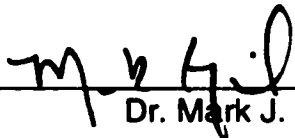
Andrea Gotzmann
3512 139 Avenue
Edmonton, Alberta
T5Y 2J4

July 23, 2001

University of Alberta

Faculty of Graduate Studies and Research

The undersigned certify that they have read and recommended to the Faculty of Graduate Studies and Research for acceptance, a thesis entitled **The Effect of Large Ability Differences on Type I Error and Power Rates using SIBTEST and TESTGRAF DIF Detection Procedures** submitted by **Andrea Julie Gotzmann** in partial fulfillment of the requirements for the degree of Master of Education.



Dr. Mark J. Gierl



Dr. W. Todd Rogers



Dr. Connie K. Varnhagen

July 20, 2001

Abstract

A simulation study was conducted to examine the effect of large ability differences using two differential item functioning (DIF) detection procedures, SIBTEST and TESTGRAF. DIF items are hard to identify when group ability differences are large (Gotzmann, Vandenberghe, & Gierl, 2000; Hambleton & Rogers, 1989). This problem was investigated in the current study for the two DIF detection procedures considered. Four ability differences (0.0, -1.0, -1.5, -2.0) and eight sample sizes (500/500, 750/1000, 1000/1000, 750/1500, 1000/1500, 1500/1500, 1000/2000, 2000/2000) were manipulated in a simulation study. Type I error and power rates were computed. The SIBTEST Type I error rates were inflated at the larger ability differences. Conversely, the TESTGRAF Type I error rates remained low for most ability differences and sample sizes. The SIBTEST power rates remained high, even with larger ability differences. The TESTGRAF power rates dropped as ability differences were introduced.

Acknowledgements

I would like to thank my committee members Dr. Mark J. Gierl, Dr. W. Todd Rogers, and Dr. Connie K. Varnhagen for their support and guidance throughout my program. Connie for suggesting Measurement as a field of study and Mark and Todd for their constant support throughout the last two years. I would also like to thank the students at CRAME for their support and friendship in and out of class. At times when things got very stressful I always got a kind word. I would also like to thank my entire family for your belief in me and sustained support through my university programs. Thanks for babysitting Adina and motivating me to achieve everything I dreamed possible.

Table of Contents

PURPOSE	3
ETHNIC DIF RESEARCH	3
Empirical Studies	3
Simulation Studies	6
MULTIDIMENSIONAL DIF FRAMEWORK	7
TECHNICAL OVERVIEW OF SIBTEST AND TESTGRAF DIF DETECTION	
PROCEDURES	9
SIBTEST	9
TESTGRAF	11
METHOD	15
Type I Error Study	16
Power Study	17
Statistical Analyses	18
RESULTS	19
Type I Error Study	19
Power Study	23
Overall Summary	25
CONCLUSIONS AND DISCUSSION	27
Limitations	30

Future Directions	31
REFERENCES.....	33

List of Tables

Table 1. Summary of Simulation studies examining ability differences	38
Table 2. Item parameters for the valid subtest Type I error study	39
Table 3. Item parameters for the DIF items for the power study	41
Table 4. Type I error rates for SIBTEST and TESTGRAF $\theta_R = 0.0 \theta_F = 0.0$ target ability differences	42
Table 5. Type I error rates for SIBTEST and TESTGRAF $\theta_R = 0.0 \theta_F = -1.0$ target ability differences	43
Table 6. Type I error rates for SIBTEST and TESTGRAF $\theta_R = 0.0 \theta_F = -1.5$ target ability differences	44
Table 7. Type I error rates for SIBTEST and TESTGRAF $\theta_R = 0.0 \theta_F = -2.0$ target ability differences	45
Table 8. Power rates for SIBTEST and TESTGRAF $\theta_R = 0.0 \theta_F = 0.0$ target ability differences	46
Table 9. Power rates for SIBTEST and TESTGRAF $\theta_R = 0.0 \theta_F = -1.0$ target ability differences	47
Table 10. Power rates for SIBTEST and TESTGRAF $\theta_R = 0.0 \theta_F = -1.5$ target ability differences	48
Table 11. Power rates for SIBTEST and TESTGRAF $\theta_R = 0.0 \theta_F = -2.0$ target ability differences	49

List of Figures

- Figure 1.** Type I error and power rates for $\theta_R = 0.0$, $\theta_F = 0.0$ and $\theta_R = 0.0$, $\theta_F = -1.0$ difference using effect size measures for SIBTEST and TESTGRAF... 50
- Figure 2.** Type I error and power rates for $\theta_R = 0.0$, $\theta_F = -1.5$ and $\theta_R = 0.0$, $\theta_F = -2.0$ difference using effect size measures for SIBTEST and TESTGRAF... 51
- Figure 3.** Type I error and power rates for $\theta_R = 0.0$, $\theta_F = 0.0$ and $\theta_R = 0.0$, $\theta_F = -1.0$ difference using the combined effect size measure and B statistic..... 52
- Figure 4.** Type I error and power rates for $\theta_R = 0.0$, $\theta_F = -1.5$ and $\theta_R = 0.0$, $\theta_F = -2.0$ difference using the combined effect size measure and B statistic..... 53

**The Effects of Large Ability Differences on Type I Error and Power Rates using
the SIBTEST and TESTGRAF DIF Detection Procedures**

Educational practitioners and test developers often find large test scores differences when comparing examinees with diverse ethnic backgrounds (Berends & Koretz, 1996; Cameron, 1990; Freedle & Kostin, 1990; Scheuneman & Grima, 1997; Schmitt & Dorans, 1990). Reducing these differences is one goal in the educational reform movement (Barron & Koretz, 1996). These large test score differences are particularly noteworthy when Native and non-Native examinees are compared (Alberta Education, 1996; Gotzmann, Vandenberghe, & Gierl, 2000; Hambleton & Rogers, 1989; Vandenberghe & Gierl, 2001). Socioeconomic and cultural differences may contribute to these performance differences (Common & Frost, 1989; Hull, 1990; Trent & Gilman, 1985; Wood & Clay, 1996). However, few researchers have studied item-level outcomes which may explain why Native examinees score lower than non-Native examinees (Gotzmann et al., 2000; Hambleton & Rogers, 1989). Native examinee scores may be biased due to factors in test development. For example, Janzen (2000) and Krywaniuk and Das (1976) found that Native children are more likely to use simultaneous processing skills and non-Native children are more likely to use successive processing skills. If exams have a small number of items that illicit simultaneous processing skills, then these exams may put Native examinees at a disadvantage. Therefore, assessment of bias at the item level, and its contribution to the total test score differences, should be studied.

Item bias can be estimated with different methods. Traditionally, item-level differences between groups have been assessed by comparing the proportion correct for each group (Lord, 1980). However, this method has one major flaw. The proportion correct method compares all examinees, regardless of ability level. Thus, the proportion correct is dependent upon the sample of examinees (see Camilli & Shepard, 1994). To overcome this problem, statistical methods can be used to determine whether differential item functioning (DIF) is present. DIF occurs when examinees from different groups have a different probability of answering the item correctly, after controlling for overall ability. In these comparisons, the majority group is called the reference group and the minority group is called the focal group. DIF methods are used to estimate bias by matching examinees on an external measure of ability or overall test score performance and comparing these examinees at the item level. This approach removes total test score differences in the estimation process, which provides a stronger measure of the actual group differences on the item.

There are many statistical procedures to estimate DIF including Item Response Theory (IRT) area measures (Lord, 1980; Thissen, Steinberg, & Wainer, 1988), Mantel-Haenszel (Holland & Thayer, 1988), Logistic Regression (Swaminathan & Rogers, 1990), Simultaneous Item Bias Test (SIBTEST; Shealy & Stout, 1993), and TESTGRAF (Ramsay, 1991, 2000). Most of these procedures have been used to identify DIF between ethnic groups. However, only two of the procedures may be suitable when large ability differences are found. These procedures are SIBTEST and TESTGRAF. Further, both of these

DIF detection procedures can be used with small sample sizes and both yield comparable DIF measures (Ramsay, 1991; 2000; Shealy & Stout, 1993).

However, these procedures also have a noteworthy difference. SIBTEST uses a regression correction to estimate true scores when ability differences occur.

TESTGRAF, on the other hand, uses kernel smoothing to match examinees on their raw scores.

Purpose

What is unknown is the extent to which SIBTEST and TESTGRAF would yield the same results in the presence of large ability differences. Consequently, the purpose of the study was to evaluate, using data simulation procedures, the effects of large ability differences on Type I error and power rates using SIBTEST and TESTGRAF.

To begin, ethnic DIF research is reviewed. The multidimensional DIF framework is discussed next, followed by a technical overview of the SIBTEST and TESTGRAF procedures.

Ethnic DIF Research

Empirical Studies

Some researchers have assessed moderate ability differences, up to 1 standard deviation, using different DIF detection procedures. For example, African American and Hispanic examinees have been compared to White examinees on various tests (e.g., Dorans, Schmitt, & Bleistein, 1992; Parshall, & Miller, 1995; Pike, 1989; Schmitt, 1988; Shepard, Camilli, & Williams, 1985; Zwick & Ercikan, 1989). Gotzmann, Vandenberghe and Gierl (2000) and

Hambleton and Rogers (1989) found large ability differences, up to 1.7 standard deviations, between Native and non-Native examinees. In both of these studies, the agreement across the procedures was evaluated (i.e., IRT area measure and the Mantel-Haenszel procedure for the Hambleton & Rogers, 1989 study and SIBTEST and TESTGRAF procedures for the Gotzmann et al., 2000 study).

Hambleton and Rogers (1989) used IRT area measures and the Mantel-Haenszel statistic (MH) to identify DIF items when Anglo American and Native American examinees were compared. They used a cross-validation design. The 2000 respondents in the total sample were randomly divided in half. The DIF analyses were then conducted in each half sample and the results compared. The ability differences for these two groups were approximately 1.6 standard deviations. They found 61% of DIF items identified in Sample 1 using the IRT area method were flagged in Sample 2, while 47% of the items identified using the MH method in Sample 1 were identified in Sample 2. Further, 56%, and 64% of the DIF items identified in Sample 2 were flagged in Sample 1 using, respectively, the IRT area measure and MH method.

In addition to this cross-validation design, Hambleton and Rogers (1989) formed two matched-groups in which examinees were matched on total test scores. The total number of examinees in each of the matched samples was 650. Consistency in flagging DIF items for the matched-group comparison was comparable to that found in the cross-validation study. Hambleton and Rogers (1989) interpreted "this moderate level of consistency [as] somewhat surprising, considering that all results were based on 1,000 examinees in each group, and

disturbing in view of the fact that, in most situations, the practitioner would not have the luxury of a cross—validation sample” (p. 324). Large ability differences in this study may have contributed to different DIF detection results.

Unfortunately the reliability of either procedure in detecting true DIF items is unknown since real data were used. Thus, the validity the IRT area measure and MH DIF detection procedure in correctly identifying DIF items is questionable.

Gotzmann et al. (2000) used the SIBTEST and TESTGRAF procedures to assess DIF between Native and non-Native examinees in two grade levels and in two subject areas. The sample sizes for the Native examinees ranged from 637 to 971. The non-Native examinees sample sizes were fixed at 2000. The ability differences ranged from 1.3 to 1.7 standard deviations between the two groups. They used effect sizes ($\hat{\beta}_U, \hat{\beta}_F$) to identify DIF items. They used correlations between the effect sizes to determine consistency between the two procedures. The effect size measures for SIBTEST and TESTGRAF are on the same scale and correlations preserve the ranks of the effect size measures. They found that the consistency in flagging DIF items between procedures was, at best, moderate: the correlations between the effect size measures yielded by the two procedures ranged from 0.61 to 0.77. Gotzmann et al. (2000) concluded that, “Since there was not any clear consistency between the two procedures, Type I error seems to be of paramount concern for users of the programs. Consequently, it is difficult to determine which items truly displayed DIF” (p. 12).

