University of Alberta

Evaluating the Performance of SIBTEST and MULTISIB for a Multidimensional Test

by

Jiawen Zhou ©

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
Fall, 2006

Your file Votre référence
ISBN: 978-0-494-22187-7
Our file Notre référence
ISBN: 978-0-494-22187-7

Abstract

Previous research on the matching criterion in differential item functioning (DIF) analyses emphasized that conditioning on multiple subtest scores reflecting the primary dimensions could raise the validity of the analysis for a multidimensional test. To assess this claim, the DIF detection performance of SIBTEST, a unidimensional procedure, and MULTISIB, a multidimensional procedure, were compared using simulated two-dimensional data of appropriate simple structure. The Type I error rates, power rates, and item correct classification results for both procedures were compared. It was found that SIBTEST, which utilized the number-correct score matching criterion, performed as well as, if not better than, MULTISIB that used a multidimensional matching criterion. The limitations of the current study are also discussed.

Acknowledgements

I would like to take this opportunity to acknowledge a number of people who have helped me through my masters program. Thanks are due first to my supervisor, Dr. Mark J. Gierl, for his great insights, perspectives, and guidance. As a great mentor, Mark provided supportive supervision in my study and the writing of this thesis. My sincere thanks go to Dr. W. Todd Rogers, who directed me to the educational testing area; otherwise my life would undoubtedly have followed a much different path. Todd helped me in various ways to clarify the things related to my research in time with excellent guidance. I will always be grateful for their kind assistance and support.

I would like to thank Dr. Samira ElAtia for serving as my committee member and providing valuable comments. Sincere gratitude is also extended to my colleagues in the Centre for Research in Applied Measurement and Evaluation, particularly Xuan Tan, Ying Cui, and Changjiang Wang, for their helpful discussion.

This piece of work would not have been possible without the support of my family, especially from my mother and father. Their understanding and unconditional help throughout my masters study have been deeply appreciated.

To My Mother and Father

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

Background to the Problem

The topic of differential item functioning (DIF) has motivated considerable attention in educational and psychological testing over the past decades. As one step in the test validity process, DIF analysis plays an important role in efforts to achieve test fairness across demographic populations. Fairness in testing is regarded as fairly and equitably assessing examinees of different groups without bias (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999; Principles for Fair Student Assessment Practices for Education in Canada, 1993). Issues of test fairness are important because high-stakes testing occurs worldwide and leads to educational admissions, placement, and honors/awards decision-making processes (Stout, 2002). As a response to the need to attain test fairness, DIF analysis and its accuracy are therefore receiving increased emphasis in the educational measurement literature (American Educational Research Association, et al., 1999).

Generally, in a DIF analysis, it is necessary to first identify two groups of examinees that will be compared. Often the two groups are different in terms of race, gender, or language (e.g., English, French). DIF analysis involves administering a test to examinees of the two groups, matching members of the two groups on a measure of target construct(s) measured by the test, and identifying group differences on test items using statistical procedures. An item exhibits DIF when examinees belonging to different groups have differing probabilities of answering that item correctly, after controlling for the ability of interest. The last clause of the definition, requiring that differences in ability do not exist after matching on the ability of interest, reveals validly matching examinees is a fundamental prerequisite of DIF analyses. DIF can be detected by a variety of statistical methods (Clauser & Mazor, 1998). Among the methodologies available, most of them employ number-correct score (NC) as the matching criterion (e.g., Simultaneous Item Bias Test [SIBTEST], Shealy & Stout, 1993a). The NC score used with SIBTEST in a DIF analysis is the total test score minus the studied item(s) score. For instance, in a test with 40 items, the scores on items 2 to 40 are used as the matching criteria when item 1 is treated as the studied item. It is reasonable to use the NC score as matching criterion if the assumption of unidimensionality is tenable.

In practice, however, some tests are not unidimensional. According to the test specifications, some tests are intentionally developed to be multidimensional, which means these tests are designed to measure multiple traits or dimensions (Clauser, Nungester, Mazor, & Ripkey, 1996). As test items intentionally measure more than one primary dimension, the NC score may no longer be sufficient for matching examinees since there is no dimensional consideration in the selection of the matching criterion. For instance, assume a test consists of three primary dimensions, logical reasoning (LR), reading comprehension (RC), and analytical reasoning (AR), which the test is intended to measure. The number of items in the three subsets is 30, 40, and 30, respectively. For a subgroup of test takers with 70% correct on the total test, such examinees could obtain this score by correctly answering many different combinations of the three subsets of items (e.g., the combination of 18, 30, and 22 or the combination of 10, 36, and 24). It is evident that misleading and inconsistent results could occur when NC is used as a matching criterion in such a multidimensional situation. Moreover, adding the score of these three different primary dimensions together results in a loss of information as differences between LR, RC, and AR cannot be evaluated. Hence, in order to promote the accuracy of the DIF analysis in a multidimensional test, it is critical to match examinees on multiple primary dimensions so that examinees are comparable on all primary dimensions the test is intended to measure before examinees are compared (Mazor, Hambleton, & Clauser, 1998; Clauser et al., 1996).

MULTISIB, a direct extension of the SIBTEST DIF analysis procedure proposed by Stout, Li, Nandakumar, and Bolt (1997), is a DIF detection program which can match examinees on two primary dimensions. Stout et al. (1997) reported that "MULTISIB demonstrated good Type I error behavior and reasonable power across a wide range of sample sizes" based on the results of the simulation studies.

Purpose of Current Study

The purpose of present study was 1) to compare the DIF detection performance of SIBTEST and MULTISIB in a multidimensional testing situation with respect to Type I error and power rates and 2) to investigate the impact of two factors on the performance of SIBTEST and MULTISIB—sample size and correlation between two primary dimensions. The specific research questions addressed in this study include:

1. Is the DIF detection performance of SIBTEST influenced by sample size?
2. Is the DIF detection performance of SIBTEST influenced by different degrees of correlation between two primary dimensions?
3. Is the DIF detection performance of MULTISIB influenced by sample size?
4. Is the DIF detection performance of MULTISIB influenced by different degrees of correlation between two primary dimensions?

5. What are the correct and incorrect classification rates of DIF and non-DIF items for SIBTEST?
6. What are the correct and incorrect classification rates of DIF and non-DIF items for MULTISIB?
7. What are the differences between SIBTEST and MULTISIB for a multidimensional test in terms of DIF detection performance?

To begin, related literature and the technical frameworks for SIBTEST and MULTISIB are reviewed in Chapter 2. This review is then followed by the methods in Chapter 3 and the results in Chapter 4. Conclusions and discussion are presented in Chapter 5.

Chapter 2: Overview OF SIBTEST and MULTISIB

Dimensionality and the Matching Criterion for DIF analysis

Shealy and Stout (1993b) presented an elaborate, in-depth, theoretical description of the multidimensional model for DIF (MMD) that provides a rigorous framework for understanding how DIF occurs in a multidimensional test. The term dimension should be defined to clarify the underlying cause of DIF. According to Shealy and Stout (1993b), *dimension* refers to any substantive characteristic of an item that can affect the probability of correctly answering the item. DIF occurs because an item is sensitive not only to the intended primary dimension(s) but also to secondary dimension(s) and a difference exists on secondary construct(s) between two demographic groups of interest after controlling for ability on the primary dimension(s). That is, DIF is present only when examinees with equal ability but belonging to either the *reference group* or the *focal group* have differing probabilities of answering an item correctly. Generally, the reference group is a majority group whereas the focal group is a minority group. In other words, DIF is attributable to multidimensionality because the dimension(s) not intended to be measured on the test distinctly affects the performance of examinees in different groups (Ackerman, 1992; Camilli & Shepard, 1994; Gierl, 2005; Lord, 1980; Roussos & Stout, 1996a; Shealy & Stout, 1993a).

The dimensions that a test is designed to measure are defined as the *primary dimensions* of the test. The general cause of DIF is the presence of the items that measure at least one dimension other than the primary dimensions in a test. These additional potential DIF-causing dimensions are referred to as *secondary dimensions.* A secondary dimension is also called a *nuisance dimension* because it is not the one designated to be measured on a test.

In addition, matching examinees on the primary dimensions is also a critical issue of DIF analysis because the examinees belonging to different groups with equivalent ability are supposed to perform equally on a same item. That is, the examinees of different groups with equal ability should obtain the same score on an item which is free from DIF. Alternatively, DIF occurs if this assumption is violated. Therefore, an appropriate matching criterion that can validly match examinees with equal ability is a key element for a DIF analysis procedure.

Results from both real data studies (Clauser, Mazor, & Hambleton, 1991; Mazor, Kanjee, & Clauser, 1995) and simulated data studies (Ackerman, 1992; Mazor et al, 1998) have highlighted the importance of the choice of the matching criterion for DIF analysis approaches. Clauser et al. (1991), for instance, indicated that 32% of the items (7 out of 22) were no longer identified to be DIF items when conditioning on a single subtest score instead of the total test score when the Mantel-Haenszel (MH, Holland & Thayer, 1988) DIF procedure was used. Ackerman (1992) presented an empirical example of how conditioning on a valid subtest score rather than total test score can substantially vary the results of a MH DIF analyses. Six out of seven invalid (DIF) items were clearly identified when the matching criterion was the valid subtest score. When the total test score was used as the matching condition, 10 out of 18 valid items and six out of seven invalid (DIF) items were labeled as DIF items, which indicates that Type I error rates were dramatically increased. More recently, researchers have found that simultaneously conditioning on the multiple ability estimates for an intentional multidimensional test led to the detection of substantially fewer DIF items (Clauser et al., 1996; Mazor et al., 1998). Clauser et al. (1996) indicated that the MH procedure showed 60% agreement and the logistic regression (LR, Swaminathan & Rogers, 1990) procedure showed only 57% agreement, in identified DIF items between matching on multiple subtests score and total test score. In the Mazor et al. study (1998), the MH and LR DIF analysis procedures matching on multiple subtests yielded lower Type I error rates compared to rates from analysis matching on NC score. For example, the average LR Type I error rate was 34.6% when the NC score was adopted as the matching criterion while it was 1.6% as matching on *a priori* subtests. These findings emphasize that failure to condition on multiple valid dimensions may allow intended multidimensional items to be falsely identified as DIF items.

However, an analogous study considering the impact of different matching criteria on the performance of SIBTEST, and its multidimensional version, MULTISIB, is lacking in the literature. SIBTEST is a powerful procedure that estimates the amount of DIF by controlling inflated Type I error using a regression correction technique. SIBTEST matches examinees using the NC score which is the total score minus the score on the studied item, while MULTISIB simultaneously matches examinees on the scores of two subtests which subtract the score on the studied item. MULTISIB, therefore, is deemed to be a direct extension of SIBTEST because the difference between the two approaches is the matching criterion adopted. An overview of SIBTEST and MULTISIB is presented next.

