

Evaluating the Performance of SIBTEST and MULTISIB Using Different Matching

Criteria

Jiawen Zhou, Mark J. Gierl, Xuan Tan

University of Alberta

Introduction

Fairness in testing is a broad topic of importance which has received considerable attention in educational and psychological testing over the past decade. Concerns for differential item functioning (DIF) analyses reflect a response to the need to fairly and equitably assess examinees without bias (Standards for Educational and Psychological Testing, 1999). DIF is present when examinees with equal ability but belonging to different groups have differing probabilities of answering an item correctly. The multidimensional model for DIF (MMD, Roussos & Stout, 1996a; Shealy & Stout, 1993a) provides a definition of the cause of DIF, that is, DIF is produced by the presence of multidimensionality.

Among the various DIF detection methodologies available, some employ number-correct score (NC) as the matching criterion, e. g., Simultaneous Item Bias Test (SIBTEST) (Shealy & Stout, 1993b). The NC score used in DIF analyses are the total test score minus the studied item(s) score. It is reasonable to use NC score as the matching criterion if the assumption of unidimensionality is tenable. In practice, according to the test specification, some tests are intentionally developed to be multidimensional, which means these tests are supposed to measure multiple traits (Clauser, Nungester, Mazor, & Ripkey, 1996). 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 intended to be measured by the test before they are compared (Mazor, Hambleton, & Clauser, 1998; Clauser et al., 1996).

MULTISIB, a direct extension of SIBTEST DIF analysis procedure proposed by Stout, Li, Nandakumar, and Bolt in 1997, is a DIF detection program which can match examinees on different primary dimensions for multidimensional tests. The purpose of present study was to evaluate and compare the performance of SIBTEST and MULTISIB with respect to Type I error and power rates for a multidimensional test.

Dimensionality and the Matching Criterion in DIF Detection

Shealy and Stout (1993a) presented an elaborate in-depth theoretical description of MMD which provided a rigorous framework for understanding how DIF occurs. DIF is present only when examinees with equal ability but belonging to *reference group* and *focal group* have differing probabilities of answering an item correctly. Generally, the focal group is a minority group whereas the reference group is a majority group. The term dimension should be defined to clarify the underlying cause of DIF. According to Shealy & Stout (1993a), *dimension* refers to any substantive characteristic of an item that can affect the probability of correctly answering the item. MMD provides a definition of the cause of DIF; that is, DIF is produced in the presence of multidimensionality. Typically, DIF is attributed to multidimensionality because the dimension(s) not intended to measure in 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, 1993b). The dimension(s) that a test is designed to measure is the *primary dimension(s)* of the test. The additional potential DIF-causing dimension(s) is referred to as *secondary dimension(s)*.

An appropriate criterion which matches examinees with equal ability is an essential step for DIF analysis. Results from both real data studies (Clauser, Mazor, &

Hambleton, 1991; Mazor, Kanjee, & Clauser, 1995) and simulated data studies (Ackerman, 1992; Mazor, Hambleton, & Clauser, 1998) have highlighted the importance of the choice of matching criterion for DIF analysis. Clauser et al. (1991), for instance, indicated the benefits of conditioning on single subtest score instead of total test score. 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 DIF analysis. More recent studies have concluded that conditioning on the multiple ability estimates simultaneously for an intentional multidimensional test led to the outcome that substantially fewer items were detected to exhibit DIF (Clauser et al., 1996; Mazor et al., 1998; Mazor et al., 1995). This perspective emphasized that failure to condition on multiple valid dimensions influencing item responses may allow intended multidimensional item impact to be falsely identified as DIF.

Ample studies have also been conducted to evaluate the performance of Mantel-Haenszel (MH) and logistic regression (LR, Swaminathan & Rogers, 1990) using different matching criteria (Clauser et al., 1996; Mazor et al., 1998; Mazor et al., 1995). However, the performance comparison between SIBTEST, and its multidimensional version, MULTISIB, is lacking. Thus, the present study will compare these two procedures using simulation study. The impact of two independent variables, correlation between primary dimensions and sample size, will be evaluated by assessing the Type I error and power rates using different matching criteria for the SIBTEST and MULTISIB procedures.