However, for both the Hambleton and Rogers (1989) and Gotzmann et al. (2000) DIF detection studies, matching examinees was problematic. The test

score distributions for the reference and focal groups were markedly different. In addition, both studies used real data to compare the DIF detection procedures and therefore the researchers could not determine which procedure was more valid in the sense that the true DIF items were identified.

Simulation Studies

Alternatively, simulation studies can be used to identify which DIF detection procedure is suitable with large ability differences. In a simulation study, variables are systematically manipulated. Data are generated for each condition and replicated many times for each condition. A statistic is calculated for each replication and averaged across the replications for each condition. In simulation studies, the accuracy of DIF detection procedures is assessed frequently by comparing empirical Type I error rates and power with their corresponding nominal values (e.g., Roussos & Stout, 1996b; Shealy & Stout, 1993). Type I error is the probability of falsely rejecting a true null hypothesis. Power is the probability of identifying a true alternative hypothesis.

Simulation studies have been used to assess small to moderate ability differences between the reference and the focal group with various DIF detection procedures (e.g., Chang, Mazzeo, & Roussos, 1996; Clauser, Mazor, & Hambleton, 1994; Jiang & Stout, 1998; Mazor, Clauser, & Hambleton, 1992; Narayanan & Swaminathan, 1994; Oshima, & Miller, 1992; Roussos & Stout, 1996b; Shealy & Stout, 1993; Zwick, Thayer, & Mazzeo, 1997). These studies are summarized in Table 1. As can be seen in Table 1, small ability differences, up to .5 standard deviation, have been considered. Generally, these differences

did not affect the accuracy of the DIF detection procedures (Oshima & Miller, 1992). Larger ability differences, up to 1 standard deviation, have been considered. In these cases, the Type I error rates increased modestly as ability differences increased (Clauser et al., 1994) while power decreased slightly with small sample sizes (Mazor et al., 1992). SIBTEST, one of the DIF detection procedures considered in the present study, has been thoroughly evaluated with small ability differences by Roussos and Stout (1996b) and Shealy and Stout (1993). They found low Type I error rates and moderate to high power rates when ability differences up to 1 standard deviation were present.

Overall, small to moderate ability differences do not seem to have an important differential effect on DIF detection procedures. However, large ability differences, such as those that occur when Native and non-Native examinees are compared, have not been empirically evaluated in a simulation study. Since test developers must address this problem, the empirical study of DIF procedures under the condition of large ability differences and using simulation procedures is clearly needed.

Multidimensional DIF Framework

The presence of DIF suggests that a multidimensional framework is needed instead of a unidimensional framework. The multidimensional DIF framework is an extension of the unidimensional three-parameter logistic model outlined by Lord (1980). The unidimensional three-parameter logistic model is given by,

$$P(\theta)_i = c_i + \frac{1 - c_i}{1 + e^{-1.7a_i(\theta - b_i)}}$$

where $P(\theta)_i$ is the probability of a correct response to item i , θ is the latent ability measured, and a_i is the discrimination, b_i is the difficulty, and c_i is the pseudo-guessing parameter of item i (Lord, 1980). The $P(\theta)_i$ values are calculated across the θ scale to produce an item characteristic curve (ICC). The multidimensional DIF framework is an extension of this model and conceptually separates group differences on the primary and secondary dimensions.

Multidimensionality is a recognized and generally accepted cause of DIF (i.e., Berk, 1982; Jensen, 1980; Lord, 1980; Messick, 1989; Scheuneman, 1982; as cited in Roussos & Stout, 1996a). Multidimensionality has also been used to model DIF in simulated and real data analyses (Ackerman, 1992; Bolt & Stout, 1996; Oshima and Miller, 1992; Oshima, Raju, & Flowers, 1997; Roussos & Stout, 1996a; Shealy & Stout, 1993). Within this framework, DIF items measure at least one dimension in addition to the primary dimension. The primary dimension, also called the target ability, is the construct that the item is intended to measure. The secondary dimension, also called the nuisance ability, is the construct that the item is not intended to measure (Roussos & Stout, 1996b). Shealy and Stout (1993) operationalized this framework by extending the unidimensional framework so that it included a term for the possible presence of the nuisance variable. The formula for the three-parameter multidimensional IRT model becomes:

$$P(\theta, \eta)_i = c_i + \frac{1 - c_i}{1 + \exp(-1.7(a_{i\theta}(\theta - b_{i\theta}) + a_{i\eta}(\eta - b_{i\eta})))}, i = n + 1, \dots, N,$$

where $P(\theta, \eta)_i$ is the probability of a correct response to item i , θ is the target ability, η is the nuisance ability, and a_i , b_i , and c_i are defined as before.

Technical Overview of SIBTEST and TESTGRAF DIF Detection Procedures

SIBTEST

The Simultaneous Item Bias Test (SIBTEST) is a nonparametric, model-based procedure. It provides an effect size measure and a test of significance.

The SIBTEST effect size measure, $\hat{\beta}_U$ is estimated by

$$\hat{\beta}_U = \sum_{k=0}^n \hat{p}_k (\bar{Y}^*_{Rk} - \bar{Y}^*_{Fk}),$$

where \hat{p}_k is the proportion of focal examinees at each score point k , and \bar{Y}^*_{Rk} is the estimated true score for the reference group and \bar{Y}^*_{Fk} is the estimated true score for the focal group at each score point k . The estimated true scores are produced using a regression correction described by Shealy and Stout (1993). The regression correction will be described in the next section. If the estimated effect size, $\hat{\beta}_U$ is positive, then the item favors the reference group. In contrast, if $\hat{\beta}_U$ is negative, then the item favors the focal group.

SIBTEST yields an overall statistical test of the hypothesis $H_0 : B = 0$ and $H_1 : B \neq 0$. The test statistic \hat{B} is given by

$$\hat{B} = \frac{\hat{\beta}_U}{\hat{\sigma}(\hat{\beta}_U)},$$

where $\hat{\sigma}(\hat{\beta}_U)$ is given by

$$\hat{\sigma}(\hat{\beta}_U) = \left(\sum_{k=0}^n \hat{p}_k \left(\frac{1}{J_{Rk}} \hat{\sigma}^2(Y|k, R) + \frac{1}{J_{Fk}} \hat{\sigma}^2(Y|k, F) \right) \right)^{1/2},$$

where \hat{p}_k is defined as before, J_{Rk} and J_{Fk} are the numbers of examinees with k correct on the valid subtest for, respectively, the reference and focal groups, and $\hat{\sigma}^2(Y|k, R)$ and $\hat{\sigma}^2(Y|k, F)$ are the sample variances of the studied subtest scores for examinees with the valid subtest score k for, respectively, the reference and the focal groups (Shealy & Stout, 1993). Shealy and Stout (1993) demonstrated that \hat{B} is standard normal with a mean 0 and variance 1 under the null hypothesis of no DIF. If \hat{B} exceeds the $|100(1 - \alpha/2)|$ percentile point in the unit normal distribution, then $H_0 : B = 0$ is rejected in favor of $H_1 : B \neq 0$.

To further identify DIF by size, the hypothesis test is used in conjunction with the effect size measure to control Type I error rates. Roussos and Stout (1996b) adopted guidelines based on research conducted at the Educational Testing Service (Zieky, 1993, p.342; Zwick & Ercikan, 1989). The guidelines proposed by Roussos and Stout (1996b) are: (a) negligible or A-level DIF: absolute value of $\hat{\beta}_U < 0.059$ and $H_0 : B = 0$ is rejected, (b) moderate or B-level DIF: absolute value of $0.059 \leq \hat{\beta}_U < 0.088$ and $H_0 : B = 0$ is rejected, and (c) large or C-level DIF: absolute value of $\hat{\beta}_U \geq 0.088$ and $H_0 : B = 0$ is rejected. These guidelines were used in the current study to identify DIF items.

Regression correction. The regression correction is used to transform each raw score estimate to its true score estimate. The transformation is used to correct for measurement error (Shealy & Stout, 1993). The correction uses the

classical true score model, $X = \tau + \varepsilon$, where X is the observed score, τ is the true score, and ε is the random error (Crocker & Algina, 1986). Each true score is estimated from the observed score using a linear regression transformation where the slope of the regression equation is the reliability of the modified test formed by deleting the item being evaluated for DIF from the total number of test items. Thus, $\bar{Y}_{Rk}^* - \bar{Y}_{Fk}^*$, as given in the SIBTEST equation, is an estimate of the difference in studied subtest true scores for the two groups when the examinees are matched on overall ability.

TESTGRAF

TESTGRAF is a non-parametric statistical procedure that uses kernel smoothing to estimate item characteristic curves (ICCs). TESTGRAF uses kernel smoothing and the concept of local averaging to estimate $P(\theta)_i$ (Ramsay, 2000). The estimation procedure, itself, has four sequential steps. First, the examinees are ranked according to an estimate of ability, which is total test score. Second, the ranks are replaced by quantiles of the standard normal distribution. These quantiles are calculated by dividing the area under the standard normal density function into $N + 1$ equal areas of size $1/(N + 1)$. The quantiles are used as the latent ability values θ_a , $a = 1, \dots, N$. Third, the examinees are sorted by response patterns, denoted y_{ia} , and by estimated latent ability values, θ_a . The indicator items in each scalar, y_{ia} , take on the value of 1 if examinee a chooses the correct option, or 0 if examinee a chooses the incorrect option. Fourth, $P(\theta)_i$ is

estimated by smoothing the relationship between the indicator scalar y_{ia} and the ability value $\theta_1, \dots, \theta_N$ close to the evaluation point, denoted θ_q .

TESTGRAF uses the *Gaussian kernel* smoothing technique (Ramsay, 2000). The probability of $P(\theta)_i$ is a weighted average of the values of y_{ia} for examinees with θ values close to θ_q . The number of evaluation points, θ_q , $q = 1, \dots, Q$ can be set as high as 101 but the default for TESTGRAF is 51. To estimate $P(\theta)_i$, TESTGRAF uses the equation

$$P(\theta)_{iq} = \sum_{a=1}^N w_{aq} y_{ia},$$

where y_{ia} is the indicator scalar and w_{aq} is the weight used at the particular score point. The weights are computed from

$$w_{aq} = \frac{K[(\theta_a - \theta_q)/h]}{\sum_{b=1}^N K[(\theta_b - \theta_q)/h]},$$

where K is the kernel smoothing function defined by an exponential function, θ_a is ability estimate for examinee a , θ_b is the ability estimate for examinee b , and h is the bandwidth parameter set at $h = N^{-1/5}$.

TESTGRAF uses local averaging by computing the average of the indicator values y_{ia} for values of θ_a falling within the limits $(\theta_{q-1} + \theta_q)/2$ and $(\theta_q + \theta_{q+1})/2$ (between the centers of the adjacent intervals, $[\theta_{q-1}, \theta_q]$ and $[\theta_q, \theta_{q+1}]$). For the smallest value of y_{ia} , the average is taken for all values below the centre of the first interval. For the largest value of y_{ia} , the average is taken

for all values above the centre of the last interval. These averages are denoted Q and are indicated by p_{iq} . The area under the standard normal curve between these interval centers is also computed and denoted by ϕ_q . ϕ_q is then smoothed by the equation

$$P(\theta)_{iq} = \sum_{r=1}^Q w_{rq} p_{iq},$$

where p_{iq} is the estimated average y_{ia} value, and

$$w_{rq} = \frac{\phi_r K[(\theta_r - \theta_q) / h]}{\sum_{s=1}^Q K[(\theta_s - \theta_q) / h]}$$

is the weight of the values in interval r for θ_q , r is the interval of values that are close to q , ϕ_r is the area under the curve between the two intervals, θ_r is value of the center of the r interval, and θ_q is the index of the summed r intervals.