Overview of the SIBTEST DIF Detection Procedure

With respect to MMD, the primary dimension(s) and all the secondary dimension(s) in a test represent the complete latent space, which is viewed as multidimensional. SIBTEST, proposed by Shealy and Stout (1993a), is designed to statistically test DIF hypotheses and identify items in the secondary dimension(s) that produce group differences when there is only one primary dimension. It can also be used to qualify the size of DIF. Two groups of examinees, the reference and focal groups, are administered items on a test. Items on the study test are divided into the *matching subtest* and the *studied subtest*. The matching subtest contains the items believed to measure only the primary dimension while the studied subtest contains the items believed to measure not only the primary but also the secondary dimensions. Matching subtest scores are used to place the reference and focal group examinees into subgroups at distinct score levels so they are judged approximately equivalently on the intended primary dimension. Therefore, their performances on a studied item or the studied subtest can be compared on the secondary dimension(s).

In the case of a traditional single-item DIF analysis, only one item is included in the studied subtest and the matching subtest contains the remaining test items. It is often conducted when the researcher or practitioner has either none or only a few *a priori* ideas about which items elicit group differences. This is the reason why this commonly used approach for DIF detection is also called an exploratory analysis of single item. In the case of a bundle analysis, two or more items are included in the studied subtest. In either case, the number of the items in the matching subtest is fixed. Only single-item DIF analyses were conducted in this study.

The statistical hypothesis tested by SIBTEST is:

$$H_0 : \beta_{UNI} = 0 \ vs. \ H_1 : \beta_{UNI} \neq 0,$$

where β_{UNI} is the parameter specifying the magnitude of DIF for an item. β_{UNI} is defined as:

$$\beta_{UNI} = \int_{\theta} B(\theta) f_F(\theta) d\theta,$$

where $B(\theta) = P(\theta, R) - P(\theta, F)$ is the difference in the probabilities of correct response for examinees from the reference and focal groups with ability θ . $f_F(\theta)$ is the density function for θ_F in the focal group, and $d\theta$ is the differential of theta. β_{UNI} is integrated over θ to produce a weighted expected mean difference in the probability of a correct response on an item between reference and focal group examinees of the same ability.

More specifically, let N denote the total number of items in a study test, items $1, \ldots, n$ denote the matching subtest items, and $n+1, \ldots, N$ denote the studied subtest items. Let U_i denote the response to item i scored as 0 or 1. For each examinee, $X = \sum_{i=1}^{n} U_i$ to specify the total score on the matching subtest and $Y = \sum_{i=n+1}^{N} U_i$ to specify the total score on the studied subtest. The matching subgroups are indexed by total score k, $k = 0, \ldots n,$ on the matching subtest. Examinees in the reference and focal groups are then grouped into these k subgroups with respect to their matching subtest scores. Examinees within each subgroup k are treated equivalently on θ, hence their performance on the studied subtest can be compared between reference and focal group to assess whether DIF is present.

The actual weighted mean difference between the reference and focal groups on the studied subtest item across the k subgroups is given by:

$$\hat{\beta}_{UNI} = \sum_{k=0}^{K} p_k d_k \ ,$$

which provides an estimate of β_{UNI}. p_k in this equation is the proportion of focal group examinees in subgroup k among all focal group examinees, and d_k denotes $(\bar{Y}^*_{Rk} - \bar{Y}^*_{Fk})$, which is the difference in the adjusted means on the studied subtest scores for examinees in the reference and focal groups for each subgroup k. The means on the studied subtest item are adjusted to correct for any mean differences in the ability distributions of the reference and focal groups using a regression correction described in Shealy and Stout (1993a). In most cases, the reference and focal groups have different distributions of target ability. Hence, aligning examinees of different groups in terms of equal scores on the matching subtest will fail because examinees do not have equal ability levels across the two groups. Any between-group distribution difference on the target construct can statistically inflate β_{UNI}, thereby inaccurately indicating that DIF is present. The regression correction procedure helps to overcome this inherent limitation. The weakness of using observed scores in the matching subtest given these scores contain measurement error is then controlled and minimized. The regression correction estimates the matching subtest true score in each subgroup k, respectively, for the reference and focal groups. Correspondingly, the mean scores on the target dimension of studied subtest for reference and focal groups are adjusted to the new values, $\bar{Y}^*_{Rk}$ and $\bar{Y}^*_{Fk}$. The differences between adjusted means, d_k, across valid subtest score levels then become the basis for evaluating DIF on the studied subtest item.

The estimates of the valid subtest score means, variances, and reliabilities for examinees in the reference and focal groups are calculated using the observed test scores in each subgroup. These statistics are then used to estimate $V_R(k)$ and $V_F(k)$, which are the theoretical regression of the matching subtest true score with observed score k, respectively, for the reference and focal groups. Denoting the mean of the two true scores $\hat{V}_R(k)$ and $\hat{V}_F(k)$ by $\hat{V}(k)$, the slope of the item response function in the region of score k for group g can be estimated with

$$\hat{M}_{gk} = \frac{\bar{Y}_{g,k+1} - \bar{Y}_{g,k-1}}{\hat{V}_g(k+1) - \hat{V}_g(k-1)}.$$

An adjusted conditional proportion correct score on the studied subtest for each group can then be computed as

$$\bar{Y}_{gk}^{*} = \bar{Y}_{gk} + \hat{M}_{gk}[\hat{V}(k) - \hat{V}_g(k)],$$

where $\bar{Y}_{gk}$ is the observed mean on the studied subtest for examinees in group g (reference or focal group) with $X = k$ on the matching subtest.

SIBTEST also yields an overall statistical test for $\hat{\beta}_{UNI}$. The test statistic for evaluating the null hypothesis is:

$$SIB = \frac{\hat{\beta}_{UNI}}{\hat{\sigma}(\hat{\beta}_{UNI})},$$

where $\hat{\sigma}(\hat{\beta}_{UNI})$ is the estimated standard error of $\hat{\beta}_{UNI}$. Shealy and Stout (1993a) demonstrated that SIB has a normal distribution with mean 0 and variance 1 under the null hypothesis. The null hypothesis is rejected if SIB exceeds the $100(1-\alpha)/2$ percentile point from the normal distribution using a non-directional hypothesis test.

Positive values of $\hat{\beta}_{UNI}$ indicate DIF favoring the reference group and negative values indicate DIF favoring the focal group. Roussos and Stout (1996b) proposed guidelines for interpreting DIF by combining the SIBTEST statistical results with values for the $\hat{\beta}_{UNI}$ parameter estimate to classify DIF on a single item: (a) negligible DIF: $\left|\hat{\beta}_{UNI}\right| < 0.059$ and $H_0 : \beta_{UNI} = 0$ is rejected, (b) moderate DIF: $0.059 \leq \left|\hat{\beta}_{UNI}\right| < 0.088$ and $H_0 : \beta_{UNI} = 0$ is rejected, (c) large DIF: $\left|\hat{\beta}_{UNI}\right| > 0.088$ and $H_0 : \beta_{UNI} = 0$ is rejected. Alternatively, as $H_0 : \beta_{UNI} = 0$ is not rejected, there is no DIF found in the studied item.

As a popular DIF detection procedure, the performance of SIBTEST was evaluated in different studies. A simulation study was conducted by Shealy and Stout (1993a). Sample size and the level of DIF items were manipulated. The results indicated that SIBTEST displayed acceptable Type I error and power rates. In particular, the average Type I error rate for all conditions in the simulation study was 6.0%; that is, SIBTEST displayed reasonable adherence to the nominal level of significance of 0.05. The average power rate was 75% across all sample size conditions. Moreover, a higher power rate was produced in larger sample sizes conditions with moderate or large DIF items. For example, the power rate was 89% for 1,500 examinees per group while it was 71% for 500 sample size condition with moderate or large DIF items.

Roussos and Stout (1996b) conducted a simulation study with small sample size to explore the Type I error performance of SIBTEST. Item parameters taken from the Armed Services Vocational Aptitude Battery (ASVAB) were selected to simulate a 25-item test. One item chosen from the test served as the non-DIF studied item. The mean differences between reference and focal groups were set at 0.0, 0.5, and 1.0 modeled on a normal distribution with a variance of 1.0. The results revealed that with small sample size such as 100, 200, 500, and 1,000, the Type I error rates adhered quite well to the nominal level of significance of 0.05. For instance, the average observed significance level for SIBTEST was 0.049 across all the combinations of sample size and group mean difference conditions.

A simulation study for probing the effects of large DIF on SIBTEST was executed by Gierl, Gotzmann, and Boughton (2004). The number of DIF items, direction of DIF, sample size, and ability distribution differences were four variables manipulated in the study. Gierl et al. (2004) concluded that SIBTEST provided adequate DIF detection which means the Type I error rates were less than 5% and the power rates were greater than 80% when DIF was balanced and sample sizes were 1,000 and over.

Overview of the MULTISIB DIF Detection Procedure

MULTISIB, the nature extension of SIBTEST DIF detection procedure proposed by Stout et al. (1997), is designed to identify items evaluating the secondary dimension(s) and estimate the magnitude of DIF for two-dimensional tests. As a multidimensional counterpart of SIBTEST, MULTISIB can match examinees of different groups with equal ability so that the examinees' score on the studied item can be compared and to determine if DIF is present in that studied item. The same statistical hypothesis as in SIBTEST is tested by MULTISIB:

$$H_0 : \beta_{UNI} = 0 \; vs. \; H_1 : \beta_{UNI} \neq 0 .$$

The basic logic in MULTISB is that, as the multidimensional version of SIBTEST, as long as examinees from different groups are *simultaneously* matched on the intended primary dimension one and dimension two, their score on the studied item can be compared to determine whether DIF is present in the item. The examinees from the reference and focal groups are administered a test known to be two-dimensional. N denotes the total number of items in a two-dimensional test. The items mainly measuring primary dimension one, θ_1, are grouped in the first matching subtest and n_1 denotes the item number in this matching subtest. The second matching subtest contains items believed to assess primary dimension two, θ_2, more than θ_1. Let n_2 denotes the item number in the second matching subtest. In the case of a single-item DIF analysis, only one item is included in the studied subtest while two or more items are included in the studied subtest in the case of a bundle analysis. In either case, the matching subtest is fixed. X_1 and X_2 are the total scores on the Matching subtest 1 and subtest 2, respectively. Let Y denote the score on the studied subtest which contains either one or more items that potentially cause DIF.

Examinees from the reference and focal groups are divided into subgroups based on their scores on matching subtests, X_1 and X_2. Examinees are first grouped into k_1 subgroups in terms of their score X_1 and grouped into k_2 subgroups regarding to their score X_2. Examinees on the two matching subtests are then combined to set up joint subgroups so that all examinees in each subgroup have the same scores on X_1 and X_2.