Overview of SIBTEST and MULTISIB

SIBTEST (Shealy & Stout, 1993b), is designed to statistically test DIF hypotheses and identify items in the secondary dimension that produces group differences. It can also be used to estimate the amount of DIF. The items in the matching subtest are assumed to measure an unidimensional trait, θ , which is the primary dimension in the test. Items in the studied subtest are assumed to measure additional dimension(s) that impact examinees' item responses called secondary dimension(s), which is denoted by η .

Matching subtest scores are used to place the reference and focal group examinees into subgroups at distinct score level so they are judged approximately equivalent on the intended primary dimension. Therefore, their performances on a studied item can be compared. In the case of a single-item DIF analysis, only one item is included in the studied subtest and the matching subtest contains the remaining test items. To operationalize SIBTEST, the statistical hypothesis tested is

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

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, respectively. $f_F(\Theta)$ 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, N denotes the total number of items in a study test, items 1, ..., n denotes the matching subtest items, and $n+1, \dots, N$ denotes the studied subtest items. Let U_i denote the response to item i scored as 0 or 1. For each examinee,

$$X = \sum_{i=0}^n U_i \text{ to specify the total score on the matching subtest and } Y = \sum_{i=n+1}^N U_i \text{ to specify}$$

the total score on the studied subtest. The matching subgroups are indexed by total score k on the matching subtest. Examinees in the reference and focal groups are then grouped into 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

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

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. 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, respectively, 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 (1993b).

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

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

where $\widehat{\sigma}(\widehat{\beta}_{UNI})$ is the estimated standard error of $\widehat{\beta}_{UNI}$. Shealy and Stout (1993b) 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.

In addition, $\widehat{\beta}_{UNI}$ 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. Roussos and Stout (1996b) proposed guidelines for interpreting DIF by combining the SIBTEST statistical results with values for the $\widehat{\beta}_{UNI}$ parameter estimate to classify DIF on a single item: (a) negligible DIF: absolute value of $\widehat{\beta}_{UNI} < 0.059$ and $H_0 : B = 0$ is rejected, (b) moderate DIF: absolute value of $0.059 \leq \widehat{\beta}_{UNI} < 0.088$ and $H_0 : B = 0$ is rejected, (c) large DIF: absolute value of $\widehat{\beta}_{UNI} > 0.088$ and $H_0 : B = 0$ is rejected.

Overview of MULTISIB DIF Detection Procedure

A multidimensional test consists of items designed to measure more than one primary dimension. Each item in a multidimensional test can be designed to tap one or all of the primary dimensions. DIF is assumed to occur when examinees, matched on the intended dimensions, perform differentially depending on their group membership. That is, in a multidimensional test an item is a DIF causing item if it measures the intended primary dimensions and one or more secondary dimensions differentially impacting the performance of examinees in one of the two groups.

As the multidimensional counterpart of SIBTEST, the same statistical hypothesis is tested by MULTISIB. MULTISIB is also designed to identify items evaluating the secondary dimension(s) and estimate the magnitude of DIF for two-dimensional tests. The guidelines for interpreting DIF proposed by Roussos and Stout (1996b) are applied to MULTISIB in the current study. The basic logic in MULTISIB 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 should response in a similar manner on a two-dimensional test. N denotes the total number of items in a two-dimensional test. The items 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 . 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 Matching 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 develop 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.

Method

A simulation study was conducted to evaluate and compare the performance of SIBTEST and MULTISIB to detect DIF with two-dimensional test data, in terms of Type I error and power rates. 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 dimension ($\rho_{12} = 0.20, 0.40, 0.60, \text{ and } 0.80$). The levels of each factor were designed to reflect those that might be found in real test data. Test length was consistent: 70-item tests with 50 matching items and 20 studied items.

Sample Size

Previous research indicated that sample size does affect DIF items detection (Stout, Li, Nandakumar, & Bolt, 1997; Gierl, Gotzmann, & Boughton, 2004). In actual testing situations, sample size deserves attention because it can vary dramatically in applied settings and thus affect the performance of DIF detection procedures. Therefore, to explore the effect of sample size on DIF detection rates, it 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 in both the reference and focal groups. 500 is a relative small sample size while 2000 is a large one. The systematical increment of 500 makes it easy to

explain the differences between each condition. 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 another factor considered. This variable was evaluated as the primary dimensions can be perceived as one dimension when their correlation is high, making the benefit of matching on different primary dimensions negligible. Four levels of this factor were manipulated, 0.20, 0.40, 0.60, and 0.80. The four levels with an increment of 0.20 were evaluated in present study because a zero correlation between primary dimensions is unrealistic. 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 distinct while the large correlation $\rho_{12} = 0.80$ denoted that the two primary dimensions are very similar. The interaction of the two manipulated elements, sample size and the correlation between two primary dimensions, was also of interest in present study.