This process of local averaging is used to compute the item characteristic curves. It is also used to estimate DIF between two or more groups.

Simulation studies have been conducted on the accuracy of item parameter estimation process used in TESTGRAF. Ramsay (1991), for example, found that the estimated ICCs, as computed with TESTGRAF, and true ICCs, as simulated from known parameters, were very similar using the root-mean-square-error as the criterion measure. Patsula and Gessaroli (1995) conducted a simulation study designed to compare the item parameter estimates and item characteristic curves produced by TESTGRAF with corresponding item parameter estimated ICCs produced by BILOG, a well known IRT calibration

program (Mislevy & Bock, 1990). They found that TESTGRAF was more accurate at low and middle ability levels and slightly less accurate at high ability levels compared to BILOG. These findings strongly suggest that TESTGRAF yields accurate item parameter estimates. This outcome also suggests that the TESTGRAF DIF detection procedure should yield accurate results, although this outcome has not been empirically evaluated.

The DIF detection procedure within TESTGRAF compares the ICCs for reference and focal group examinees. The estimate of DIF for TESTGRAF is denoted, $\hat{\beta}_F$, and is given by:

$$\hat{\beta}_F = \sum_{q=1}^Q p_{Fq} [P^{(F)}(\theta)_i - P^{(R)}(\theta)_i],$$

where p_{Fq} is the proportion of the focal group displaying value θ_q , and $P^{(R)}(\theta)_i$ and $P^{(F)}(\theta)_i$ are the item characteristic curve estimates for the reference and focal groups for item i , respectively. TESTGRAF uses the actual differences in the item characteristic curves rather than a regressed true score estimate for the difference between the two groups. Also, the TESTGRAF procedure does not have a statistical test. Further, the direction of the statistic is different from the SIBTEST procedure: negative $\hat{\beta}_F$ values favor the reference group and positive $\hat{\beta}_F$ values favor the focal group. However, the effect size portion of the guidelines proposed by Roussos and Stout (1996b) for the SIBTEST procedure are used with the TESTGRAF procedure in this study to identify DIF items: (a) negligible or A-level DIF: absolute value of $\hat{\beta}_F < 0.059$, (b) moderate or B-level

DIF: absolute value of $0.059 \leq \hat{\beta}_F < 0.088$, and (c) large or C-level DIF: absolute value of $\hat{\beta}_F \geq 0.088$.

In summary, Type I error and power rates have not been evaluated for SIBTEST and TESTGRAF in a common simulation study in which large ability differences between the reference and focal groups are produced. The SIBTEST DIF detection procedure has been evaluated with ability differences up to 1 standard deviation and should produce more accurate results with larger ability differences than TESTGRAF due to the regression correction it includes. The TESTGRAF DIF detection procedure has not been evaluated in a simulation study. The similarity of the effect size measures between the two procedures and the use of regression correction in SIBTEST without a comparable correction in TESTGRAF provides two points of comparison for comparatively evaluating these procedures when ability differences are large.

Method

A simulation study was conducted to compare Type I error and power rates when target ability differences and sample size were manipulated. The DIFSIM program created in the Stout Research Lab was used to create the data. DIFSIM generates dichotomous item response scalars for two distinct samples of examinees based on the Shealy and Stout (1993) and Roussos and Stout (1996a) latent multidimensional IRT model. DIFSIM generates two types of items: target ability items that are the non-DIF items and secondary ability items that are the DIF items. The non-DIF items were generated using the unidimensional three-parameter logistic model (i.e., $\eta = 0$). DIF items were

generated using the multidimensional three-parameter logistic model (i.e., $\eta \neq 0$). DIFSIM generates the examinee abilities θ and η from a bivariate normal distribution for each secondary ability $\eta_1, \eta_2, \eta_3, \eta_4$. Each secondary ability is created for a specific direction and magnitude for separate DIF items, with means $(\mu_\theta, \mu_{\eta_1}, \mu_{\eta_2}, \mu_{\eta_3}, \mu_{\eta_4})$, standard deviations $(\sigma_\theta, \sigma_{\eta_1}, \sigma_{\eta_2}, \sigma_{\eta_3}, \sigma_{\eta_4})$, and correlations $(\rho_{\theta\eta_1}, \rho_{\theta\eta_2}, \rho_{\theta\eta_3}, \rho_{\theta\eta_4})$ for both the reference and focal groups. In the present study, the means for the target ability were set to 0.0 for the reference group. The means for the target ability were set at 0.0, -1.0, -1.5, -2.0 for the focal group. The standard deviations were set to 1. Correlations, for both the reference and focal groups, were set at 0.5. These values were used in previous studies to simulate DIF in a multidimensional context (Nandakumar, 1993; Shealy and Stout, 1993). One hundred replications were generated for each condition, and Type I error and power rates were calculated.

Type I Error Study

The Type I error study contained a valid subtest with 50 non-DIF items using parameters from the SAT-Verbal subtest (Drasgow, 1987). The a -, b -, and c - parameters are reported in Table 2. The reference group target ability remained constant at $\theta_R = 0.0$, but the focal group target ability θ_F varied ($\theta_F = 0.0, -1.0, -1.5, \text{ and } -2.0$). The range of values of θ_F were chosen to allow an assessment of both SIBTEST and TESTGRAF under conditions of no ability differences through large ability differences actually found in real data studies (i.e., Gotzmann et al., 2000; Hambleton & Rogers, 1989; Vandenberghe & Gierl,

2001). The sample sizes for the reference and focal groups were 500/500, 750/1000, 1000/1000, 750/1500, 1000/1500, 1500/1500, 1000/2000, and 2000/2000. These sample size conditions were chosen because they are frequently found in actual testing situations. Sample size and target ability differences were fully crossed in a 4 (Ability differences) X 8 (Sample size) design resulting in 32 conditions for the Type I error study.

Power Study

The power study contained a valid subtest consisting of the first 38 items in the Type I error study test (see Table 2). The remaining 12 items were DIF items with the item parameters reported in Table 3. The nuisance dimensions were set at a specific value resulting in average effect size values according to the A-, B-, and C- DIF level guidelines outlined by Roussos and Stout (1996b; Nandakumar, 1993). The first nuisance dimension was simulated to create three B- level items favoring the reference group and, where the ability difference was set at 0.6, and the second nuisance dimension three B-level items favoring the focal group, where the ability difference was set at 0.6. The third nuisance dimension contained three C-level items favoring the reference group, where the ability difference was set at 0.8, and the fourth nuisance dimension contained three C-level items favoring the focal group, where the ability difference was set at 0.8. The sample size conditions and target ability differences remained unchanged from the Type I error study. Items meeting the A-level criteria were considered non-DIF items whereas items meeting the B- or C- level criteria were considered DIF items. This interpretation seems justified since B- and C- level

criteria are often considered for potential item bias in test reviews (e.g., Zieky, 1993).

Statistical Analyses

The SIBTEST and TESTGRAF DIF detection procedures were used for the analyses. Items were flagged as non-DIF if they met the A-level criteria and DIF if they met the B- or C- level criteria. The effect size measure, null hypothesis test, and combined measures were used to classify DIF items for both procedures in the present study. An alpha level of .05 was used for all hypothesis testing. Type I error rates were calculated by taking the average for the simulated non-DIF items flagged as either B- or C- level across replications. Type I error rates were considered liberal if above the .05 level and conservative if below the .05 level. Power rates were calculated by taking the average for the simulated DIF items flagged as at least B- level across replications. Cohen (1992) interpreted power rates as excellent if above 0.80, and poor below 0.80. This criterion was used in this study but an additional distinction of moderate power rates for values between 0.80 and 0.70 was also used. For SIBTEST, the $\hat{\beta}_U$ effect size measure, B statistic (null hypothesis test of $H_0 : B = 0$, where $p < .05$), and combined use of the $\hat{\beta}_U$ effect size measure and B statistic were used for the Type I error and power analyses. The effect size measure is the magnitude of the difference between the reference and focal groups. The combined use of the $\hat{\beta}_U$ effect size measure and B statistic was used to distinguish statistical significance from practical significance. The combined effect size measure and null hypothesis test was flagged as DIF when the effect size measure and null

hypothesis were both flagged as DIF. For TESTGRAF, the effect size measure, denoted $\hat{\beta}_F$, was used for the Type I error and power analyses. Type I error rates for $\hat{\beta}_U$, the B statistic, and the combined use of the $\hat{\beta}_U$ and the B statistic were compared to $\hat{\beta}_F$ at each target ability difference for the Type I error study. Power rates for $\hat{\beta}_U$, the B statistic, and the combined use of the $\hat{\beta}_U$ and the B statistic were compared to the $\hat{\beta}_F$ at each target ability difference for the power study.

Results

Type I Error Study

The results for the Type I error study for each ability difference condition are presented in Tables 4 to 7. For the $\theta_R = 0.0$ and $\theta_F = 0.0$ (i.e., no ability difference), the empirical Type I error rates for $\hat{\beta}_U$ were, with one exception (500/500), considerably less the nominal .05 level (see Table 4). In contrast, the empirical Type I error rates for the B statistic, which ranged from .05 to .06, were very close to the nominal level. Again, with the exception of the 500/500 sample size condition, the empirical Type I error rates for the combined use of the $\hat{\beta}_U$ and the B statistic were considerably less than .05. Likewise, with the exception of the 500/500 sample size condition, the empirical Type I error rates for $\hat{\beta}_F$ were considerably below .05. Thus, with the exception of the B statistic for SIBTEST, the DIF identification procedure for SIBTEST and for TESTGRAF generally yielded very conservative Type I error rates.

For the $\theta_R = 0.0$ and $\theta_F = -1.0$, the empirical Type I error rates for $\hat{\beta}_U$ for the 500/500 sample size condition was twice the nominal level (see Table 5). When the sample sizes were 750/1000, the empirical and nominal Type I error rates were comparable. For the remaining sample size conditions the empirical Type I error rates were somewhat less than the nominal level, with the discrepancy between the empirical and nominal alphas generally increasing as the sample sizes increased. In contrast, the empirical Type I error rates for the B statistic, which ranged from .07 to .12, exceeded the nominal level. Further, the discrepancy between the empirical and nominal rates increased as the sample size increased to the point that for the six largest sample size conditions, the empirical rates were approximately twice the nominal level. The empirical Type I error rates for the combined use of the $\hat{\beta}_U$ and the B statistic were, with the exception of the smallest sample size condition, quite comparable to the empirical rates of $\hat{\beta}_U$ alone. Further, the empirical Type I error rates for $\hat{\beta}_F$ followed a similar pattern, although they were approximately half the size. Thus, with the exception of the B statistic for SIBTEST, the DIF identification procedure for SIBTEST and for TESTGRAF generally yielded very conservative Type I error rates with sample sizes greater than or equal to 750/1000. In contrast, the B statistic yielded liberal Type I error rates for all of the sample sizes considered.