Due to the possible distribution difference on the target traits between the reference and focal groups, regression theory is applied to correct and, therefore, to minimize the inflated Type I error, as in SIBTEST. $V_{gj}(k_j)$ denotes the expected true score on the matching subtest j ($j = 1,2$) for group g (reference or focal group) in the subgroup with $X_j = k_j$. As for the unidimensional case, the estimated studied item true score for examinees in the subgroup (k_1, k_2) is then given by

$$\overline{Y}^*_{g(k_1,k_2)} = \overline{Y}_{g(k_1,k_2)} + \widehat{M}_{g1(k_1,k_2)}[\widehat{V}_1(k_1) - \widehat{V}_{g1}(k_1)] + \widehat{M}_{g2(k_1,k_2)}[\widehat{V}_2(k_2) - \widehat{V}_{g2}(k_2)],$$

where

$$\widehat{M}_{g1(k_1,k_2)} = \frac{\overline{Y}_{g(k_1+1,k_2)} - \overline{Y}_{g(k_1-1,k_2)}}{\widehat{V}_{g1}(k_1+1) - \widehat{V}_{g1}(k_1-1)},$$

and

$$\widehat{M}_{g2(k_1,k_2)} = \frac{\overline{Y}_{g(k_1,k_2+1)} - \overline{Y}_{g(k_1,k_2-1)}}{\widehat{V}_{g2}(k_2+1) - \widehat{V}_{g2}(k_2-1)},$$

for group g (reference or focal group).

The weighted mean difference between the reference and focal groups on the studied subtest item, $\widehat{\beta}_{UNI}$, which provides an estimate of β_{UNI}, can be interpreted as the magnitude of DIF for each item. Positive values of $\widehat{\beta}_{UNI}$ indicate DIF favoring the reference group and negative values indicate DIF favoring the focal group. The same guidelines as for SIBTEST can be applied to MULTISIB, as the two procedures are different dimensional versions of one method. The guidelines to classify DIF by combining the MULTISIB statistical results on a single item are: (a) negligible DIF: $|\widehat{\beta}_{UNI}| < 0.059$ and $H_0 : \beta_{UNI} = 0$ is rejected, (b) moderate DIF: $0.059 \leq |\widehat{\beta}_{UNI}| < 0.088$ and $H_0 : \beta_{UNI} = 0$ is rejected, (c) large DIF: $|\widehat{\beta}_{UNI}| > 0.088$ and $H_0 : \beta_{UNI} = 0$ is rejected. Alternatively, as $H_0 : \beta_{UNI} = 0$ is not rejected, there is no DIF found in the studied item.

The DIF detection performance of MILTISIB regarding to Type I error and power rates was evaluated by Stout et al. (1997) using a simulation study. Three factors were manipulated: sample size for reference and focal groups, the mean difference between the ability distributions of reference and focal groups, and the level of item discrimination, item difficulty, and item guessing parameters. The matching subtest consisted of 40 items, with the first 20 items strictly measuring primary dimension one and the last 20 items strictly measuring primary dimension two. The correlation between the two primary dimensions was 0.50. Fourteen items which measured some composite of the two primary dimensions served as studied items for the Type I error study. An additional 19 items served as studied items for the power analysis. The Type I error rate was 5.7% for 300 examinees per group, 6.6% for 500, 6.0% for 1,000, 5.9% for 1,500, and 4.5% for 3,000 examinees per group, respectively. The power rate was 26% for 300 examinees per group, 36% for 500, 50% for 1,000, 63% for 1,500, and 82% for 3,000 examinees per group. Obviously, the Type I error rates adhered quite well to the nominal level of significance of 0.05 and was relatively unaffected by the increasing sample size. However, the power rate increased with the increasing sample size. Stout et al. (1997) concluded that "MULTISIB demonstrated good Type I error behavior and reasonable power across a wide range of sample sizes" based on the results of their study.

Chapter 3: Method

A simulation study was conducted to compare and evaluate the Type I error and power rates of SIBTEST and MULTISIB to detect DIF in two-dimensional test data. Examinee response data were simulated under specific conditions expected to affect DIF detection rates. Two factors were manipulated: sample size (500, 1,000, 1,500, and 2,000 examinees in each group) and the correlation between the two primary dimensions ($\rho_{12} = 0.20$, 0.40, 0.60, and 0.80). The levels of each factor were designed to reflect those that might be found in real data. Test length was consistent: 70 items with 50 matching items and 20 studied items were constructed.

Manipulated Factors

Sample Size

Previous research indicated sample size affects DIF items detection (Stout et al., 1997; Gierl et al., 2004). In the simulation studies conducted by Stout et al. (1997), sample size was a key variable that impacted the DIF detection rates for MULTISIB. In actual testing situations, sample size is a condition that deserves attention because it can vary dramatically with both small and large sample sizes occurring. Thus, to explore the effect of sample size on DIF detection rates for the two procedures, sample size was considered as a factor in this simulation study. Four levels of sample sizes were evaluated: 500, 1,000, 1,500, and 2,000 examinees. Five hundred is a relatively small sample size while 2000 is a large one. The reference and focal groups had the same number of examinees; hence sample size was balanced in all conditions.

Correlation Between Dimension

The correlation between the two primary dimensions, θ_1 and θ_2, was the second factor considered. This variable was evaluated because the primary dimensions can be perceived as one single dimension when their correlation is high, thereby making the potential benefit of matching on different primary dimensions negligible. Four levels of this factor were considered: 0.20, 0.40, 0.60, and 0.80. A zero correlation between primary dimensions is unrealistic in any practical testing situation and was therefore not considered. Similarly, correlations greater than 0.80 are also unusual. The small correlation $\rho_{12} = 0.20$ implied the two primary dimensions in the simulated test are quite distinct while the large correlation $\rho_{12} = 0.80$ implied that the two primary dimensions are very similar.

Thus, the design for this DIF analyses study was a 4×4 crossed design, with four levels of sample size and four levels of correlation between primary dimensions to produce 16 conditions in total. Each condition was replicated 100 times to facilitate calculations of Type I error and power rates (Harwell, Stone, Hsu, & Kirisci, 1996).

Data Generation and Analysis

The examinee item responses to the 70 items were simulated by MULTISIM using the compensatory multidimensional item response theory (MIRT) model (Reckase, 1997). The 3PL item response function (IRF) for the compensatory MIRT model can be expressed by the following formula

$$P_i[U_i = 1 \left| (\theta_{1,\ldots,}\theta_k) \right|] = c_i + \frac{1 - c_i}{1 + e^{-1.7(a_{i1}\theta_1 + a_{i2}\theta_2 + \ldots a_{ik}\theta_k + d_i)}},$$

where U_i is the response to item i, $\vec{\theta}^T = (\theta_{1,...,}\theta_k)$ is the vector of examinee ability, $\vec{a}_i^T = (a_{1,...,}a_k)$ is the vector of discrimination parameter, d_i is the multidimensional difficulty parameter, c_i is the guessing parameter, and k is the number of dimensions underlying the test, which in the present study was two. The examinee abilities were assumed to have a bivariate normal distribution with a mean of (0, 0) and a standard deviation of (1, 1).

The number of items in the current simulated test was 70, with 50 matching items and 20 studied items. The 50 matching items were evenly distributed across two primary dimensions; that is, 25 items measured primary dimension θ_1 and 25 items measured primary dimension θ_2. The a_1-parameters of items measuring primary dimension θ_1 were set in the range of 0.35 to 1.55 with an increment of 0.30, while the a_2-parameters were restricted in the range of 0.05 to 0.30 to ensure the directions of the 25 items measuring θ_1 were within the range of 1.85° to 17.53°. The angular direction of an item was calculated using

$$\alpha = \arccos\frac{a_1}{MDISC},$$

where $MDISC = \sqrt{a_1^2 + a_2^2}$ is the multidimensional discrimination which represents the multidimensional slope of the surface in different directions. For items measuring primary dimension θ_2, the values of the a_1-and a_2- parameters were set in the reverse order hence the directions of the 25 items measuring primary dimension θ_2 were bounded in the range of 74.05° to 88.15°. The 50 items approximated simple structure as the angular directions were both within 20.00 degrees from the x- or y- axis (Froelich &

Habing, 2001). The *d*-parameters ranged from -1.00 to 1.00 with an increment of 0.50. The guessing parameter was set at 0.20 for all matching items. Table 1 contains the item parameters and the angular directions of each item for the 50 matching items. The vector plots of the items measuring dimension θ_1 and θ_2 are shown in Figures 1 and 2, respectively.

The remaining 20 items in the simulated test were studied items intended to test the DIF detection rates of SIBTEST and MULTISIB. The first eight of these 20 studied items were non-DIF items which measured the two primary dimensions. These items are denoted as non-DIF items because they do not measure a secondary dimension and, therefore, do not differentially impact the performances of the examinees in reference and focal groups. Three of these items were referenced to primary dimension θ_1, three to primary dimension θ_2, and two equally to both θ_1 and θ_2. The *d*-parameter of the 8 items ranged from -1.00 to 1.10. The guessing parameter remained at 0.20. The item parameters and the angular directions of the eight designed non-DIF items are listed in Table 2.

The remaining 12 items were DIF items that dominantly measured one of the two primary dimensions as well as a secondary nuisance dimension, θ_3. The a_1-parameters and a_2-parameters of the 12 items were set in the range of 0.10 to 1.50 and 0.05 to 1.25, while the a3-parameters, the discrimination parameter for the nuisance dimension θ_3, were restricted to the range 0.30 to 1.55, thereby producing differential item responses. The 12 DIF items were simulated as four negligible DIF items, four moderate DIF items, and four large DIF items. The difference in the mean of the distributions on the secondary dimension between the reference and focal groups was set within the range of 0.50 and