Thus, the item response data for DIF analyses were crossed 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.

Date Generation and Analysis

The examinee item responses to the 70 items were simulated using the compensatory multidimensional item response theory (MIRT) model (Reckase, 1997). The 3PL item response function (IRF) for the compensatory MIRT model can be presented as:

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

where U_i is the response to item i , $\vec{\theta}^T = (\theta_1, \dots, \theta_k)$ is the vector of examinee ability, $\vec{a}_i^T = (a_{i1}, \dots, a_{ik})$ 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. Examinee ability was 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 measuring primary dimension θ_1 and 25 items measuring 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 slopes 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 reversed 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 in

that the angular directions were both within 20° 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 plot of the items measuring dimension θ_1 and θ_2 is shown in Figure 1.

The other 20 items in the simulated test were studied items which were intended to test the DIF detection rates of SIBTEST and MULTISIB. The first eight of these 20 studied items were non-DIF items which only measured the two primary dimensions. These items do not measure the 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 non-DIF items were used to calculate the Type I error rates on both SIBTEST and MULTISIB. The d -parameter of the eight 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 mainly measured one of the two primary dimensions and a secondary 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 a_3 -parameters were restricted in the range of 0.30 to 1.55, which were the potential causes of 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 1.00 by Stout et al. (1997). However, as they mentioned, when the mean difference between groups equals 1, it is typically the largest value obtained in real testing situations. Thus, in the 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 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 group, were 0.20. Table 3 and Table 4 contain the item parameters and the angular directions of the 12 DIF items for reference and focal group, respectively.

The computer programs SIBTEST and MULTISIB were used for the DIF analyses with the simulated data sets. The guidelines for interpreting DIF by Roussos and Stout (1996b) are used to classify DIF items in the present study. Two-tailed hypothesis tests were conducted for all analyses using an alpha level of 0.05 in present study. Two types of DIF detection rates were assessed in this study. Type I error occurred when $H_0 : B = 0$ of a non-DIF item was incorrectly rejected. Conversely, power occurred if a DIF item was correctly identified where $H_0 : B = 0$ was rejected. Furthermore, the identified DIF items by both procedures, correctly or incorrectly, were flagged using the conventions for negligible, moderate, or large DIF. Also, the proportion of correct classification of the non-DIF items and different magnitudes DIF items was explored in this study.

Results

The results of the simulation study are presented for the 0.20, 0.40, 0.60, and 0.80 correlation conditions in Table 5 and 6, which contains Type I error rates and power rates

respectively. In each condition, results are displayed based on the increase of the sample size variable. Table 7 to Table 10 contains 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 as classified by the SIBTEST and MULTISIB procedures. The proportions of correct classification for each correlation and the corresponding sample size condition are listed in the four tables. The comparisons of SIBTEST and MULTISIB across correlation and sample size conditions are illustrated in Figure 2 to 5.

Table 5 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. With small correlation rate such as $\rho_{12} = 0.20$, the Type I error rates for SIBTEST decreased from 0.38% to 0.00% as the sample size increased from 500 to 2,000. Similarly, with the increase of the sample size, the Type I error rates decreased from 0.75% to 0.00% with 0.40 correlation condition and decreased from 0.63% to 0.00% with 0.60 correlation condition. However, with a large correlation condition, 0.80, the Type I error rates varied across four sample sizes, making it difficult to evaluate the trend.

The Type I error rates on MULTISIB decreased as the sample size increased across all correlation conditions. Moreover, the incorrect detection rates for MULTISIB were greater than or equal to SIBTEST across all correlation with only a small number of exceptions. Figure 2 contains the Type I error rates comparison on SIBTEST and MULTISIB across the four correlation conditions.