For the $\theta_R = 0.0$ and $\theta_F = -1.5$, the empirical Type I error rate for $\hat{\beta}_U$ was four times the nominal level for the 500/500 sample size condition and approximately three times the nominal level for the 750/1000 sample size

condition (see Table 6). The empirical Type I error rates were approximately twice the nominal level for the remaining sample size conditions, with the exception of the 2000/2000 sample size condition, which was slightly above the nominal level. The empirical Type I error rates for the B statistic likewise exceeded the nominal rate. For the sample size combinations 500/500, 750/1500, 1000/2000, and 2000/2000 sample size conditions, the empirical rate was approximately twice the nominal rate. For the sample size combinations 750/1000, 1000/1000, and 1000/1500 sample size conditions, the empirical Type I error rates were approximately three times the nominal rate. Further, for the 1500/1500 sample size condition, the empirical Type I error rate was four times the nominal rate. The empirical Type I error rates for the combined use of the $\hat{\beta}_U$ and the B statistic were again liberal for the first six sample size conditions shown in Table 6, ranging from 0.09 (1500/1500) to 0.12 (750/1000). The empirical Type I error for the 1000/2000 sample size condition was close to the nominal level, while for the 2000/2000 sample size condition, was somewhat conservative. The empirical Type I error rates for $\hat{\beta}_F$ were somewhat lower than the nominal rate for all sample size conditions, with the exception of the 500/500 and 1000/1000 sample size conditions, which were approximately twice the nominal level. Thus, SIBTEST generally yielded moderately liberal to very liberal empirical Type I error rates, with the exception of the combined use of the $\hat{\beta}_U$ and the B statistic with the two largest sample sizes (1000/2000, 2000/2000). The TESTGRAF procedure generally yielded conservative Type I error rates for

the remaining sample size conditions, with the exception of the 500/500 and 1000/1000 sample size conditions.

For the $\theta_R = 0.0$ and $\theta_F = -2.0$, the empirical Type I error rates for $\hat{\beta}_U$ were considerably higher than the nominal level (see Table 7). They were approximately four times the nominal level for the 1000/1500, 1500/1500, 1000/2000, and 2000/2000 sample size conditions, five times the nominal level for the 750/1000, 1000/1000 and 750/1500 sample size conditions, and approximately six times the nominal level for the 500/500 sample size condition. The empirical Type I error rates for B statistic were approximately twice the nominal rate, with the exception of the 1500/1500 and 2000/2000 sample size condition, which were triple the nominal level. The empirical Type I error rates for the combined use of the $\hat{\beta}_U$ and the B statistic were approximately twice the nominal level for all sample size conditions. Lastly, the empirical Type I error rate for $\hat{\beta}_F$ were approximately twice the nominal level for 500/500 sample size condition, equal to the nominal level for the 750/1000 and 1000/1000 sample size conditions, and less than the nominal rate for the remaining five sample size conditions. Thus, the SIBTEST procedure generally yielded inflated Type I error rates for all sample size conditions. The TESTGRAF procedure generally yielded conservative Type I error rates for the larger sample size conditions (i.e., greater than 750/1500, approximately at the 750/1000 and 1000/1000 sample size conditions).

To summarize, the Type I error rates for TESTGRAF and SIBTEST were comparable and generally conservative when there was no difference in ability

and up to 1 standard deviation difference between the reference and the focal groups. However, beginning with the case in which there was a 1.5 standard deviation difference in ability between the two groups, TESTGRAF Type I error rates tended to be very conservative while the SIBTEST rates tended to be liberal. This pattern increased, especially as the largest ability differences were introduced (i.e., -2.0). However, the Type I error rates for the combined effect size measure and B statistic were somewhat comparable to the $\hat{\beta}_F$ rates with small to moderate target ability differences (0.0, -1.0).

Power Study

The results for the power study for each target ability difference are presented in Tables 8 to 11. When $\theta_R = 0.0$ and $\theta_F = 0.0$, the power for all of the procedures exceeded the minimal power of .80 suggested by Cohen (1992; see Table 8). The power ranged from: .87 to .96 for $\hat{\beta}_U$, .90 to 1.00 for the B statistic, and .86 to .95 for the combined use of the $\hat{\beta}_U$ and the B statistic, and .83 to .93 for $\hat{\beta}_F$. Generally, the power for the B statistic was the highest, followed by $\hat{\beta}_U$ and the combined use of the $\hat{\beta}_U$ and the B statistic conditions, which were quite comparable, and $\hat{\beta}_F$, which yielded the lowest values. However, in all cases the power for each DIF detection procedure was considered excellent for all sample sizes conditions.

For the $\theta_R = 0.0$ and $\theta_F = -1.0$, the power for the three SIBTEST DIF detection procedures exceeded the .80 criteria for all sample size conditions with one exception. The power for the 500/500 sample size condition was in the

moderate range (.70 to .80; see Table 9). In contrast, the power for $\hat{\beta}_F$ was less than .5 for all the sample size conditions. Thus, the SIBTEST procedure yielded excellent power, with the exception of the 500/500 sample size condition. The TESTGRAF procedure yielded poor power for all the sample size conditions.

For $\theta_R = 0.0$ and $\theta_F = -1.5$, power for $\hat{\beta}_U$ and the combined use of $\hat{\beta}_U$ and the B statistic were moderate and close to the .80 criteria for all sample size conditions, with the exception of the 500/500 sample size condition, in which the power rate was poor (see Table 10). The power for the B statistic was excellent for the sample size conditions 1000/1500, 1500/1500, 1000/2000, 2000/2000, in the moderate range for the 750/1000, 1000/1000, and 750/1500 sample size conditions, and in the poor range for the 500/500 sample size condition. Again, the power for $\hat{\beta}_F$ was poor for all conditions, with values between .17 and .28. Thus, the SIBTEST procedure yielded moderate power for all sample size conditions, with the exception of the 500/500 sample size condition, and excellent power for the B statistic power for larger sample sizes (i.e., greater than or equal to 1000/1500). The TESTGRAF procedure yielded poor power for all the sample size conditions.

For the $\theta_R = 0.0$ and $\theta_F = -2.0$, the power was poor for all the DIF detection procedures and sample sizes, with one exception (2000/2000 for the B statistic; see Table 11). Thus, for the largest ability difference, as for the previous values of θ_F , with one exception, both SIBTEST and TESTGRAF the power was less than adequate.

To summarize, the power rates for SIBTEST and TESTGRAF were very comparable when there were no target ability difference. The power rates for SIBTEST remained excellent for the middle target ability difference (-1.0) and poor for the larger target ability differences (-1.5, -2.0). The power rates for TESTGRAF dropped when ability differences were introduced and were considered poor for all ability differences greater than zero.

Overall Summary

The results from the Type I error and power studies reported in Tables 4 to 11 are displayed on a common graph in Figures 1 and 2 for both effect size measures, $\hat{\beta}_U$ and $\hat{\beta}_F$. The Type I error results are shown on the primary Y axis, and the power results are shown on the secondary Y axis, with sample size constant for both Y axes shown on the X axis. As shown in Figure 1A for no target ability differences, the Type I error rates for both SIBTEST and TESTGRAF were less than .05 and power for both procedures was excellent. For the target ability difference of -1.0, the Type I error rates for both SIBTEST and TESTGRAF remained low, with the exception of the 500/500 sample size condition (see Figure 1B). The power for SIBTEST was all above 0.80, but dropped dramatically for TESTGRAF. For the target ability difference of -1.5, the Type I error rates for SIBTEST were liberal, though only slightly at the largest sample size condition 2000/2000 (see Figure 2A). The Type I error rates were conservative for TESTGRAF, with the exception of the 500/500 and 1000/1000 sample size conditions. Power was considered moderate with the exception of the 500/500 sample size for SIBTEST, and poor for TESTGRAF across all

sample size conditions. For the target ability difference of -2.0 , the Type I error rates for the SIBTEST procedure were very liberal (see Figure 2B). The Type I error rates for TESTGRAF were above .05 for the smallest sample size condition 500/500, and slightly higher than .05 for the 750/1000, and 1000/1000 sample size conditions, and less than .05 for the 750/1500, 1000/1500, 1500/1500, 1000/2000, and 2000/2000 sample size conditions. Power for both DIF detection procedures was poor for all sample size conditions. Overall, the SIBTEST procedure produced liberal Type I error rates when ability differences increased, but also maintained moderate power for most ability difference conditions. The TESTGRAF procedure produced lower Type I error rates when larger ability differences occurred and also less power in detecting DIF.

The results from the Type I error and power studies reported in Tables 4 to 11 are displayed in Figures 3 and 4 for the combined use of the $\hat{\beta}_U$ and the B statistic for the SIBTEST procedure. These graphs are presented to indicate which sample sizes maintained a relatively small Type I error rate and still maintained high power. The TESTGRAF procedure was not included in these graphs as the power rates were poor for all ability differences. Moreover, the results from this study indicate the combined SIBTEST procedure may work better when large ability differences occur. For no target ability differences, the combined procedure maintained Type I error rates below the nominal level and excellent power for all sample size conditions (see Figure 3A). Therefore, when no ability differences are present all the sample sizes shown are suitable for the combined procedure. For -1.0 target ability differences, Type I error rates were

conservative and power were excellent for all sample size conditions, with the exception of the 500/500 sample size condition. In this condition, the Type I error rate was liberal and the power was poor (see Figure 3B). Larger sample sizes, greater than 500/500, maintained reasonable Type I error and power with middle ability differences for DIF analyses. For -1.5 target ability differences, Type I error rates were liberal for most of the sample sizes, with the exception of the two largest sample size conditions 1000/2000 and 2000/2000, and power was moderate for all sample size conditions, with the exception of the 500/500 sample size condition (see Figure 4A). For the -2.0 target ability difference condition, Type I error rates for all sample sizes were liberal, and the power for those same sample sizes was poor (see Figure 4B).

Conclusions and Discussion

The purpose of this research was to evaluate the effects of large ability differences on two DIF detection procedures, SIBTEST and TESTGRAF. Four ability differences (0.0, -1.0, -1.5, -2.0) and eight sample sizes (500/500, 750/1000, 1000/1000, 750/1500, 1000/1500, 1500/1500, 1000/2000, 2000/2000) were manipulated in this simulation study. The Type I error and power rates were compared for the SIBTEST effect size measure, the B statistic, and combined effect size measure and B statistic and the TESTGRAF effect size measure. Simulation studies are used to assess the accuracy of a DIF detection procedure by evaluating Type I error rates, which is falsely rejecting a true null hypothesis, and power rates, which is accepting a true alternative hypothesis when the true state is known.

The Type I error rates for the SIBTEST effect size measure were conservative until large ability differences were introduced (i.e., $\theta_F = -1.5$). The Type I error rates decreased for the combined effect size measure and B statistic when compared to the effect size measure alone, with larger target ability differences (i.e., $\theta_F = -1.5; -2.0$). The Type I error rates for TESTGRAF remained conservative for most sample sizes across the ability difference conditions. In summary, the SIBTEST Type I error rates were highly inflated at the larger ability differences and the TESTGRAF Type I error rates remained conservative for most ability differences and sample sizes.

Power was quite comparable for the TESTGRAF and SIBTEST procedures when no target ability differences were present. However, the SIBTEST procedure had much higher power than the TESTGRAF procedure when ability differences were introduced. The SIBTEST power remained high, even with larger ability differences, and the TESTGRAF power dropped when ability differences were introduced (i.e., $\theta_F = -1.0$).