Table 1

Item parameters and angular direction of 50 matching items

Item	a_1	a_2	d	c	MDISC	D	α
1	0.35	0.05	-1.00	0.20	0.35	2.83	8.13°
2	0.65	0.05	-1.00	0.20	0.65	1.53	4.40°
3	0.95	0.05	-1.00	0.20	0.95	1.05	3.01°
4	1.25	0.10	-1.00	0.20	1.25	0.80	4.57°
5	1.55	0.10	-1.00	0.20	1.55	0.64	3.69°
6	0.35	0.10	-0.50	0.20	0.36	1.37	15.95°
7	0.65	0.08	-0.50	0.20	0.65	0.76	7.02°
8	0.95	0.10	-0.50	0.20	0.96	0.52	6.01°
9	1.25	0.10	-0.50	0.20	1.25	0.40	4.57°
10	1.55	0.15	-0.50	0.20	1.56	0.32	5.53°
11	0.35	0.05	0.00	0.20	0.35	0.00	8.13°
12	0.65	0.20	0.00	0.20	0.68	0.00	17.10°
13	0.95	0.10	0.00	0.20	0.96	0.00	6.01°
14	1.25	0.12	0.00	0.20	1.26	0.00	5.48°
15	1.55	0.15	0.00	0.20	1.56	0.00	5.53°
16	0.35	0.05	0.50	0.20	0.35	-1.41	8.13°
17	0.65	0.15	0.50	0.20	0.67	-0.75	12.99°
18	0.95	0.30	0.50	0.20	1.00	-0.50	17.53°
19	1.25	0.12	0.50	0.20	1.26	-0.40	5.48°
20	1.55	0.05	0.50	0.20	1.55	-0.32	1.85°
21	0.35	0.08	1.00	0.20	0.36	-2.79	12.88°
22	0.65	0.10	1.00	0.20	0.66	-1.52	8.75°
23	0.95	0.05	1.00	0.20	0.95	-1.05	3.01°
24	1.25	0.15	1.00	0.20	1.26	-0.79	6.84°
25	1.55	0.30	1.00	0.20	1.58	-0.63	10.95°
26	0.05	0.35	-1.00	0.20	0.35	2.83	81.87°
27	0.05	0.65	-1.00	0.20	0.65	1.54	85.60°
28	0.05	0.95	-1.00	0.20	0.95	1.05	86.99°
29	0.05	1.25	-1.00	0.20	1.25	0.80	87.71°
30	0.05	1.55	-1.00	0.20	1.55	0.64	88.15°
31	0.05	0.35	-0.50	0.20	0.35	1.41	81.87°
32	0.10	0.65	-0.50	0.20	0.66	0.76	81.25°
33	0.05	0.95	-0.50	0.20	0.95	0.53	86.99°
34	0.10	1.25	-0.50	0.20	1.25	0.40	85.43°
35	0.10	1.55	-0.50	0.20	1.55	0.32	86.31°
36	0.10	0.35	0.00	0.20	0.36	0.00	74.05°
37	0.10	0.65	0.00	0.20	0.66	0.00	81.25°
38	0.15	0.95	0.00	0.20	0.96	0.00	81.03°
39	0.10	1.25	0.00	0.20	1.25	0.00	85.43°
40	0.15	1.55	0.00	0.20	1.56	0.00	84.47°
41	0.05	0.35	0.50	0.20	0.35	-1.41	81.87°
42	0.15	0.65	0.50	0.20	0.67	-0.75	77.01°
43	0.20	0.95	0.50	0.20	0.97	-0.52	78.11°
44	0.09	1.25	0.50	0.20	1.25	-0.40	85.88°
45	0.15	1.55	0.50	0.20	1.56	-0.32	84.47°
46	0.05	0.35	1.00	0.20	0.35	-2.83	81.87°
47	0.10	0.65	1.00	0.20	0.66	-1.52	81.25°
48	0.05	0.95	1.00	0.20	0.95	-1.05	86.99°
49	0.10	1.25	1.00	0.20	1.25	-0.80	85.43°
50	0.09	1.55	1.00	0.20	1.55	-0.64	86.68°

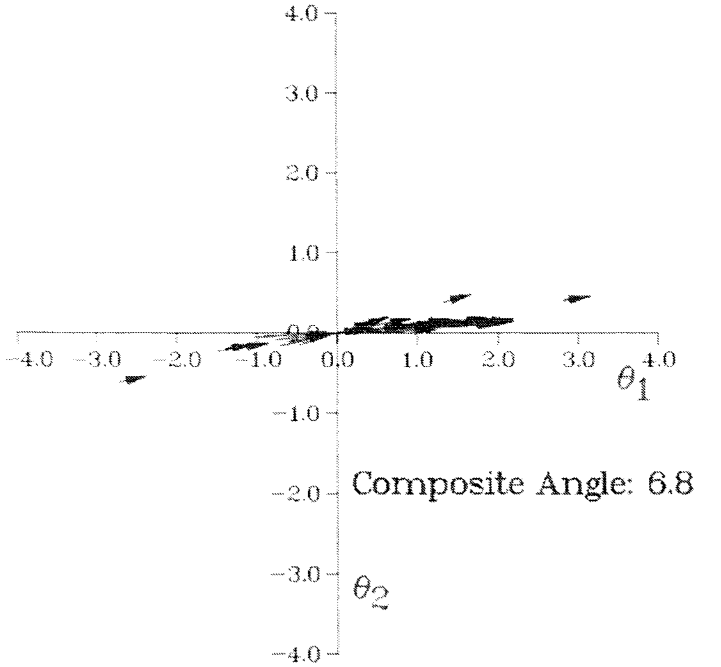

Figure 1. Vector plot of simulated items measuring primary dimension 1.

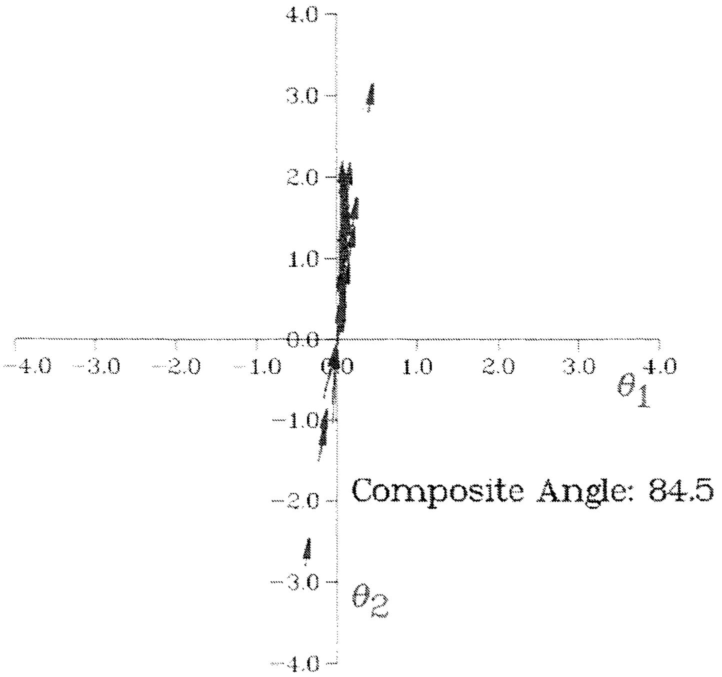

Figure 2. Vector plot of simulated items measuring primary dimension 2.

Table 2

Item parameters and angular direction of 8 non-DIF items

Item	a_1	a_2	a_3	d	c	MDISC	D	α
1	0.40	0.00	0.00	0.80	0.2	0.40	-2.00	0.00°
2	0.00	1.00	0.00	-0.50	0.2	1.00	0.50	90.00°
3	1.20	0.10	0.00	0.10	0.2	1.20	-0.08	4.76°
4	1.80	0.15	0.00	-0.50	0.2	1.81	0.28	4.76°
5	0.15	1.15	0.00	0.80	0.2	1.16	-0.69	82.57°
6	0.15	1.55	0.00	-0.90	0.2	1.56	0.58	84.47°
7	1.20	1.20	0.00	1.10	0.2	1.70	-0.65	45.00°
8	0.75	0.75	0.00	-1.00	0.2	1.06	0.94	45.00°

1.00 by Stout et al. (1997). However, as they noted, the mean difference between groups of one approximately corresponds to the largest value obtained in actual applications. Thus, in present study, the differences of d-parameters on negligible, moderate, and large DIF items were 0.05, 0.20, and 0.40, respectively, across reference and focal groups. The d-parameters of the 12 DIF items, therefore, for the reference group ranged from -0.70 to 1.00, whereas those for focal group were within the range of -0.75 to 0.95. The guessing parameters of the 12 items, for both reference and focal groups, were set to 0.20. Table 3 contains the item parameters and the angular directions of the 12 DIF items for reference and focal groups.

The computer programs SIBTEST and MULTISIB were used for the DIF analyses with the simulated data sets. The procedure to test for DIF using SIBTEST involved dividing the examinees from the reference and focal groups into subgroups based on a single test score derived from the 50 matching subtest. In each of these subgroups, examinees' scores on the studied item were compared to determine whether DIF is caused by the item. The procedure to test for DIF using MULTISIB involved two matching subtests. The items primarily measuring dimension θ_1 were grouped into

Table 3

Item parameters and angular direction of 12 DIF items for reference group and focal group

Group	Item	a_1	a_2	a_3	d	c	MDISC	D	α
Reference group	1	0.90	0.20	1.15	0.50	0.20	0.92	-0.54	12.53°
	2	0.55	0.15	1.20	1.00	0.20	0.57	-1.75	15.26°
	3	0.10	0.35	0.85	-0.70	0.20	0.36	1.92	74.05°
	4	0.15	1.20	1.20	0.10	0.20	1.21	-0.08	82.87°
	5	1.50	0.30	1.25	0.20	0.20	1.53	-0.13	11.31°
	6	0.90	0.15	0.30	-0.50	0.20	0.91	0.55	9.46°
	7	0.30	1.25	1.55	-0.35	0.20	1.29	0.27	76.50°
	8	0.15	0.50	0.35	0.85	0.20	0.52	-1.63	73.30°
	9	1.00	0.05	0.50	0.60	0.20	1.00	-0.60	2.86°
	10	0.60	0.20	1.10	1.00	0.20	0.63	-1.58	18.43°
	11	0.15	1.00	0.80	-0.30	0.20	1.01	0.30	81.47°
	12	0.15	0.50	0.90	-0.50	0.20	0.52	0.96	73.30°
Focal group	1	0.90	0.20	1.15	0.45	0.2	0.92	-0.49	12.53°
	2	0.55	0.15	1.20	0.95	0.2	0.57	-1.67	15.26°
	3	0.10	0.35	0.85	-0.75	0.2	0.36	2.06	74.05°
	4	0.15	1.20	1.20	0.05	0.2	1.21	-0.04	82.87°
	5	1.50	0.30	1.25	0.00	0.2	1.53	0.00	11.31°
	6	0.90	0.15	0.30	-0.70	0.2	0.91	0.77	9.46°
	7	0.30	1.25	1.55	-0.55	0.2	1.29	0.43	76.50°
	8	0.15	0.50	0.35	0.65	0.2	0.52	-1.25	73.30°
	9	1.00	0.05	0.50	0.20	0.2	1.00	-0.20	2.86°
	10	0.60	0.20	1.10	0.60	0.2	0.63	-0.95	18.43°
	11	0.15	1.00	0.80	-0.70	0.2	1.01	0.69	81.47°
	12	0.15	0.50	0.90	-0.90	0.2	0.52	1.72	73.30°

matching subtest 1 while the items mainly measuring dimension θ_2 were grouped into matching subtest 2. Once examinees from the reference and focal groups are simultaneously matched on the subtest scores of matching subtest 1 and subtest 2, the performance of examinees from the reference and focal groups on the item of interest was compared.

The guidelines for interpreting DIF by Roussos and Stout (1996b; see Chapter 2) were used to classify DIF items for SIBTEST and MULTISIB. Two-tailed hypothesis tests were conducted for all analyses using an alpha level of 0.05. Two types of DIF detection rates were assessed. Type I error occurred when $H_0 : \beta_{UNI} = 0$ for a non-DIF item was incorrectly rejected. Conversely, power occurred when $H_0 : \beta_{UNI} = 0$ was correctly rejected. Furthermore, the identified DIF items by both procedures, correctly or incorrectly, were flagged using the conventions for negligible, moderate, or large DIF. The proportion of correct classification of the non-DIF items and different magnitudes DIF items was also evaluated in this study.