Table 6 presents the power rates for SIBTEST and MULTISIB with the four levels of correlation. The power rates on SIBTEST consistently increased as sample size

increased across all correlation conditions. Similar results were found in the MULTISIB cases. The power rates on MULTISIB increased when the sample size increased (except for the 1,500 examinee per group condition where the correct DIF detection rate was slightly smaller compared to the 1,000 examinee per group condition in 0.20, 0.40, and 0.60 correlation cases). Furthermore, the power rates on MULTISIB of each level of sample size were smaller compared to SIBTEST in all correlation conditions. Figure 3 contains the power rates comparison on SIBTEST and MULTISIB across the four correlation conditions.

Please note that, when sample size was fixed, correlation between dimensions did not show a consistent influence on the Type I error rates as shown in Figure 4. For example, the Type I error rate on SIBTEST with 500 examinee per group was 0.38% for the 0.20 correlation condition, 0.75% for 0.40, 0.63% for 0.60, and 0.50% for the 0.80 correlation condition, respectively. Similarly, the power rates on MULTISIB varied unsystematically across the four correlation conditions with 1,000 and 2,000 sample size conditions. However, the power rates on SIBTEST decreased as the correlation between dimensions systematically increased as the sample size was 1,000 and larger (see Figure 5).

Table 7 to 10 contains the classification of the 20 study items across the two DIF detection procedures with four levels of correlation. Similar results were found for both SIBTEST and MULTISIB: the eight non-DIF items were correctly classified by both SIBTEST and MULTISIB, as the low Type I error rates indicates. The negligible DIF items were primarily grouped as non-DIF items when the sample size was 500 while the proportion of correct grouping for the negligible DIF items increased as the sample size

increased. The proportion of correct classifications for moderate and large DIF items increased when the sample size increased for both SIBTEST and MULTISIB. Note, however, that the proportions of correct classification for both non-DIF and DIF items were consistently larger with SIBTEST compared to MULTISIB across each sample size condition.

Conclusions and Discussion

The influences of matching criterion on the performance of DIF analysis procedure have been raised by previous research. The evaluation of the corresponding DIF detection methods for multidimensional tests is therefore required. MULTISIB was, thus, evaluated and compared with its unidimensional version, SIBTEST, for an intentional multidimensional test in this study. Two independent variables, correlation between primary dimensions and sample size, were included in the simulation design to investigate the impact of the two factors to the performance of SIBTEST and MULTISIB. The Type I error rates and power rates generated by SIBTEST and MULTISIB were evaluated and compared.

All Type I error rates for both SIBTEST and MULTISIB in this study were less than 5%, which was acceptable. The eight simulated non-DIF items were correctly identified by both procedures. Type I error rates systematically decreased as the sample size increased in all correlation conditions (with one exception as $\rho_{12} = 0.80$ for SIBTEST) for both procedures indicating that sample size is an important factor affecting detection rates. In contrast, in each sample size condition, no systematical variation based on increasing correlation could be found across the two procedures. We concluded that

the factor correlation between primary dimensions did not affect the performance of the two procedures.

In addition, Type I error rates for MULTISIB were greater than or equal to SIBTEST across all correlation conditions with several exceptions. Though these rates for both procedures were acceptable, incorrect detection rates of MULTISIB were expected to be less than SIBTEST. The NC score used in SIBTEST was assumed to be less effective for matching examinees on two distinct dimensions simultaneously, especially with a low correlation between two primary dimensions, and therefore perform more poorly as a matching criterion compared to MULTISIB. However, the comparison outcomes revealed that the multidimensional matching criterion did not positively influence the performance, which was an unexpected result.

The power rates (except two power rates on 0.20 and 0.40 correlation crossed with 1,500 sample size conditions) for both SIBTEST and MULTISIB were acceptable, which were greater than 80.00% when the sample size was 1,000 and larger. Sample size was an important factor that positively affected the power rates. The power rates for SIBTEST systematically increased as the sample size increased across the four correlation conditions. Similar results were found for MULTISIB with several exceptions. However, with the increase of the correlation, there was no systematical variation on power rates for MULTISIB in some sample size conditions. Moreover, the power rates, either acceptable or not, for SIBTEST were consistently greater than those for MULTISIB across all conditions. This result was, again, not as anticipated.