The best outcome-meaning low or conservative Type I error and high power-was achieved for the SIBTEST procedure when both the effect size measure and statistical test were used. The Type I error rates for the SIBTEST combined effect size measure and null hypothesis were conservative for target ability differences up to -1.0 , and larger sample size conditions for ability differences of -1.5 . Power was high for target ability differences up to -1.0 , and poor for larger target ability differences of -1.5 and -2.0 . In summary, the Type I error rates for the combined effect size measure and the B statistic were

reduced for larger ability differences and the power was poor for these same conditions using the SIBTEST procedure. Combining the effect size measure with the null hypothesis test reduced Type I error rates, but maintained high power for the -1.0 ability difference sample size conditions and poor power for the -1.5 and -2.0 ability difference conditions (see Figure 4).

This study indicates that both procedures generally maintained low or conservative Type I error rates for ability differences up to -1.0. The TESTGRAF procedure had lower Type I error rates than the SIBTEST procedure when large ability differences were present. However, combining the effect size measure and null hypothesis test for SIBTEST produces similar Type I error rates when compared to TESTGRAF.

The power results indicate that the SIBTEST procedure with combined effect size and null hypothesis test generally produced good estimates of DIF for ability differences up to -1.0 for most sample sizes, and moderate to excellent power for samples sizes greater than 500 for ability differences of -1.5. Power was poor for the -2.0 ability difference condition, but power improved with larger sample sizes. The power results indicates that TESTGRAF produced good estimates of DIF when no ability differences were present and poor estimates when any ability differences were introduced. This outcome was likely due to the regression correction in the SIBTEST procedure. The TESTGRAF power might be improved by implementing a similar regression correction.

The combined effect size measure and null hypothesis test results suggest that the SIBTEST procedure is suitable for ability differences of -1.0 and

might be suitable when large samples are available for ability differences of -1.5. For ability differences as large as -2.0, no clear pattern was found making it difficult to recommend either DIF procedure in such a testing situation. Further study of the effect of larger sample sizes with ability differences of -1.5 and -2.0 to improve Type I error and power is suggested.

Limitations

There are at least two limitations to this study. First, the TESTGRAF procedure does not have a significance test or a regression correction so a direct comparison between SIBTEST and TESTGRAF is difficult, because the criteria for flagging DIF items are different. The SIBTEST procedure yielded three different criteria for flagging DIF and the TESTGRAF procedure had only one. However, the results reveal that both procedure produce relatively low Type I error rates across most conditions in this study. Further improvement to the power of TESTGRAF is needed. Perhaps TESTGRAF, with a regression corrected true score estimate, could produce similar power to the SIBTEST procedure. Regression correction improves the conditioning variable making it easier to match examinees. Improper matching may produce low DIF detection because the TESTGRAF effect size measure is not optimally estimated (i.e., it contains measurement error).

Second, the Type I error rates and power are *not* directly comparable. There is a trade off in conducting a study in which the Type I error calculations are based on data that do not contain DIF items. The resulting Type I error rates may be higher or lower than when DIF items are present. Analysis of the Type I

error rates in conjunction with power would be ideal. However due to the fluctuation of Type I error rates in this particular study, interpretation of power would not possible for many conditions. When a Type I error rate is not within a 95 % confidence interval the corresponding power is not interpretable. As a result, many of the power results would not be reported due to inflated Type I error rates (c.f. Ankenmann, Witt, & Dunbar, 1999, p. 289). This problem was overcome by conducting separate the Type I error and power studies.

Future Directions

There are at least three directions for future research resulting from this study. One direction is to evaluate Type I error in conjunction with power where loss of information would occur, as just described, but where the Type I error rates and power could be directly compared. Evaluation of Type I error rates with DIF items would indicate if the rates reported in this study were similar.

A second direction is to apply a regression correction to the TESTGRAF procedure and evaluate the results of the new statistic under the conditions simulated in this study. With regression correction the matching variable is improved which would likely increase power while maintaining low Type I error rates (Bolt & Gierl, in preparation). Although this work is very technical, it has the potential to dramatically improve the statistical properties of TESTGRAF.

A third direction is to use examinees in the reference and focal groups who have overlapping total test scores and reanalyze the data for DIF detection consistency as suggested by Hambleton and Rogers (1989). A simulation study to compare the effects of the overlapped sample versus using the entire sample

is currently being conducted by the author. This procedure may maintain lower Type I error rates regardless of ability difference and could be used by test developers to accurately identify DIF items. Again, matching of examinees is very important for all DIF detection procedures. Any method that can be used to better estimate the true DIF between groups should be studied and empirically evaluated.

In conclusion, this study has demonstrated that matching examinees is very important in detecting group differences. Results from the SIBTEST procedure indicate that the proper matching of examinees result in low Type I error rates and high power. However, the SIBTEST matching procedure for ability differences greater than -1.5 was not adequate. Hence, more research is needed to refine the SIBTEST matching procedure when large ability differences occur. Results from the TESTGRAF procedure indicate that poor matching of examinees will produce low Type I error rates and power. However, other DIF detection procedures, such as Logistic Regression, should be evaluated for large ability differences. The results from this study indicate that proper matching of examinees is needed when conducting DIF analyses. To date, the best method to overcome this problem has not been identified. Improved matching procedures may assist test developers in maintaining low Type I errors and increasing power and result in DIF detection procedures that can be reliably used to estimate DIF between groups with ability differences. Consequently, much more research is needed to address this practical testing problem.

References

Ackerman, T. A. (1992). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. Journal of Educational Measurement, 29, 67-91.

Alberta Education. (1996). Jurisdiction profile report: Northern lights school division #69. Edmonton, AB: Educational Information Exchange.

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

Barron, S. I., & Koretz, D. M. (1996). An evaluation of the robustness of the National Assessment of Educational Progress trend estimates for racial-ethnic subgroups. Educational Assessment, 3, 209-248.

Berends, M., & Koretz, D. M. (1996). Reporting minority students' test scores: How well can the National Assessment of Educational Progress account for differences in social context? Educational Assessment, 3, 249-285.

Berk, R. A. Editor. (1982). Handbook of methods for detecting test bias. Baltimore, MD: Johns Hopkins University Press.

Bolt, D., & Gierl, M. J. (in preparation). Application of a regression correction to three nonparametric tests of DIF: Implications for local and global DIF assessment. Unpublished manuscript.

Bolt, D., & Stout, W. (1996). Differential item functioning: Its multidimensional model and resulting SIBTEST detection procedure. Behaviormetrika, 23, 67-95.

Cameron, I. (1990). Student achievement among Native students in British Columbia. Canadian Journal of Native Education, 17, 36-43.

Camilli, G., & Shepard, L. A. (1994). Methods for identifying biased test items. Thousand Oaks, CA: Sage Publications.

Chang, H. H., Mazzeo, J., & Roussos, L. (1996). Detecting DIF for polytomously scored items: An adaptation of the SIBTEST procedure. Journal of Educational Measurement, 33, 333-353.

Clauser, B. E., Mazor, K. M., & Hambleton, R. K. (1994). The effects of score group width on the Mantel-Haenszel procedure. Journal of Educational Measurement, 31, 67-78.

- Cohen, J. (1992). A power primer. Psychological Bulletin, 112, 155-159.
- Common, R. W., & Frost, L. G. (1989). In search of equity education: A case study of native and non-native student progress in a public school system. Multiculturalism, 12, (3) 3-14.
- Crocker, L., & Algina, J. (1986). Introduction to classical and modern test theory. Orlando, FL: Harcourt Brace Jovanovich College Publishers.
- Dorans, N. J., Schmitt, A. P., & Bleistein, C. A. (1992). The standardization approach to assessing comprehensive differential item functioning. Journal of Educational Measurement, 29, 309-319.
- Drasgow, F. (1987). Study of the measurement bias of two standardized psychological tests. Journal of Applied Psychology, 72, 19-29.
- Freedle, R., & Kostin, I. (1990). Item difficulty of four verbal item types and an index of differential item functioning for black and white examinees. Journal of Educational Measurement, 27, 329-343.
- Gotzmann, A., Vandenberghe, C., & Gierl, M. (2000, May) Differential item functioning on Alberta achievement tests: A comparison of SIBTEST and TestGraf using data from native and non-native students. Poster presented at the Canadian Society for the Study of Education Annual Conference, Edmonton, AB.
- Hambleton, R. K., & Rogers, H. J. (1989). Detecting potentially biased test items: Comparison of IRT area and Mantel-Haenszel methods. Applied Measurement in Education, 2, 313-334.
- Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In Wainer, H. & Braun, H. I. (Eds.), Test validity (pp.129-145). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Hull, J. (1990). Socioeconomic status and native education in Canada. Canadian Journal of Native Education, 17, 1-14.
- Janzen, T. M. (2000). Assessment and remediation using the PASS theory with Canadian Natives. Unpublished doctoral dissertation, University of Alberta, Edmonton, AB.
- Jensen, A. R. (1980). Bias in mental testing. New York, NY: Macmillan Publishing Co.
- Jiang, H., & Stout, W. (1998). Improved type I error control and reduced estimation bias for DIF detection using SIBTEST. Journal of Educational and Behavioral Statistics, 23, 291-322.

Krywaniuk, L. W., & Das, J. P. (1976). Cognitive strategies in Native children: Analysis and intervention. The Alberta Journal of Educational Research, 22, 271-280.

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

Mazor, K. M., Clauser, B. E., & Hambleton, R. K. (1992). The effect of sample size on the functioning of the Mantel-Haenszel statistic. Educational and Psychological Measurement, 52, 443-451.

Messick, S. (1989). Validity. In Linn, R. L. Editor. Educational measurement. (3rd ed.). New York, NY: Macmillan Publishing Company.

Mislevy, R. J., & Bock, R. D. (1990). Item analysis and test scoring with binary logistic models. (2nd Edition). Mooresville, IN: Scientific Software Inc.

Nandakumar, R. (1993). Simultaneous DIF amplification and cancellation: Shealy-Stout's test for DIF. Journal of Educational Measurement, 30, 293-311.

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

Oshima, T. C., & Miller, M. D. (1992). Multidimensionality and item bias in item response theory. Applied Psychological Measurement, 16, 237-248.

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

Parshall, C. G., & Miller, T. R. (1995). Exact versus asymptotic Mantel-Haenszel DIF statistics: A comparison of performance under small-sample conditions. Journal of Educational Measurement, 32, 302-316.

Patsula, L. N., & Gessaroli, M. E. (April, 1995). A comparison of item parameter estimates and ICCs produced with TESTGRAF and BILOG under different test lengths and sample sizes. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

Pike, G. (1989). The performance of black and white students on the ACT-COMP exam: An analysis of differential item functioning using Samejima's graded model. Research Report 89-11. Knoxville, TN: The University of Tennessee.

Ramsay, J. O. (1991). Kernel smoothing approaches to nonparametric item characteristic curve estimation. Psychometrika, *56*, 611-630.

Ramsay, J. O. (2000). TESTGRAF: A program for the graphical analysis of multiple choice and questionnaire data. (Technical Manual). Montreal, PQ: McGill University, Department of Psychology.

Roussos, L., & Stout, W. (1996a). A multidimensionality-based DIF analysis paradigm. Applied Psychological Measurement, *20*, 355-371.

Roussos, L., & Stout, W. (1996b). 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.

Scheuneman, J. D. (1982). A posteriori analyses of biased items. In Berk, R. A. Editor. Handbook of methods for detecting test bias. Baltimore, MD: Johns Hopkins University Press.

Scheuneman, J. D., & Grima, A. (1997). Characteristics of quantitative word items associated with differential performance for female and black examinees. Applied Measurement in Education, *10*, 299-319.

Schmitt, A. P. (1988). Language and cultural characteristics that explain differential item functioning for Hispanic examinees on the Scholastic Aptitude Test. Journal of Educational Measurement, *25*, 1-13.