Chapter 4: Results

The results of the simulation study are presented for the 0.20, 0.40, 0.60, and 0.80 correlation conditions in Tables 4 and 5, which contain Type I error and power rates, respectively. In each condition, results are displayed based on increasing sample size. Tables 6 to 9 contain the classification results of the 20 study items consisting of eight non-DIF items, four negligible DIF items, four moderate DIF items, and four large DIF items for SIBTEST and MULTISIB. The proportions of correct classification for each correlation and corresponding sample size condition are listed in the four tables.

Type I Error Results

Table 4 contains the results of SIBTEST and MULTISIB on the Type I error rates for each level of correlation between two primary dimensions across four levels of sample size. For SIBTEST, with a correlation of $\rho_{12} = 0.20$, the Type I error rates decreased from 0.4% to 0.0% as the sample size increased from 500 to 2,000. With a correlation of $\rho_{12} = 0.40$, the Type I error rates decreased from 0.8% to 0.0% as the sample size increased from 500 to 2,000. With a correlation of $\rho_{12} = 0.60$, the Type I error rates decreased from 0.6% to 0.0% as the sample size increased from 500 to 2,000. With a correlation of $\rho_{12} = 0.80$, the Type I error rates were different from those in other correlation conditions. They varied inconsistently across four sample sizes, making it difficult to evaluate the trend in this condition.

For MULTISIB, with a correlation of $\rho_{12} = 0.20$, the Type I error rates decreased from 0.9% to 0.0% as the sample size increased from 500 to 2,000. With a correlation of $\rho_{12} = 0.40$, the Type I error rates decreased from 0.9% to 0.0% as the sample size increased from 500 to 2,000. With a correlation of $\rho_{12} = 0.60$, the Type I error rates

Table 4

Type I error rates for SIBTEST and MULTISIB

Correlation between dimensions	Sample size	SIBTEST(%)	MULTISIB(%)
0.20	$N_R=N_F=500$	0.38	0.88
	$N_R=N_F=1000$	0.25	0.38
	$N_R=N_F=1500$	0.00	0.25
	$N_R=N_F=2000$	0.00	0.00
0.40	$N_R=N_F=500$	0.75	0.88
	$N_R=N_F=1000$	0.50	0.50
	$N_R=N_F=1500$	0.13	0.00
	$N_R=N_F=2000$	0.00	0.00
0.60	$N_R=N_F=500$	0.63	1.63
	$N_R=N_F=1000$	0.38	0.38
	$N_R=N_F=1500$	0.13	0.00
	$N_R=N_F=2000$	0.00	0.00
0.80	$N_R=N_F=500$	0.50	1.38
	$N_R=N_F=1000$	0.75	0.63
	$N_R=N_F=1500$	0.00	0.00
	$N_R=N_F=2000$	0.13	0.00

decreased from 1.6% to 0.0% as the sample size increased from 500 to 2,000. With a correlation of $\rho_{12} = 0.80$, the Type I error rates decreased from 1.4% to 0.0% as the sample size increased from 500 to 2,000.

In addition, the incorrect detection rates for MULTISIB were greater than or equal to SIBTEST across all conditions with four exceptions. For instance, in Table 4, with a correlation of $\rho_{12} = 0.20$, the Type I error rate for MULTISIB, 0.9%, was greater than the corresponding Type I error rate for SIBTEST, 0.4%, in the 500 sample size condition. The four exceptions occurred in the 1,500 sample size with a $\rho_{12} = 0.40$; the 1,500 sample size with a $\rho_{12} = 0.60$; the 1,000 sample size with a $\rho_{12} = 0.80$; and the 2,000 sample size with a $\rho_{12} = 0.80$ conditions.

Moreover, the variation of the Type I error rates on both SIBTEST and MULTISIB did not produce a consistent pattern with the increase of the correlation rates in each sample size condition. For example, the Type I error rate on SIBTEST with 500 examinees per group was 0.4% for the 0.20 correlation condition, 0.8% for 0.40, 0.6% for 0.60, and 0.5% for the 0.80 correlation condition, respectively.

Power Results

Table 5 presents the power rates for SIBTEST and MULTISIB. The power rates for SIBTEST consistently increased as sample size increased across all correlation conditions. With a correlation of $\rho_{12} = 0.20$, the power rates increased from 60.8% to 99.8% as the sample size increased from 500 to 2,000. With a correlation of $\rho_{12} = 0.40$, the power rates increased from 61.5% to 99.7% as the sample size increased from 500 to 2,000. With a correlation of $\rho_{12} = 0.60$, the power rates increased from 61.3% to 99.5% as the sample size increased from 500 to 2,000. With a correlation of $\rho_{12} = 0.80$, the power rates increased from 61.9% to 98.5% as the sample size increased from 500 to 2,000.

For MULTISIB, the power rates increased as sample size increased across all correlation conditions except for the 1,500 examinees per group conditions. With a correlation of $\rho_{12} = 0.20$, the power rates increased from 48.0% to 96.7% as the sample size increased from 500 to 2,000 except that the power rate for 1,500 examinees per group condition, 78.2%, was smaller than the power rate for the 1,000 examinees per group condition, 81.8%. With a correlation of $\rho_{12} = 0.40$, the power rates increased from 48.3% to 97.8% as the sample size increased from 500 to 2,000 while the power rate

Table 5

Power rates for SIBTEST and MULTISIB

Correlation between dimensions	Sample size	SIBTEST(%)	MULTISIB(%)
0.20	N_R=N_F=500	60.83	48.00
	N_R=N_F=1000	87.75	81.83
	N_R=N_F=1500	97.75	78.18
	N_R=N_F=2000	99.75	96.67
0.40	N_R=N_F=500	61.50	48.33
	N_R=N_F=1000	86.67	83.00
	N_R=N_F=1500	97.58	78.75
	N_R=N_F=2000	99.67	97.75
0.60	N_R=N_F=500	61.25	49.50
	N_R=N_F=1000	86.06	80.67
	N_R=N_F=1500	95.33	80.50
	N_R=N_F=2000	99.50	95.92
0.80	N_R=N_F=500	61.92	52.25
	N_R=N_F=1000	84.17	80.15
	N_R=N_F=1500	92.50	81.33
	N_R=N_F=2000	98.50	95.33

for 1,500 examinees per group condition, 78.8%, was smaller than the one for the 1,000 examinees per group condition, 83.0%. With a correlation of $\rho_{12} = 0.60$, the power rates increased from 49.5% to 95.9% as the sample size increased from 500 to 2,000 while the power rate for 1,500 examinees per group condition, 80.5%, was slightly smaller than the one for the 1,000 examinees per group condition, 80.7%. With a correlation of $\rho_{12} = 0.80$, the power rates consistently increased from 52.3% to 95.3% as the sample size increased from 500 to 2,000.

Furthermore, the power rates on MULTISIB were consistently smaller than SIBTEST in all conditions. For instance, with a correlation of $\rho_{12} = 0.20$, the power rate for MULTISIB, 81.8%, was less than the corresponding power rate for SIBTEST, 87.8%, in the 1,000 sample size condition.

In addition, the variation of the power rates for both procedures did not produce a consistent pattern with the increase of the correlation rates for each sample size condition. The power rates on SIBTEST decreased with increasing correlations when the sample size was 1,000 and over, for example, from 87.8% to 84.2% in the 1,000 sample size condition. However, it varied unsystematically with the increasing correlation in 500 sample size condition. The power rates on MULTISIB increased with the increase of the correlations only in 500 and 1,500 sample size conditions. The power rates for MULTISIB varied unsystematically with the increasing correlation in 1,000 and 2,000 sample size conditions.

Non-DIF and DIF Item Classification Results

Table 6 contains the classification results for the 20 study items across the two DIF detection procedures for the 0.20 correlation conditions. Similar results were found for both SIBTEST and MULTISIB. The eight non-DIF items were correctly classified by SIBTEST and MULTISIB across four different sample size conditions. For instance, with 500 sample size, the correct classification rate of non-DIF items was 99.6% and 99.1% for SIBTEST and MULTISIB, respectively.

The negligible DIF items were primarily grouped as non-DIF items when the sample size was 500 by both SIBTEST and MULTISIB. For example, the correct classification rate of the negligible DIF items was 69.3% and 79.0% for SIBTEST and MULTISIB, respectively. The proportion of correct grouping of the negligible DIF items increased as the sample size increased for both procedures. For example, the correct

Table 6

Study Items Classification Results across SIBTEST & MULTISIB with $\rho_{12} = 0.20$ conditions

		SIBTEST				MULTISIB			
Sample Size	Designed Study Items	non-DIF	Negligible DIF	Moderate DIF	Large DIF	non-DIF	Negligible DIF	Moderate DIF	Large DIF
500	8 non-DIF Items	99.63%	0.00%	0.38%	0.00%	99.13%	0.00%	0.88%	0.00%
	4 negligible DIF Items	69.25%	0.50%	26.25%	4.00%	79.00%	0.00%	12.50%	8.50%
	4 moderate DIF Items	45.00%	2.25%	37.75%	15.00%	63.25%	0.00%	19.75%	17.00%
	4 large DIF Items	3.25%	0.00%	17.75%	79.00%	13.75%	0.00%	14.25%	72.00%
1000	8 non-DIF Items	99.75%	0.25%	0.00%	0.00%	99.63%	0.38%	0.00%	0.00%
	4 negligible DIF Items	33.75%	49.75%	16.50%	0.00%	45.00%	33.75%	20.50%	0.75%
	4 moderate DIF Items	3.00%	30.75%	62.25%	4.00%	9.50%	21.75%	62.50%	6.25%
	4 large DIF Items	0.00%	0.00%	5.75%	94.25%	0.00%	0.25%	7.00%	92.75%
1500	8 non-DIF Items	100.00%	0.00%	0.00%	0.00%	99.75%	0.25%	0.00%	0.00%
	4 negligible DIF Items	6.25%	85.50%	8.25%	0.00%	50.25%	30.25%	19.50%	0.00%
	4 moderate DIF Items	0.50%	33.50%	65.50%	0.50%	15.25%	16.00%	59.75%	9.00%
	4 large DIF Items	0.00%	0.00%	2.50%	97.50%	0.00%	0.00%	6.25%	93.75%
2000	8 non-DIF Items	100.00%	0.00%	0.00%	0.00%	100.00%	0.00%	0.00%	0.00%
	4 negligible DIF Items	0.75%	95.25%	4.00%	0.00%	9.00%	84.00%	7.00%	0.00%
	4 moderate DIF Items	0.00%	27.25%	72.75%	0.00%	1.00%	25.75%	72.75%	0.50%
	4 large DIF Items	0.00%	0.00%	1.25%	98.75%	0.00%	0.00%	1.75%	98.25%

Note: The cells of correct classifications are highlighted.

classification rate of the negligible DIF items was 0.0% for MULTISIB while it was 84.0% as the sample size increased from 500 to 2,000.