This study also evaluated the proportion of correct and incorrect classification of non-DIF and DIF items by SIBTEST and MULTISIB. Non-DIF items were correctly

classified across all the conditions while negligible DIF items were sensitive to small sample size. Negligible DIF items had a greater chance at being identified as non-DIF items when sample size was 500. With the increase in the sample size, the correct classification of DIF items on all levels increased. Consistent with the detection rates, the proportion of correct classifications of both non-DIF and DIF items was greater with SIBTEST compared to MULTISIB across sample size conditions.

Given the results of this study, sample size is a factor that positively influences the performance of SIBTEST and MULTISIB while the correlation between primary dimensions is not. Both procedures performed well for DIF detection when sample size was 1,000 and larger. The multidimensional matching criterion did not impact the two procedures as was expected. We conclude that SIBTEST provided superior DIF detection results compared to MULTISIB for intentional multidimensional tests.

Future Directions

The results from this study provide researchers and practitioners with some insights into the detection rates for SBITEST and MULTISIB where the test is designed to be multidimensional. However, only simulated data were analyzed in this study. The item parameters used for simulation were systematically manipulated, thus results obtained should be generalized to realistic testing situations with caution. More research is needed to evaluate SIBTEST and MULTISIB using item parameters derived from real tests. This expansion will, in turn, make the study more generalizable to real testing situations. Supplementing simulation studies with real data analysis has become more and more common. Due to limited resources available, this study was not supplemented by a real data analysis. However, future efforts should be made to incorporate real test data

into simulation procedures using items that exhibit two clear dimensions that, in turn, can be split to two subtests. Adopting realistic item parameters as well as conducting real data analysis will help make future studies more generalizable.

The guessing parameter was set fixed in this simulation study. That is, the influence of this factor to both SIBTEST and MULTISIB was not explored. Further study can treat this issue as one factor to see if it has differential influence on the performance of SIBTEST and MULTISIB.

Other DIF detection methods can also be used by matching on multiple subtest scores, such as MH and LR. Comparisons between MULTISIB and these DIF detection procedures are lacking. Hence, future research using alternative DIF procedures is also promising.

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.
- Ackerman, T. A., & Evans, J. A. (1993). *A didactic example of conditioning on the complete latent ability space when performing DIF analyses*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Atlanta, GA.
- Camilli, G., & Shepard, L. (1994). *Methods for identifying biased test items*. Newbury Park, CA: Sage.
- 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., Bisanz, J., Bisanz, G., Boughton, K., & Khaliq, S. (2001). Illustrating the utility of differential bundle functioning analyses to identify and interpret group differences on achievement test. *Educational Measurement: Issues and Practice*, 20, 26-36.
- Gierl, m. J., & Khaliq, S. N. (2001). Identifying sources of differential item and bundle functioning on translated achievement tests: a confirmatory analysis. *Journal of Educational Measurement*. Vol. 38(2), 164-187.
- 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.
- Reckase, M. D. (1997). A linear logistic multidimensional model for dichotomous item response data. In van der Linder, W. J., & Hambleton, R. K. (Eds.), *Handbook of Modern Item Response Theory*, 271-286. New York: Springer-Verlag.
- 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 samle 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). An item response theory model for test bias. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 281-315). Hillsdale NJ: Erlbaum.
- Shealy, R., & Stout, W. F. (1993b). 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.
- Standards for Educational and Psychological Testing*. (1999). Washington, DC: American Educational Research Association, American Psychological Association, National Council on Measurement in Education.
- 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.