Schmitt, A. P., & Dorans, N. J. (1990). Differential item functioning for minority examinees on the SAT. Journal of Educational Measurement, *27*, 67-81.

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

Shepard, L. A., Camilli, G., & Williams, D. M. (1985). Validity of approximation techniques for detecting item bias. Journal of Educational Measurement, *22*, 77-105.

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. (1988). Use of item response theory in the study of group differences in trace lines. In Wainer, H. & Braun, H. I. (Eds.), Test validity (pp.147-169). Hillsdale, NJ: Lawrence Erlbaum Associates.

Trent, J. H., & Gilman, R. A. (1985). Math achievement of Native Americans in Nevada. Journal of American Indian Education, 24, 39-45.

Vandenberghe, C. N., & Gierl, M. J. (April, 2001). Differential bundle functioning on three achievement tests: A comparison of aboriginal and non-aboriginal examinees. Paper presented at the Annual meeting of the American Educational Research Association, Seattle, WA.

Wood, P. B., & Clay, W. C. (1996). Perceived structural barriers and academic performance among American Indian high school students. Youth and Society, 28, 40-61.

Zieky, M. (1993). Practical questions in the use of DIF statistics in test development. In Holland, P. W. & Wainer, H. (Eds.), Differential item functioning (pp.337-348). Hillsdale, NJ: Lawrence Erlbaum Associates.

Zwick, R., & Ercikan, K. (1989). Analysis of differential item functioning in the NAEP history assessment. Journal of Educational Measurement, 26, 55-66.

Zwick, R., Thayer, D. T., & Mazzeo, J. (1997). Descriptive and inferential procedures for assessing differential item functioning in polytomous items. Applied Measurement in Education, 10, 321-344.

Table 1

Summary of simulation studies examining ability differences

Study Authors	Size of ability difference	DIF procedure examined
1. Chang, Mazzeo, and Roussos (1996)	1 SD	Modified SIBTEST procedure, MH, Standardized mean difference (SMD)
2. Clauser, Mazor, and Hambleton (1994)	1 SD	MH
3. Jiang and Stout (1998)	0, 1 SD	New Regression corrected SIBTEST, SIBTEST
4. Mazor, Clauser and Hambleton (1992)	0, 1 SD	MH
5. Narayanan and Swaminathan (1994)	0, 0.5, 1 SD	MH, SIBTEST
6. Oshima and Miller (1992)	0, 0.5 SD	Four IRT areas measure methods
7. Roussos and Stout (1996b)	0, 0.5, 1 SD	MH, SIBTEST
8. Shealy and Stout (1993)	0, 0.5, 1 SD	MH, SIBTEST
9. Zwick, Thayer, & Mazzeo (1997)	0, 1 SD	SMD divided by standard error under the hypergeometric model, SMD divided by the standard error under the multinomial model, Mantel procedure, Standard SIBTEST, Modified SIBTEST

Table 2
Item parameters for the valid subtest Type I error study

Item	$a -$	$b -$	$c -$
1	1.20	1.40	0.11
2	1.40	1.60	0.11
3	0.90	0.80	0.20
4	2.00	1.40	0.11
5	1.50	2.00	0.06
6	0.50	-0.80	0.20
7	1.00	1.60	0.13
8	0.70	-1.00	0.20
9	0.50	0.40	0.20
10	1.10	1.40	0.04
11	0.90	-0.40	0.20
12	0.50	0.50	0.20
13	0.90	0.30	0.20
14	0.90	0.20	0.20
15	1.10	2.00	0.16
16	1.20	-0.30	0.20
17	1.70	1.30	0.17
18	1.20	0.50	0.20
19	0.90	0.50	0.14
20	0.70	-0.40	0.20
21	0.70	-0.60	0.20
22	1.20	-0.60	0.20
23	1.30	0.40	0.18
24	1.90	1.90	0.11
25	1.20	0.70	0.12

Table 2 Continued

Item	<i>a</i> –	<i>b</i> –	<i>c</i> –
26	1.00	1.50	0.11
27	0.70	-0.50	0.20
28	0.90	0.70	0.20
29	0.60	1.20	0.12
30	0.70	-0.50	0.20
31	0.70	-0.20	0.20
32	1.30	0.20	0.20
33	0.60	2.50	0.10
34	1.10	0.80	0.12
35	0.40	0.30	0.20
36	0.80	-0.70	0.20
37	1.50	1.70	0.09
38	1.00	1.70	0.08
39	1.10	-0.70	0.20
40	1.40	0.10	0.20
41	1.20	1.60	0.09
42	0.60	0.20	0.20
43	1.00	0.70	0.15
44	0.50	-0.60	0.20
45	0.90	1.60	0.11
46	1.10	1.20	0.05
47	0.70	0.50	0.20
48	1.20	-0.50	0.20
49	0.50	0.00	0.20
50	1.30	0.80	0.18

Table 3

Item parameters for the DIF items for the power study

Item	a_{θ}	b_{θ}	a_{η}	b_{η}	c_{θ}
1 ^a	1.00	0.00	0.80	0.00	0.20
2 ^a	1.30	0.00	0.75	0.00	0.20
3 ^a	2.00	0.00	1.00	0.00	0.20
4 ^b	0.80	-0.30	0.75	0.00	0.20
5 ^b	1.00	0.00	0.85	0.00	0.20
6 ^b	1.50	0.30	0.85	0.00	0.20
7 ^c	1.20	0.00	0.80	0.00	0.20
8 ^c	1.00	0.00	0.75	0.00	0.20
9 ^c	0.80	-0.30	0.65	0.00	0.20
10 ^d	1.00	0.00	0.65	0.00	0.20
11 ^d	1.30	0.30	0.75	0.00	0.20
12 ^d	1.00	0.00	0.80	0.00	0.20

Note: ^a B- level items favoring the Reference group; ^b B- level items favoring the Focal group; ^c C- level items favoring the Reference group; ^d C- level items favoring the Focal group.

Table 4

Type I error rates for SIBTEST and TESTGRAF $\theta_R = 0.0$ $\theta_F = 0.0$ Target Ability

Differences

Sample size	SIBTEST		TESTGRAF	
	$\hat{\beta}_U$ effect size	B statistic	$\hat{\beta}_U$ effect size and B statistic	$\hat{\beta}_F$ effect size
500/500	0.0376	0.0514	0.0348	0.0316
750/1000	0.0058	0.0494	0.0058	0.0056
1000/1000	0.0026	0.0558	0.0026	0.0042
750/1500	0.0022	0.0456	0.0022	0.0030
1000/1500	0.0006	0.0516	0.0006	0.0008
1500/1500	0.0002	0.0510	0.0002	0.0002
1000/2000	0.0014	0.0456	0.0014	0.0010
2000/2000	0.0000	0.0530	0.0000	0.0000

Table 5

Type I error rates for SIBTEST and TESTGRAF $\theta_R = 0.0$ $\theta_F = -1.0$ Target Ability

Differences

Sample size	SIBTEST		TESTGRAF	
	$\hat{\beta}_U$ effect size	B statistic	$\hat{\beta}_U$ effect size and B statistic	$\hat{\beta}_F$ effect size
500/500	0.1004	0.0734	0.0648	0.0666
750/1000	0.0430	0.0866	0.0420	0.0232
1000/1000	0.0294	0.0908	0.0288	0.0152
750/1500	0.0388	0.0908	0.0384	0.0138
1000/1500	0.0182	0.0920	0.0182	0.0108
1500/1500	0.0146	0.1086	0.0144	0.0074
1000/2000	0.0146	0.0914	0.0144	0.0054
2000/2000	0.0026	0.1218	0.0026	0.0034

Table 6

Type I error rates for SIBTEST and TESTGRAF $\theta_R = 0.0$ $\theta_F = -1.5$ Target Ability

Differences

Sample size	SIBTEST		TESTGRAF	
	$\hat{\beta}_U$ effect size	B statistic	$\hat{\beta}_U$ effect size and B statistic	$\hat{\beta}_F$ effect size
500/500	0.2000	0.1120	0.1046	0.0924
750/1000	0.1428	0.1468	0.1242	0.0408
1000/1000	0.1126	0.1560	0.1098	0.0814
750/1500	0.1264	0.1384	0.1126	0.0298
1000/1500	0.1100	0.1692	0.1078	0.0218
1500/1500	0.0914	0.2064	0.0896	0.0184
1000/2000	0.0962	0.0894	0.0450	0.0162
2000/2000	0.0624	0.1228	0.0310	0.0128

Table 7

Type I error rates for SIBTEST and TESTGRAF $\theta_R = 0.0$ $\theta_F = -2.0$ Target Ability

Differences

Sample size	SIBTEST		TESTGRAF	
	$\hat{\beta}_U$ effect size	B statistic	$\hat{\beta}_U$ effect size and B statistic	$\hat{\beta}_F$ effect size
500/500	0.3154	0.0892	0.0780	0.0994
750/1000	0.2694	0.1132	0.0952	0.0516
1000/1000	0.2538	0.1244	0.1068	0.0514
750/1500	0.2634	0.1018	0.0846	0.0368
1000/1500	0.2384	0.1326	0.1104	0.0350
1500/1500	0.2216	0.1486	0.1108	0.0326
1000/2000	0.2372	0.1246	0.1008	0.0266
2000/2000	0.2066	0.1590	0.1014	0.0272

Table 8

Power for SIBTEST and TESTGRAF $\theta_R = 0.0$ $\theta_F = 0.0$ Target Ability

Differences

	SIBTEST		TESTGRAF	
Sample size	$\hat{\beta}_U$ effect size	B statistic	$\hat{\beta}_U$ effect size and B statistic	$\hat{\beta}_F$ effect size
500/500	0.8700	0.9000	0.8617	0.8317
750/1000	0.9083	0.9758	0.8983	0.8783
1000/1000	0.9308	0.9917	0.9217	0.8842
750/1500	0.9192	0.9883	0.9108	0.8950
1000/1500	0.9300	0.9942	0.9200	0.8983
1500/1500	0.9367	1.0000	0.9267	0.9092
1000/2000	0.9475	0.9975	0.9383	0.9208
2000/2000	0.9567	1.0000	0.9467	0.9325

Table 9

Power for SIBTEST and TESTGRAF $\theta_R = 0.0$ $\theta_F = -1.0$ Target Ability

Differences

Sample size	SIBTEST		TESTGRAF	
	$\hat{\beta}_U$ effect size	B statistic	$\hat{\beta}_U$ effect size and B statistic	$\hat{\beta}_F$ effect size
500/500	0.7908	0.7358	0.7283	0.4867
750/1000	0.8442	0.9042	0.8342	0.4892
1000/1000	0.8592	0.9425	0.8500	0.4742
750/1500	0.8625	0.9375	0.8542	0.4525
1000/1500	0.8892	0.9733	0.8792	0.4667
1500/1500	0.8692	0.9858	0.8592	0.4325
1000/2000	0.8825	0.9792	0.8742	0.4675
2000/2000	0.8875	0.9950	0.8783	0.4408

Table 10

Power for SIBTEST and TESTGRAF $\theta_R = 0.0$ $\theta_F = -1.5$ Target Ability

Differences

Sample size	SIBTEST		TESTGRAF	
	$\hat{\beta}_U$ effect size	B statistic	$\hat{\beta}_U$ effect size and B statistic	$\hat{\beta}_F$ effect size
500/500	0.6717	0.5283	0.5200	0.2825
750/1000	0.7258	0.7258	0.7125	0.2367
1000/1000	0.7275	0.7750	0.7208	0.2175
750/1500	0.7450	0.7533	0.7258	0.2008
1000/1500	0.7708	0.8400	0.7642	0.2117
1500/1500	0.7608	0.8908	0.7517	0.1867
1000/2000	0.7658	0.8367	0.7558	0.1975
2000/2000	0.7683	0.9475	0.7608	0.1675