The moderate DIF items were primarily classified into non-DIF items when the sample size was 500 (i.e., 45.0% for SIBTEST and 63.3% for MULTISIB). The correct classification rates of the moderate DIF items increased with the increasing sample size for both procedures. For instance, the proportion of correct grouping of the moderate DIF items by SIBTEST increased from 37.8% to 72.8% as the sample size increased from 500 to 2,000.

The majority of large DIF items were correctly classified when the sample size was 500 (i.e., 79.0% for SIBTEST and 72.0% for MULTISIB). When the sample size was 1,000 and over, the large DIF items were correctly classified by both procedures: 94.3% in the 1,000 sample size condition, 97.5% in the 1,500 sample size condition, and 98.8% in the 2,000 sample size condition for SIBTEST, for example.

The majority of the proportions of correct classifications for both non-DIF and DIF items on SIBTEST were larger than or equal to the corresponding correct classification rates for MULTISIB across each sample size condition. For example, the SIBTEST correct classification rate of non-DIF items was 99.6% while the MULTISIB correct classification rate of non-DIF items was 99.1% in the 500 sample size condition. The SIBTEST correct classification rate of the negligible DIF items was 49.8% while the MULTISIB correct classification rate of the negligible DIF items was 33.8% in the 1,000 sample size condition.

Table 7 contains the classification results for the 20 study items across the two DIF detection procedures for the 0.40 correlation conditions. Similar patterns of results were found for both SIBTEST and MULTISIB. The eight non-DIF items were correctly classified by SIBTEST and MULTISIB across four different sample size conditions. For example, with 500 sample size, the correct classification rate of non-DIF items was 99.3% and 99.1% for SIBTEST and MULTISIB, respectively.

The negligible DIF items were primarily grouped as non-DIF items when the sample size was 500 for both SIBTEST and MULTISIB. For instance, the correct classification rate of the negligible DIF items was 70.3% and 79.5% for SIBTEST and MULTISIB, respectively. The proportion of correct grouping of the negligible DIF items increased as the sample size increased for both procedures. For example, the correct classification rates of the negligible DIF items increased from 0.0% to 90.0% as the sample size increased from 500 to 2,000 for MULTISIB.

The moderate DIF items were primarily classified into non-DIF items when the sample size was 500 (i.e., 42.8% for SIBTEST and 60.3% for MULTISIB). The correct classification rates of the moderate DIF items increased with increasing sample size for both procedures with one exception. The proportion of correct grouping of the moderate DIF items by SIBTEST increased from 41.0% to 73.3% as the sample size increased from 500 to 2,000. However, the proportion of correct grouping by MULTISIB was 23.0% in the 500 sample size, 63.8% in the 1,000 sample size, 54.0% in the 1,500 sample size, and 69.0% in the 2,000 sample size.

The majority of the large DIF items were correctly classified when the sample size was 500 (i.e., 78.5% for SIBTEST and 71.3% for MULTISIB). When the sample size was 1,000 and over, the large DIF items were correctly classified by both procedures:

Table 7

Study Items Classification Results across SIBTEST & MULTISIB with $\rho_{12} = 0.40$ conditions

		SIBTEST				MULTISIB			
Sample Size	Designed Study Items	non-DIF	Negligible DIF	Moderate DIF	Large DIF	non-DIF	Negligible DIF	Moderate DIF	Large DIF
500	8 non-DIF Items	99.25%	0.38%	0.38%	0.00%	99.13%	0.00%	0.88%	0.00%
	4 negligible DIF Items	70.25%	0.50%	26.00%	3.25%	79.50%	0.00%	11.50%	9.00%
	4 moderate DIF Items	42.75%	3.25%	41.00%	13.00%	60.25%	0.00%	23.00%	16.75%
	4 large DIF Items	2.50%	0.25%	18.75%	78.50%	15.25%	0.00%	13.50%	71.25%
1000	8 non-DIF Items	99.50%	0.50%	0.00%	0.00%	99.50%	0.50%	0.00%	0.00%
	4 negligible DIF Items	35.75%	48.25%	16.00%	0.00%	42.75%	35.50%	21.25%	0.50%
	4 moderate DIF Items	4.25%	33.00%	60.25%	2.50%	8.25%	22.50%	63.75%	5.50%
	4 large DIF Items	0.00%	0.00%	8.00%	92.00%	0.00%	0.00%	11.00%	89.00%
1500	8 non-DIF Items	99.88%	0.13%	0.00%	0.00%	100.00%	0.00%	0.00%	0.00%
	4 negligible DIF Items	7.00%	84.50%	8.50%	0.00%	49.50%	35.25%	15.25%	0.00%
	4 moderate DIF Items	0.25%	35.00%	64.50%	0.25%	14.25%	23.75%	54.00%	8.00%
	4 large DIF Items	0.00%	0.00%	6.25%	93.75%	0.00%	0.00%	8.50%	91.50%
2000	8 non-DIF Items	100.00%	0.00%	0.00%	0.00%	100.00%	0.00%	0.00%	0.00%
	4 negligible DIF Items	1.00%	96.75%	2.25%	0.00%	6.75%	90.00%	3.25%	0.00%
	4 moderate DIF Items	0.00%	26.75%	73.25%	0.00%	0.00%	30.75%	69.00%	0.25%
	4 large DIF Items	0.00%	0.00%	2.00%	98.00%	0.00%	0.00%	1.00%	99.00%

Note: The cells of correct classifications are highlighted.

92.0% in the 1,000 sample size condition, 93.8% in the 1,500 sample size condition, and 98.0% in the 2,000 sample size condition for SIBTEST, for example.

The majority of correct classifications for both non-DIF and DIF items using SIBTEST were larger than or equal to the correct classification rates using MULTISIB across each sample size condition. For example, the percentage of SIBTEST correct classifications of non-DIF items was 99.3% while the percentage of MULTISIB correct classifications of non-DIF items was 99.1% in the 500 sample size condition. The SIBTEST correct classification rate for the moderate DIF items was 60.3% while the MULTISIB correct classification rate for the moderate DIF items was 63.8% in the 1,000 sample size condition.

Table 8 contains the classification results for the 20 study items across the two DIF detection procedures for the 0.60 correlation conditions. Similar results were found for both SIBTEST and MULTISIB. With few exceptions, eight non-DIF items were correctly classified by SIBTEST and MULTISIB across four different sample size conditions. For example, with the 500 sample size, the correct classification rate of non-DIF items was 99.4% and 98.4% for SIBTEST and MULTISIB, respectively.

The negligible DIF items were primarily grouped as non-DIF items when the sample size was 500 for both SIBTEST and MULTISIB (i.e., 71.3% and 81.5% for SIBTEST and MULTISIB, respectively). The proportion of correct grouping of the negligible DIF items increased as the sample size increased for both procedures. For example, the correct classification rates of negligible DIF items increased from 0.0% to 84.5% as the sample size increased from 500 to 2,000 for MULTISIB.

Table 8

Study Items Classification Results across SIBTEST & MULTISIB with $\rho_{12} = 0.60$ conditions

		SIBTEST				MULTISIB			
Sample Size	Designed Study Items	non-DIF	Negligible DIF	Moderate DIF	Large DIF	non-DIF	Negligible DIF	Moderate DIF	Large DIF
500	8 non-DIF Items	99.38%	0.38%	0.25%	0.00%	98.38%	0.13%	1.50%	0.00%
	4 negligible DIF Items	71.25%	0.50%	23.25%	5.00%	81.50%	0.00%	10.50%	8.00%
	4 moderate DIF Items	42.75%	3.25%	41.50%	12.50%	58.25%	0.00%	22.25%	19.50%
	4 large DIF Items	2.25%	0.25%	18.50%	79.00%	11.75%	0.00%	18.00%	70.25%
1000	8 non-DIF Items	99.63%	0.38%	0.00%	0.00%	99.50%	0.50%	0.00%	0.00%
	4 negligible DIF Items	37.75%	47.25%	15.00%	0.00%	50.00%	29.50%	20.25%	0.25%
	4 moderate DIF Items	4.00%	28.25%	62.25%	5.50%	8.00%	24.00%	61.00%	7.00%
	4 large DIF Items	0.00%	0.00%	7.75%	92.25%	0.00%	0.00%	11.75%	88.25%
1500	8 non-DIF Items	99.88%	0.13%	0.00%	0.00%	100.00%	0.00%	0.00%	0.00%
	4 negligible DIF Items	13.75%	81.00%	5.25%	0.00%	50.25%	35.50%	14.25%	0.00%
	4 moderate DIF Items	0.25%	34.75%	64.75%	0.25%	8.25%	31.00%	56.75%	4.00%
	4 large DIF Items	0.00%	0.00%	6.50%	93.50%	0.00%	0.00%	12.25%	87.75%
2000	8 non-DIF Items	100.00%	0.00%	0.00%	0.00%	100.00%	0.00%	0.00%	0.00%
	4 negligible DIF Items	1.50%	96.75%	1.75%	0.00%	12.25%	84.50%	3.25%	0.00%
	4 moderate DIF Items	0.00%	30.25%	69.75%	0.00%	0.00%	31.25%	68.00%	0.75%
	4 large DIF Items	0.00%	0.00%	1.75%	98.25%	0.00%	0.00%	3.75%	96.25%

Note: The cells of correct classifications are highlighted.

The moderate DIF items were primarily classified into non-DIF items when the sample size was 500 (i.e., 42.8% for SIBTEST and 58.3% for MULTISIB). The correct classification rates of the moderate DIF items increased with increasing sample size for both procedures with one exception. The proportion of correct grouping of the moderate DIF items by SIBTEST increased from 41.5% to 69.8% as the sample size increased from 500 to 2,000. However, the proportion of correct grouping by MULTISIB was 22.3% in the 500 sample size, 61.0% in the 1,000 sample size, 56.85% in the 1,500 sample size, and 68.0% in the 2,000 sample size.

The large DIF items were primarily correctly classified when the sample size was 500 (i.e., 79.0% for SIBTEST and 70.3% for MULTISIB). When the sample size was 1,000 and over, the large DIF items were correctly classified by both procedures: 92.3% in the 1,000 sample size condition, 93.5% in the 1,500 sample size condition, and 98.3% in the 2,000 sample size condition for SIBTEST, for example.

The majority of correct classifications for both non-DIF and DIF items using SIBTEST were larger than or equal to the corresponding correct classification rates using MULTISIB across each sample size condition. For example, the SIBTEST correct classification rate of non-DIF items was 99.4% while the MULTISIB correct classification rate of non-DIF items was 98.4% in the 500 sample size condition. The SIBTEST correct classification rate of moderate DIF items was 62.3% while the MULTISIB correct classification rate of moderate DIF items was 61.0% in the 1,000 sample size condition.