Table 11

Power for SIBTEST and TESTGRAF $\theta_R = 0.0$ $\theta_F = -2.0$ Target Ability

Differences

	SIBTEST		TESTGRAF	
Sample size	$\hat{\beta}_U$ effect size	B statistic	$\hat{\beta}_U$ effect size and B statistic	$\hat{\beta}_F$ effect size
500/500	0.5500	0.3633	0.3592	0.1642
750/1000	0.6017	0.5050	0.4983	0.1142
1000/1000	0.5917	0.5467	0.5400	0.1142
750/1500	0.6233	0.5425	0.5375	0.0992
1000/1500	0.6233	0.5967	0.5875	0.0875
1500/1500	0.6358	0.6875	0.6283	0.0925
1000/2000	0.6533	0.6217	0.6117	0.0917
2000/2000	0.6325	0.7542	0.6275	0.0817

Figure A. $\theta_R = 0.0, \theta_F = 0.0$

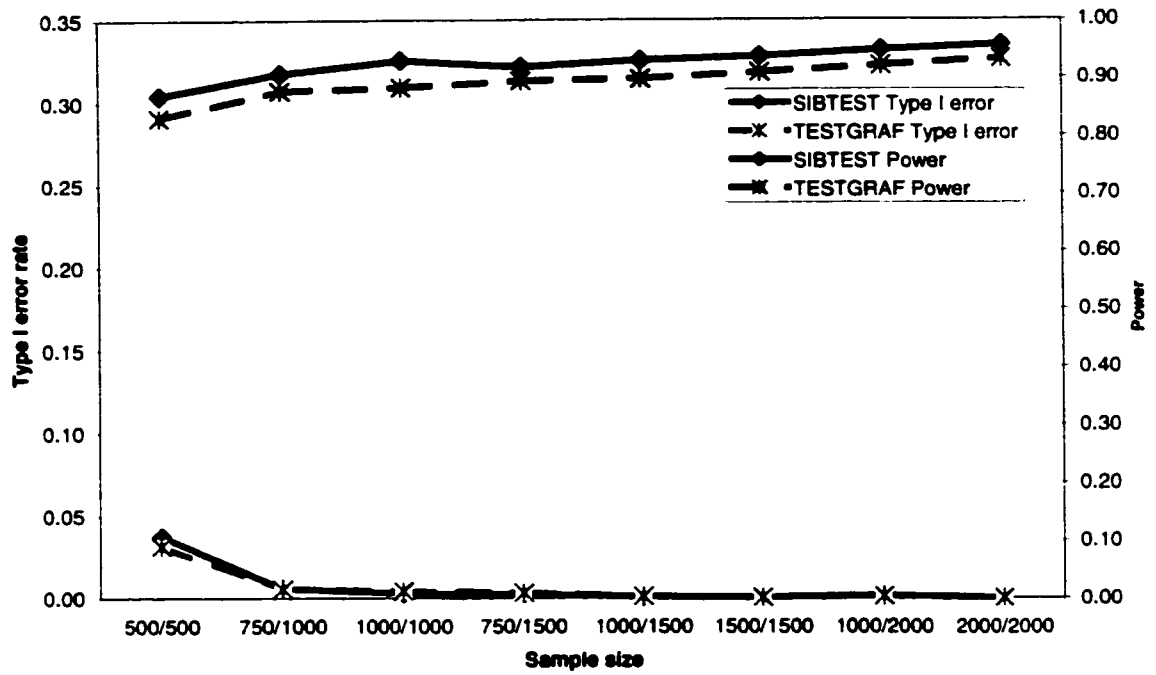


Figure B. $\theta_R = 0.0, \theta_F = -1.0$

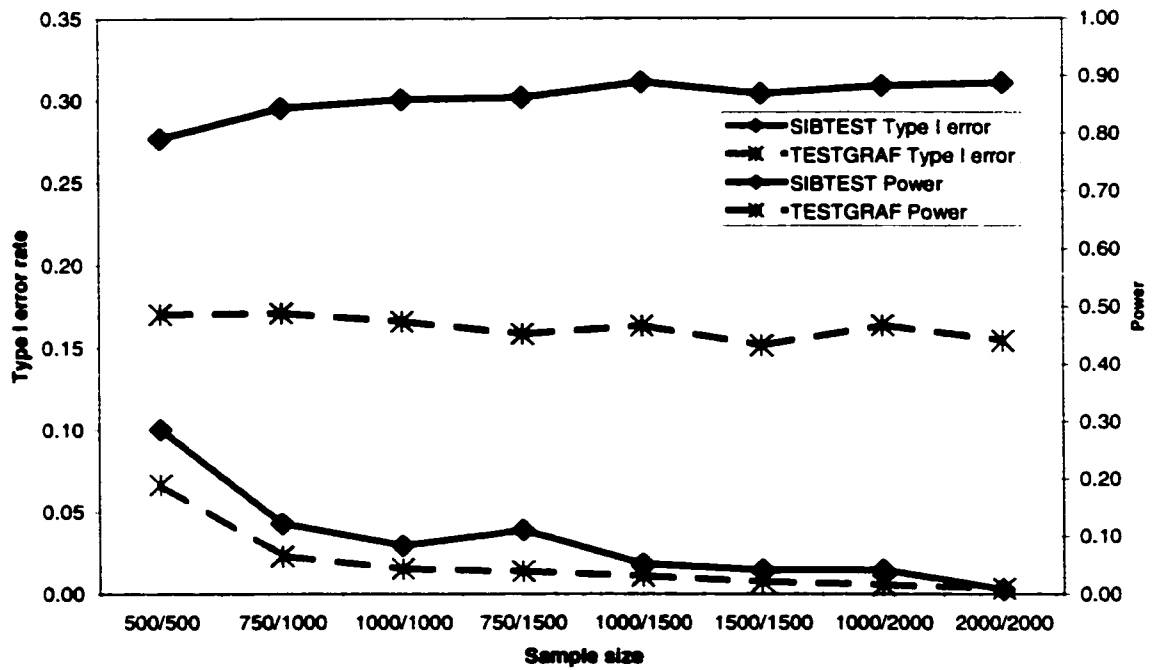


Figure 1.

Figure A. $\theta_R = 0.0, \theta_F = -1.5$

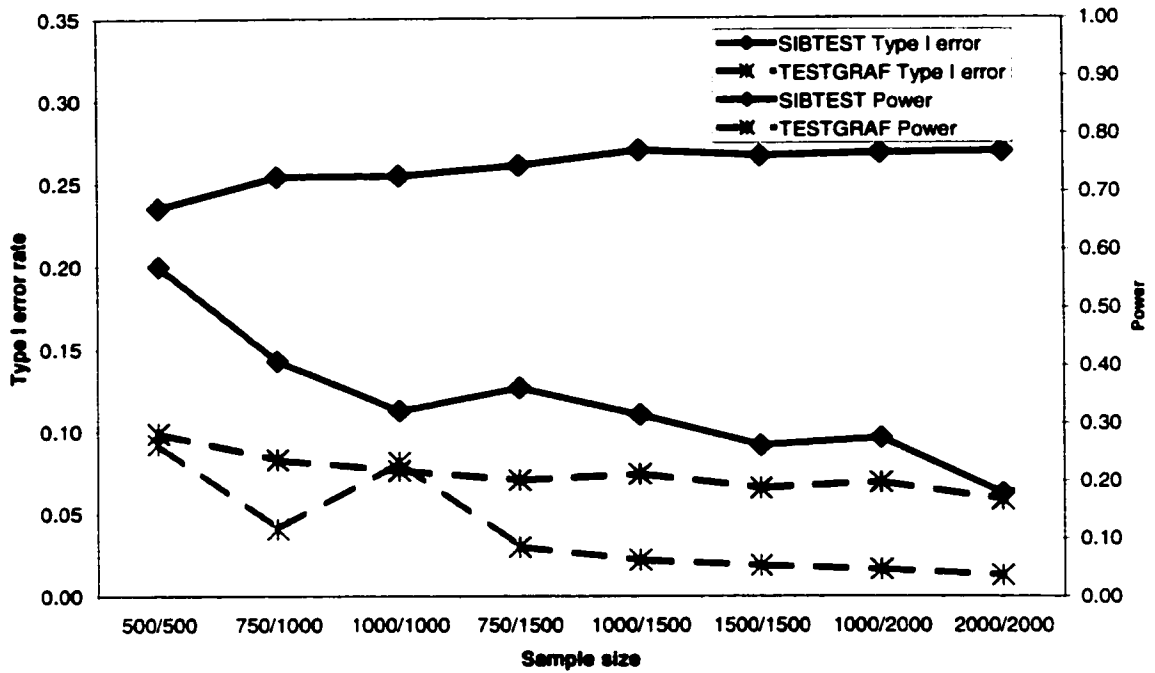


Figure B. $\theta_R = 0.0, \theta_F = -2.0$

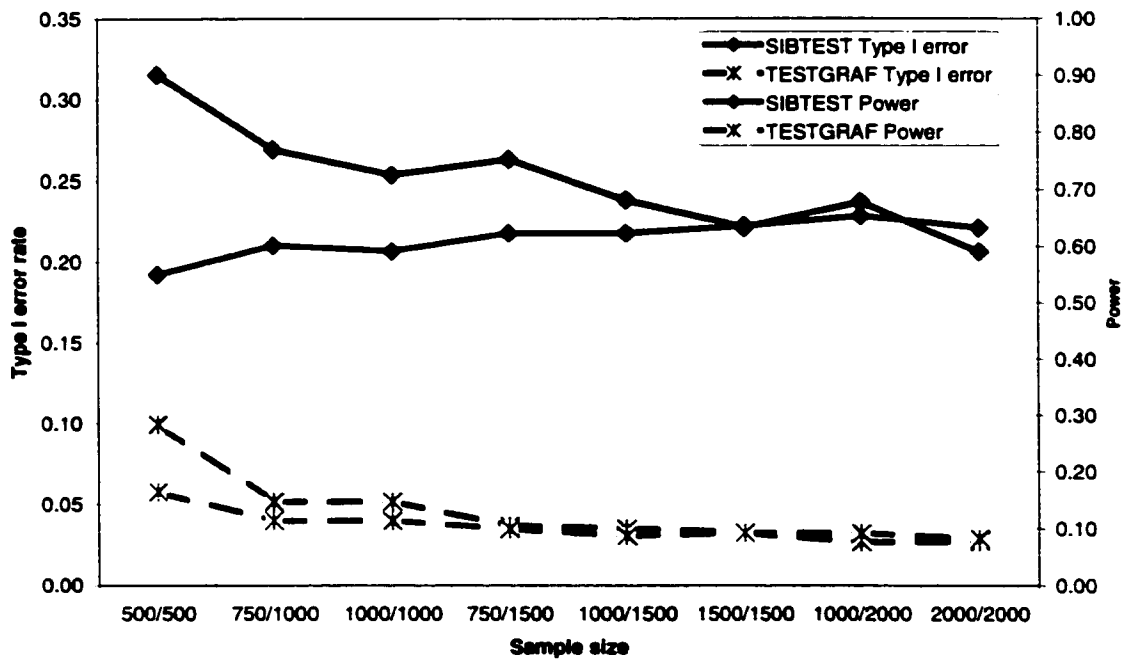


Figure 2.