Table 9 contains the classification results for the 20 study items across the two DIF detection procedures for the 0.80 correlation conditions. Similar results were

Table 9

Study Items Classification Results across SIBTEST & MULTISIB with $\rho_{12} = 0.80$ conditions

		SIBTEST				MULTISIB			
Sample Size	Designed Study Items	non-DIF	Negligible DIF	Moderate DIF	Large DIF	non-DIF	Negligible DIF	Moderate DIF	Large DIF
500	8 non-DIF Items	99.50%	0.38%	0.13%	0.00%	98.63%	0.00%	1.38%	0.00%
	4 negligible DIF Items	69.25%	1.00%	26.25%	3.50%	76.25%	0.00%	17.50%	6.25%
	4 moderate DIF Items	42.75%	4.50%	41.75%	11.00%	56.25%	0.25%	26.50%	17.00%
	4 large DIF Items	2.25%	0.00%	18.50%	79.25%	10.75%	0.00%	16.00%	73.25%
1000	8 non-DIF Items	99.25%	0.75%	0.00%	0.00%	99.38%	0.63%	0.00%	0.00%
	4 negligible DIF Items	41.25%	45.00%	13.75%	0.00%	48.50%	32.50%	18.75%	0.25%
	4 moderate DIF Items	6.25%	35.50%	53.25%	5.00%	10.00%	25.25%	57.50%	7.25%
	4 large DIF Items	0.00%	0.00%	8.25%	91.75%	0.00%	0.00%	9.75%	90.25%
1500	8 non-DIF Items	100.00%	0.00%	0.00%	0.00%	100.00%	0.00%	0.00%	0.00%
	4 negligible DIF Items	22.50%	74.75%	2.75%	0.00%	52.50%	40.00%	7.50%	0.00%
	4 moderate DIF Items	0.00%	42.00%	57.50%	0.50%	3.50%	36.00%	55.75%	4.75%
	4 large DIF Items	0.00%	0.00%	10.25%	89.75%	0.00%	0.00%	10.50%	89.50%
2000	8 non-DIF Items	99.88%	0.13%	0.00%	0.00%	100.00%	0.00%	0.00%	0.00%
	4 negligible DIF Items	4.50%	94.75%	0.75%	0.00%	14.00%	82.75%	3.25%	0.00%
	4 moderate DIF Items	0.00%	27.50%	72.50%	0.00%	0.00%	28.00%	71.25%	0.75%
	4 large DIF Items	0.00%	0.00%	12.50%	87.50%	0.00%	0.00%	9.75%	90.25%

Note: The cells of correct classifications are highlighted.

found for both SIBTEST and MULTISIB. The eight non-DIF items were essentially correctly classified by SIBTEST and MULTISIB across four different sample size conditions. For instance, with 500 sample size, the correct classification rate of non-DIF items was 99.5% and 98.6% for SIBTEST and MULTISIB, respectively.

The negligible DIF items were primarily grouped as non-DIF items when the sample size was 500 for both SIBTEST and MULTISIB (i.e., 69.3% and 76.3% for SIBTEST and MULTISIB, respectively). The proportion of correct grouping of the negligible DIF items increased as the sample size increased for both procedures. For example, the correct classification rates of negligible DIF items increased from 0.0% to 82.8% as the sample size increased from 500 to 2,000 for MULTISIB.

The moderate DIF items were primarily classified into non-DIF items when the sample size was 500 (i.e., 42.8% for SIBTEST and 56.3% for MULTISIB). The correct classification rates of the moderate DIF items increased with the increasing sample size for both procedures with one exception. The proportion of correct grouping of the moderate DIF items by SIBTEST increased from 41.8% to 72.5% as the sample size increased from 500 to 2,000. However, the proportion of correct grouping by MULTISIB was 26.5% in the 500 sample size, 57.5% in the 1,000 sample size, 55.8% in the 1,500 sample size, and 71.3% in the 2,000 sample size.

The large DIF items were primarily correctly classified when the sample size was 500 (i.e., 79.3% for SIBTEST and 73.3% for MULTISIB). When the sample size was 1,000 and over, the large DIF items were correctly classified by both procedures. For instance, 91.8% in the 1,000 sample size condition, 89.8% in the 1,500 sample size condition, and 87.5% in the 2,000 sample size condition for SIBTEST.

The majority of correct classifications for both non-DIF and DIF items using SIBTEST were larger than or equal to the corresponding correct classification rates using MULTISIB across each sample size condition. For example, the SIBTEST correct classification rate of the non-DIF items was 99.5% while the MULTISIB correct classification rate of the non-DIF items was 98.6% in the 500 sample size condition. The SIBTEST correct classification rate of negligible DIF items was 45.0% while the MULTISIB correct classification rate of moderate DIF items was 32.5% in the 1,000 sample size condition.

Chapter 5: Discussion

Conclusions and Discussion

The purpose of this study was to evaluate and compare the DIF detection performance of SIBTEST and its multidimensional version, MULTISIB, using intentional multidimensional data. This study is relevant to researchers and practitioners alike because many tests are purposefully designed to measure more than one construct. Thus, the influence of matching criterion on the DIF detection rates in a multidimensional test is an important issue as the appropriate DIF analysis procedures with dimensional consideration in the selection of the matching criterion are required (Ackerman, 1992; Clauser et al., 1991; Clauser et al., 1996; Mazor et al., 1995; Mazor et al., 1998).

Two independent variables, correlation between primary dimensions and sample size, were included in the simulation design employed in the present study to investigate the impact of these two factors on the performance of SIBTEST and MULTISIB. The Type I error and power rates generated by SIBTEST and MULTISIB were compared. The correct classification results of each level of DIF items in all conditions were also reported.

Conservative results were obtained for the Type I error rates and the correct classification rates for both SIBTEST and MULTISIB. With only two exceptions, the Type I error rates for these two procedures in this study were less than 1.0% (see Table 4). The correct classification rates of negligible and moderate DIF items were also low when the sample size was small.

Seven research questions were raised in the introduction section. The answers to the seven questions are presented next.

Is the DIF detection performance of SIBTEST influenced by sample size?

The DIF detection performance of SIBTEST was strongly affected by sample size. Tables 4 and 5 show the decreasing Type I error and increasing power rates for SIBTEST with increasing sample size. As listed in Table 4, when the sample size increased from 500 to 2,000, the Type I error rates systematically decreased, with only one exception. For instance, the Type I error rates decreased from 0.4% to 0.0% when the correlation between two primary dimensions was 0.20. As shown in Table 5, the power rates for SIBTEST increased with the increasing sample size. For example, the power rates, with a correlation of $\rho_{12} = 0.20$, were 60.8% as sample size was 500, 87.8% as sample size was 1,000, 97.8% as sample size was 1,500, and 99.8% as sample size was 2,000. Sample size is therefore deemed as a key factor in the performance of SIBTEST.

Is the DIF detection performance of SIBTEST influenced by different degrees of correlation between two primary dimensions?

The DIF detection performance of SIBTEST was not strongly affected by the correlation between primary dimensions in current study. As shown in Table 4, the Type I error rates varied unsystematically across different correlation conditions. For example, the Type I error rate on SIBTEST with 500 examinees per group was 0.4% for the 0.20 correlation condition, 0.8% for 0.40, 0.6% for 0.60, and 0.5% for the 0.80 correlation condition, respectively. The power rates varied unsystematically with the increasing correlation. In Table 5, the power rate for SIBTEST with 500 examinees per group was 60.8% for the 0.20 correlation condition, 61.5% for 0.40, 61.3% for 0.60, and 61.9% for 0.80. The power rates for SIBTEST decreased as the correlation increased when the sample size was equal to or greater than 1,000. For instance, the power rate for 1000 examinees per group decreased from 87.8% to 84.2% when the correlation increased from 0.20 to 0.80. That is, there was no consistent pattern of performance, especially on Type I error rate, for SIBTEST as a function of correlation between dimensions.

Is the DIF detection performance of MULTISIB influenced by sample size?

The DIF detection performance of MULTISIB was strongly affected by sample size. Tables 4 and 5 show the decreasing Type I error and increasing power rates for MULTISIB with increasing sample size. As listed in Table 4, when the sample size increased from 500 to 2,000, the Type I error rates systematically decreased. For instance, the Type I error rates decreased from 0.9% to 0.0% as the correlation between two primary dimensions was 0.20. As shown in Table 5, the power rates for MULTISIB increased with the enlarged sample size except for the 1,500 examinees per group conditions. For example, the power rates, with a correlation of $\rho_{12} = 0.20$, were 48.0% as sample size was 500, 81.8% as sample size was 1,000, 78.2% as sample size was 1,500, and 96.7% as sample size was 2,000. Sample size is therefore regarded as a key factor to the performance of MULTISIB.

Is the DIF detection performance of MULTISIB influenced by different degrees of correlation between two primary dimensions?

The correlation between primary dimensions did not influence the DIF detection performance of MULTISIB. Tables 4 and 5 show the variable Type I error and power rates for MULTISIB as the correlation between two primary dimensions increased from 0.20 to 0.80. For example, the Type I error rate on MULTISIB with 500 examinees per group was 0.9% for the 0.20 correlation condition, 0.9% for 0.40, 1.6% for 0.60, and 1.4% for the 0.80 correlation condition, respectively (see Table 4). In Table 5, the power rate on MULTISIB with 1,000 examinees per group was 81.8% for the 0.20 correlation condition, 83.0% for 0.40, 80.7% for 0.60, and 80.2% for 0.80. The power rates on MULTISIB increased as the correlation increased only when the sample size was 500 and 1,500. For instance, the power rate for 500 examinees per group increased from 48.0% to 52.3% as the correlation increased from 0.20 to 0.80. That is, there was no consistent pattern of performance, especially for the Type I error rate, for MULTISIB with increased correlation rates.

What are the correct and incorrect classification rates of DIF and non-DIF items for SIBTEST?

Non-DIF items were correctly classified across all the conditions by SIBTEST. Negligible and moderate DIF items were incorrectly classified when sample size was small; they both had large proportions of improperly identified non-DIF items when the sample size was 500. For example, 69.3% of negligible DIF items and 45.0% of moderate DIF items were identified by SIBTEST as non-DIF items when the number of examinees was 500 and $\rho_{12} = 0.20$ (see Table 6). However, the majority of the negligible and moderate DIF items were correctly classified when the sample size was 1,000 or larger. For example, 49.8% of negligible DIF items and 62.3% of moderate DIF items were correctly classified when the sample size was 1,000. Large DIF items were, most often, correctly identified. For instance, 79.0% of large DIF items were properly classified by SIBTEST in the 500 examinees per group, $\rho_{12} = 0.20$ condition. In addition, as the sample size increased, the proportion of correct classifications of DIF items at all levels increased. For example, as the sample size increased from 500 to 2,000, as shown in Table 6, the SIBTEST correct classification rates of negligible DIF items increased from

0.5% to 95.3%; the SIBTEST correct classification rates of moderate DIF items increased from 37.8% to 72.8%. However, with increasing correlation between primary dimensions, there was little variation on the proportion of correct classifications of DIF items for all levels of DIF. For example, when the sample size was 500, the SIBTEST correct classification rate of negligible DIF items was 0.5% for $\rho_{12} = 0.20$, 0.40, and 0.60, and 1.0% for $\rho_{12} = 0.80$. It was therefore concluded that sample size had an impact on the correct classification rate of DIF items for SIBTEST while the correlation between primary dimensions did not.

What are the correct and incorrect classification rates of DIF and non-DIF items for MULTISIB?

Non-DIF items were correctly classified across all the conditions by MULTISIB. Negligible and moderate DIF items were incorrectly classified in some sample size conditions; they both had large proportions of improperly identified non-DIF items in these sample size conditions. For example, 79.0%, 45.0%, and 50.3% of negligible DIF items were identified by MULTISIB as non-DIF items when the sample size was 500, 1,000, and 1,500 with $\rho_{12} = 0.20$. There were 63.3% of moderate DIF items identified by MULTISIB as non-DIF items in the condition of 500 examinees per group across $\rho_{12} = 0.20$ condition. However, the majority of the moderate DIF items were correctly classified by MULTISIB when the sample size was 1,000 or larger. In Table 6, 62.5% of moderate DIF items were correctly classified when the sample size was 1,000. Large DIF items were, most often, correctly identified. For instance, 72.0% of large DIF items were properly classified by MULTISIB in the 500 examinees per group and $\rho_{12} = 0.20$ condition. In addition, as the sample size increased, the proportion of correct classifications of DIF items on all magnitude levels increased with several exceptions. For example, as the sample size increased from 500 to 2,000, as shown in Table 6, the MULTISIB correct classification rates of negligible DIF items increased from 0.0% to 84.0%; the MULTISIB correct classification rates of large DIF items increased from 72.0% to 98.3%. However, with increasing correlation between primary dimensions, there was little variation on the proportion of correct classifications of DIF items on all magnitude levels. For example, when the sample size was 500, the MULTISIB correct classification rate of large DIF items was 72.0% for $\rho_{12} = 0.20$, 71.3% for $\rho_{12} = 0.40$, 70.3% for $\rho_{12} = 0.60$, and 73.3% for $\rho_{12} = 0.80$. It was therefore concluded that sample size had an impact on the correct classification rate of DIF items for MULTISIB while the correlation between primary dimensions did not.

What is the difference between SIBTEST and MULTISIB for a multidimensional test in terms of DIF detection performance?

The two procedures both performed well for DIF detection along with correct DIF item classification as long as the sample size was 1,000 and over. That is, the Type I error rates for both SIBTEST and MULTISIB were less than 5.0% in all conditions. The power rates for the two procedures were greater than 80.0% in all conditions except for the 500 examinees conditions. Consequently, the proportions of correct classification of DIF items were acceptable when the sample size was 1,000 or larger.

There was no substantive difference between incorrect detection rates for the two procedures. The Type I error rate was 0.4% for SIBTEST in the 500 examinees per group and $\rho_{12} = 0.20$ condition, and it was 0.9% for MULTISIB in the same condition. However, there was a systematic difference between the power rates of SIBTEST and

MULTISIB. The power rates for SIBTEST were consistently greater than the power rates for MULTISIB. For example, the power rate was 87.8% for SIBTEST versus 81.8% for MULTISIB for the 1,000 examinees, $\rho_{12} = 0.20$ condition, and 92.5% for SIBTEST versus 81.3% for MULTISIB for the 1,500 examinees, $\rho_{12} = 0.80$ condition. Moreover, the increase in the proportions of correct classifications of both non-DIF and DIF items were greater with SIBTEST compared to MULTISIB. As shown in Table 6, for example, when the sample size increased from 500 to 2,000, the proportion of SIBTEST correct classification of negligible items increased from 0.5% to 95.3% while that of MULTISIB increased from 0.0% to 84.0%. In a word, MULTISIB did not perform as well as SIBTEST in the current study.

Based on the results of previous relevant studies (Ackerman, 1992; Clauser et al., 1991; Clauser et al., 1996; Mazor et al., 1995; Mazor et al., 1998), the performance of MULTISIB was expected to be superior to SIBTEST due to the multidimensional matching criterion. The data simulated in current study was intentionally multidimensional. SIBTEST, which adopts the number-correct score, was expected to ineffectively match examinees on two distinct dimensions simultaneously and therefore perform as a weak matching criterion for a multidimensional test relative to MULTISIB. Nevertheless, the results listed above revealed that the number-correct score matching criterion performed as well as, if not better than, the multidimensional matching criterion under the conditions evaluated in this study.

One potential cause for these results was that the matching subtest was simulated to possess approximate simple structure. As an almost ideal state, the approximate simple structure condition simplified the relationship between items belonging to the two primary dimensions. Items that approximate simple structure only measure one of the primary dimensions. Conversely, items in a relatively complex structure condition often measure both primary dimensions. The simplification of the matching subtest in current study likely affected the DIF detection performance of SIBTEST and MULTISIB. The advantage of MULTISIB regarding its multidimensional matching criterion was cancelled out by the approximate simple structure. That is, although SIBTEST matches on the number-correct score without dimensional consideration in the matching criterion, its disadvantage with respect to the matching criterion did not impact its performance because of the simple structure of the data studied. SIBTEST, with its comparatively straightforward underlying principle, therefore performed better than MULTISIB under the conditions evaluated in this study.

Limitations of Current Study

The results from this study provide researchers and practitioners with some insights into the detection rates for SIBTEST and MULTISIB where the test is designed as multidimensional. However, only simulated data were analyzed in this study. The item parameters used for simulating matching subtest items were systematically manipulated and, in turn, the simulated matching subtest approximated simple structure. Complex structure items, which can happen in real testing situations, were lacking in the matching subtests considered in the present study. Thus, the results obtained in current study must be generalized to real testing situations with caution.

Only two variables were manipulated in current study. The two variables, sample size and correlation between primary dimensions, are fundamental ones in real testing situations and practical DIF analyses. Hence, they served as the focus in the current study.

However, other variables, such as the number of DIF items, direction of DIF, and ability distribution differences could affect the results. The use of only sample size and correlation between the primary dimensions can therefore be deemed as a limitation.

Moreover, although sample size was specified in each analysis condition, due to the requirement of meeting the minimum numbers of examinees in each score level of the matching subtest(s), the exact numbers of examinees involved in SIBTEST and MULTISIB analyses were not the same as expected. For instance, the proportion of the examinees eliminated was around 30% for 500 sample size with 0.20 correlation condition in SIBTEST. Reducing the number of examinees eliminated in the analysis will enhance the representativeness of the sample distribution, that is, examinees with different performance level can be included in the analysis. When the number of examinees retained is close to the specified sample size, the accuracy of DIF analysis using SIBTEST and MULTISIB will be promoted, theoretically. On the other hand, a sufficient sample size often leads to results which represent reality more closely.

Future Directions for Research

More research is needed to evaluate SIBTEST and MULTISIB using item parameters derived from real tests for simulation. On the other hand, supplementing simulation studies with real data analysis has become more common in the research literature. Due to limited resources of multidimensional testing data, this study was not supplemented by a real data analysis. Adopting realistic item parameters as well as conducting real data analysis will help make future studies more generalizable to real-world testing situations.

The variables manipulated were sample size and correlation between primary dimensions in the current study. These two variables are fundamental ones in DIF analyses. However, to some extent, the study design was limited in the number of variables manipulated. In future study, other variables, such as amount of DIF items, can be added to enrich the study design and emulate the testing situations in practice.

Compared with the results of a simulation study conducted by Stout et al. (1997), in which the performance of MULTISIB was evaluated, the DIF detection rates of Type I error rates in the current study were low and conservative while the power rates were greater (see page 39). The Type I error rates for MULTISIB were 5.7% for 300 sample size, 6.6% for 500 sample size, 6.0% for 1,000 sample size, 5.9% for 1,500 sample size, and 4.5% for 3,000 sample size in the Stout et al. (1997) study. The power rates were 26% for 300 sample size, 36% for 500 sample size, 50% for 1,000 sample size, 63% for 1,500 sample size, and 82% for 3,000 sample size. The difference on the results between two studies should be explored in future study.

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.

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.

Camilli, G., & Shepard, L. (1994). *Methods for identifying biased test items.* Newbury Park, CA: Sage.

Clauser, B. E., & Mazor, K. M. (1998). Using statistical procedures to identify differentially functioning test items. *Educational Measurement: Issues and Practice, 17,* 31-44.

Clauser, B. E., Mazor, K., & Hambleton, R. K. (1991). The influence of the criterion variable on the identification of differentially functioning items using the Mantel-Haenszel statistic. *Applied Psychological Measurement, 15*, 353-359.

Clauser, B. E., Nungester, R. J., Mazor, K., & Ripkey, D. (1996). A comparison of alternative matching strategies for DIF detection in tests that are multidimensional. *Journal of Educational Measurement. v33*(2), 202-14.

Froelich, A. G., & Habing, B. (2001). *Refinements of the DIMTEST methodology for testing unidimensionality and local independence.* Paper presented at the annual meeting of the National Council on Measurement in Education, Seattle, WA.

Gierl, M. J. (2005). Using dimensionality-based DIF analyses to identify and interpret constructs that elicit group differences. *Educational Measurement: Issues & Practice. Vol 24*(1), 3-14.

Gierl, M. J., Gotzmann, A., & Boughton, K.A. (2004). Performance of SIBTEST when the percentage of DIF items is large. *Applied Measurement in Education, 17*, 241-264.

Harwell, M., Stone, C. A., Hsu, T., Kirisci, L. (1996). Monte carlo studies in item response theory. *Applied Psychological Measurement. Vol 20*(2), 101-125.

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. I Braun (Eds.), *Test validity* (pp. 129-145). Hillsdale, NJ: Erlbaum.

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

Mazor, K. M., Hambleton, R. K., & Clauser, B. E. (1998). Multidimensional DIF analyses: The effects of matching on unidimensional subtest scores. *Applied Psychological Measurement. Vol 22*(4), 357-367.

Mazor, K. M., Kanjee, A., & Clauser, B. E. (1995). Using logistic regression and the Mantel-Haenszel with multiple ability estimates to detect differential item functioning. *Journal of Educational Measurement, 32*, 131-144.

Principles for Fair Student Assessment Practices for Education in Canada. (1993). Edmonton, AB: Joint Advisory Committee.

Reckase, M. D. (1997). A linear logistic multidimensional model for dichotomous item response data. In W. J. van der Linden and R. K. Hambleton (Eds.) *Handbook of modern item response theory.* New York: Springer.

Roussos, L., & Stout, W. (1996a). A multidimensionality-based DIF analysis paradigm. *Applied Psychological Measurement, 20*(4), 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 Measuremnt, 33*, 215-230.

Shealy, R., & Stout, W. F. (1993a). 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.

Shealy, R., & Stout, W. F. (1993b). An item response theory model for test bias and differential test functioning. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 281-315). Hillsdale NJ: Erlbaum.

Stout, W. (2002). Psychometrics: From practice to theory and back. *Psychometrika. Vol. 67*(4), 485-518.

Stout, W., Li, H., Nandakumar, R., & Bolt, D. (1997). MULTISIB: A procedure to investigate DIF when a test is intentionally two-dimensional. *Applied Psychological Measurement. Vol 21*(3), 195-213.

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