Figure A. $\theta_R = 0.0, \theta_F = 0.0$

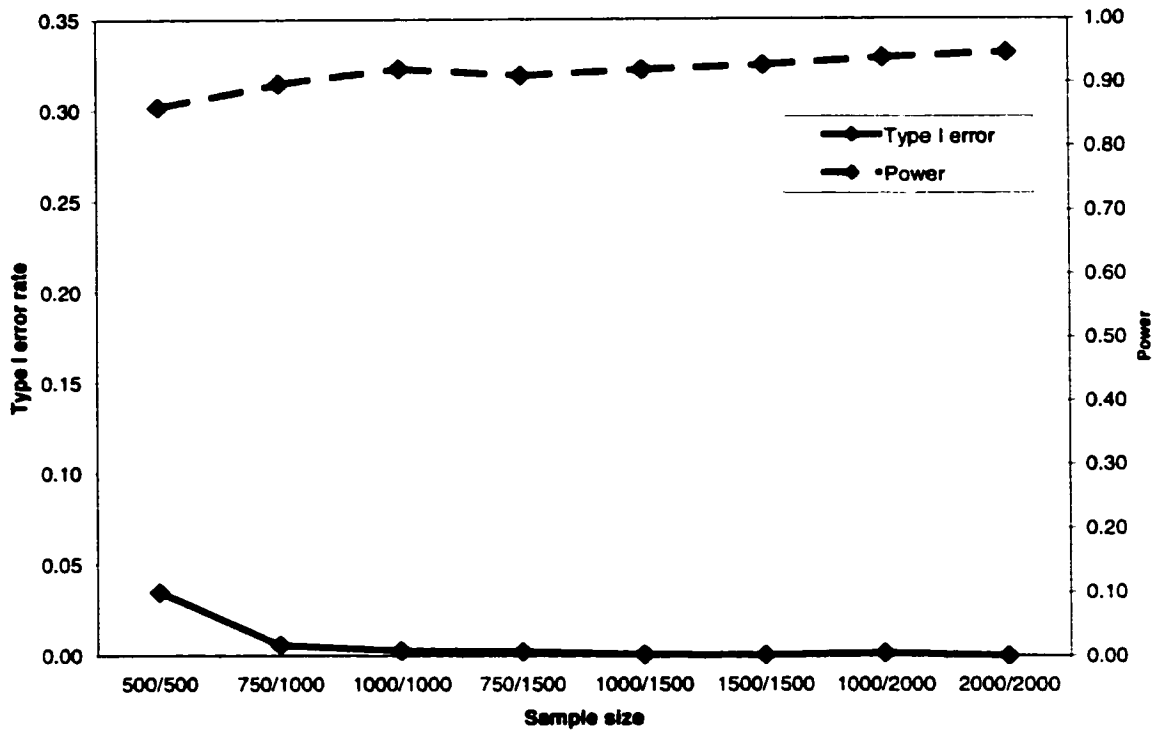


Figure B. $\theta_R = 0.0, \theta_F = -1.0$

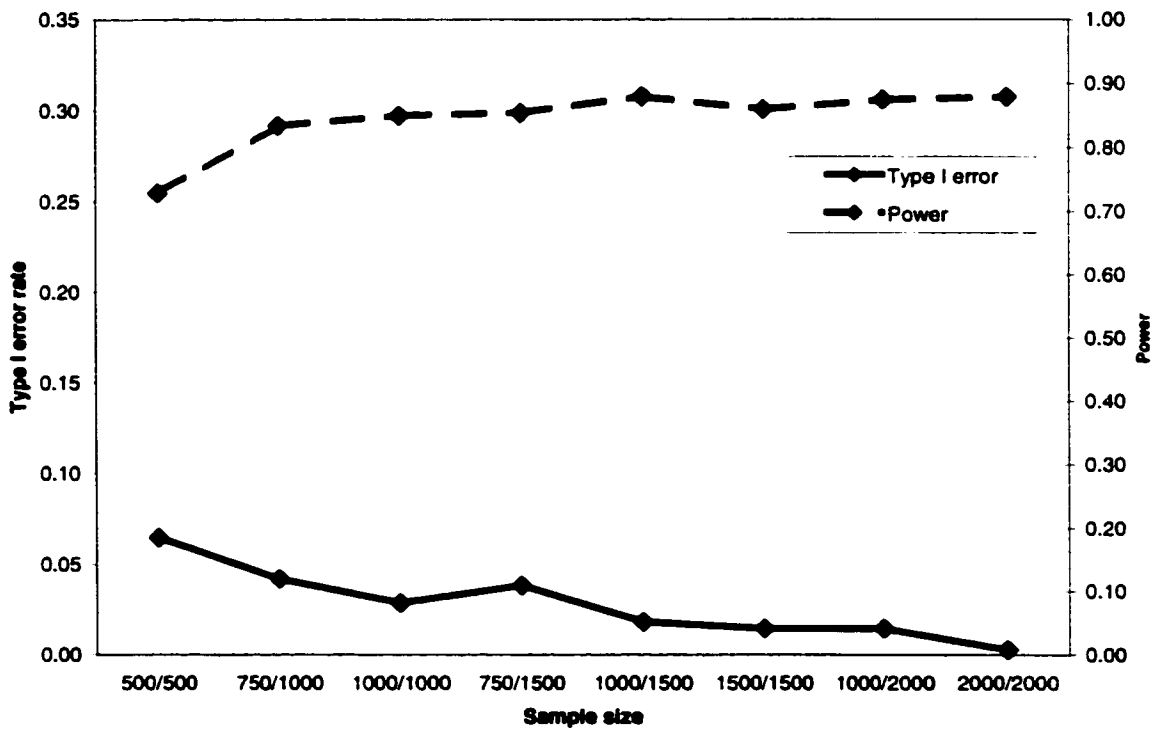


Figure 3.

Figure A. $\theta_R = 0.0, \theta_F = -1.5$

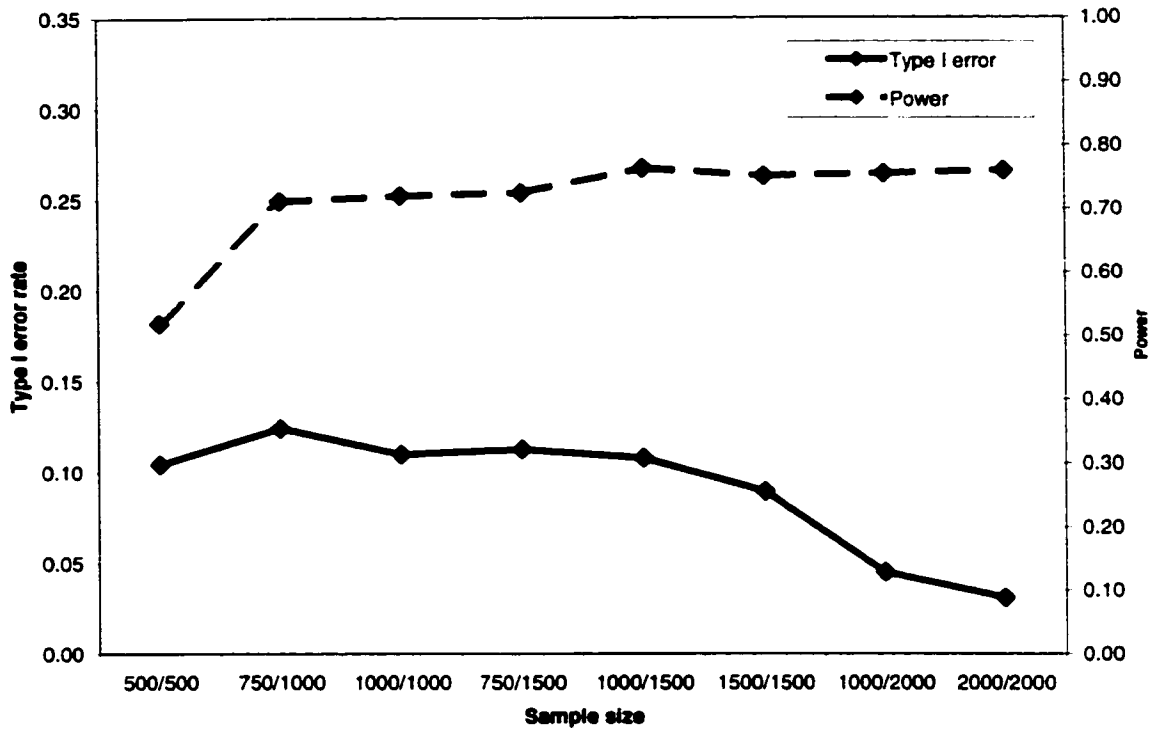


Figure B. $\theta_R = 0.0, \theta_F = -2.0$

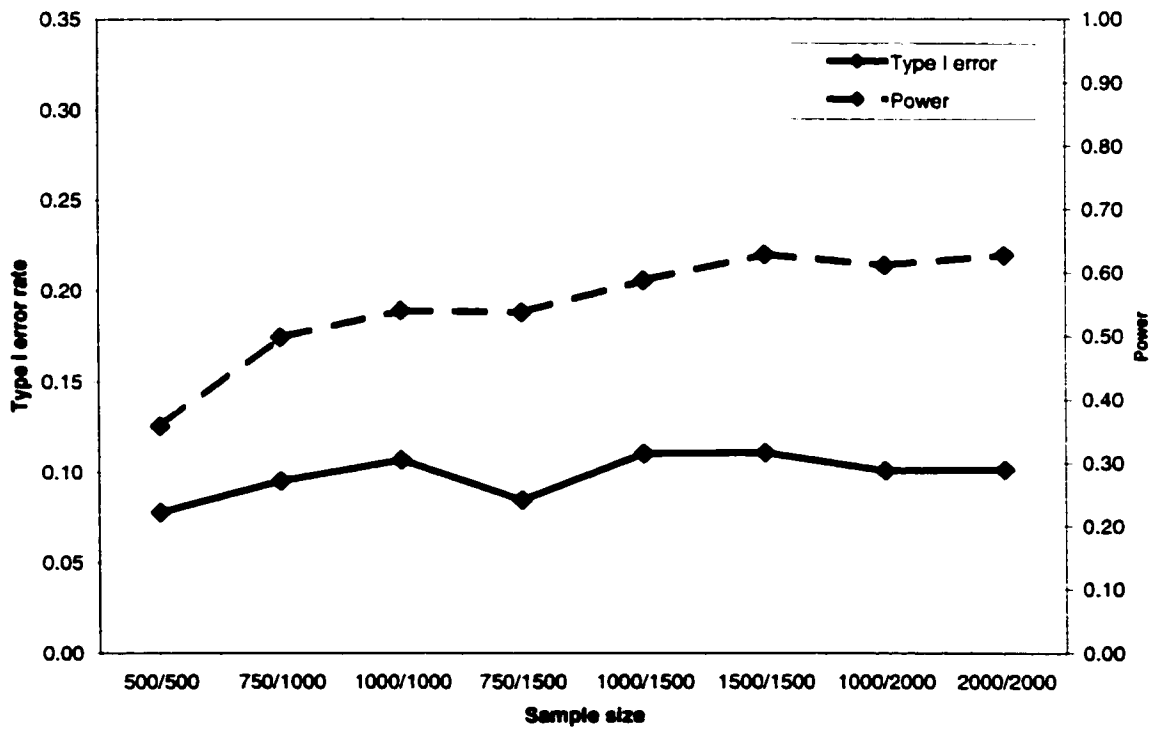


Figure 4.