Table 1

Item Parameters and Angular Direction of 50 Matching Items

	a₁	a₂	a₃	d	c	MDISC	D	α
1	0.35	0.05	0.00	-1.00	0.20	0.35	2.83	8.13°
2	0.65	0.05	0.00	-1.00	0.20	0.65	1.53	4.40°
3	0.95	0.05	0.00	-1.00	0.20	0.95	1.05	3.01°
4	1.25	0.10	0.00	-1.00	0.20	1.25	0.80	4.57°
5	1.55	0.10	0.00	-1.00	0.20	1.55	0.64	3.69°
6	0.35	0.10	0.00	-0.50	0.20	0.36	1.37	15.95°
7	0.65	0.08	0.00	-0.50	0.20	0.65	0.76	7.02°
8	0.95	0.10	0.00	-0.50	0.20	0.96	0.52	6.01°
9	1.25	0.10	0.00	-0.50	0.20	1.25	0.40	4.57°
10	1.55	0.15	0.00	-0.50	0.20	1.56	0.32	5.53°
11	0.35	0.05	0.00	0.00	0.20	0.35	0.00	8.13°
12	0.65	0.20	0.00	0.00	0.20	0.68	0.00	17.10°
13	0.95	0.10	0.00	0.00	0.20	0.96	0.00	6.01°
14	1.25	0.12	0.00	0.00	0.20	1.26	0.00	5.48°
15	1.55	0.15	0.00	0.00	0.20	1.56	0.00	5.53°
16	0.35	0.05	0.00	0.50	0.20	0.35	-1.41	8.13°
17	0.65	0.15	0.00	0.50	0.20	0.67	-0.75	12.99°
18	0.95	0.30	0.00	0.50	0.20	1.00	-0.50	17.53°
19	1.25	0.12	0.00	0.50	0.20	1.26	-0.40	5.48°
20	1.55	0.05	0.00	0.50	0.20	1.55	-0.32	1.85°
21	0.35	0.08	0.00	1.00	0.20	0.36	-2.79	12.88°
22	0.65	0.10	0.00	1.00	0.20	0.66	-1.52	8.75°
23	0.95	0.05	0.00	1.00	0.20	0.95	-1.05	3.01°
24	1.25	0.15	0.00	1.00	0.20	1.26	-0.79	6.84°
25	1.55	0.30	0.00	1.00	0.20	1.58	-0.63	10.95°
26	0.05	0.35	0.00	-1.00	0.20	0.35	2.83	81.87°
27	0.05	0.65	0.00	-1.00	0.20	0.65	1.54	85.60°
28	0.05	0.95	0.00	-1.00	0.20	0.95	1.05	86.99°
29	0.05	1.25	0.00	-1.00	0.20	1.25	0.80	87.71°
30	0.05	1.55	0.00	-1.00	0.20	1.55	0.64	88.15°
31	0.05	0.35	0.00	-0.50	0.20	0.35	1.41	81.87°
32	0.10	0.65	0.00	-0.50	0.20	0.66	0.76	81.25°
33	0.05	0.95	0.00	-0.50	0.20	0.95	0.53	86.99°
34	0.10	1.25	0.00	-0.50	0.20	1.25	0.40	85.43°
35	0.10	1.55	0.00	-0.50	0.20	1.55	0.32	86.31°
36	0.10	0.35	0.00	0.00	0.20	0.36	0.00	74.05°
37	0.10	0.65	0.00	0.00	0.20	0.66	0.00	81.25°
38	0.15	0.95	0.00	0.00	0.20	0.96	0.00	81.03°
39	0.10	1.25	0.00	0.00	0.20	1.25	0.00	85.43°
40	0.15	1.55	0.00	0.00	0.20	1.56	0.00	84.47°
41	0.05	0.35	0.00	0.50	0.20	0.35	-1.41	81.87°
42	0.15	0.65	0.00	0.50	0.20	0.67	-0.75	77.01°
43	0.20	0.95	0.00	0.50	0.20	0.97	-0.52	78.11°
44	0.09	1.25	0.00	0.50	0.20	1.25	-0.40	85.88°
45	0.15	1.55	0.00	0.50	0.20	1.56	-0.32	84.47°
46	0.05	0.35	0.00	1.00	0.20	0.35	-2.83	81.87°
47	0.10	0.65	0.00	1.00	0.20	0.66	-1.52	81.25°
48	0.05	0.95	0.00	1.00	0.20	0.95	-1.05	86.99°
49	0.10	1.25	0.00	1.00	0.20	1.25	-0.80	85.43°
50	0.09	1.55	0.00	1.00	0.20	1.55	-0.64	86.68°

Table 2

Item Parameters and Angular Direction of Eight Non-DIF Items

	a₁	a₂	a₃	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°

Table 3

Item Parameters and Angular Direction of 12 DIF Items for Reference Group

	a₁	a₂	a₃	d	c	MDISC	D	α
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°

Table 4

Item Parameters and Angular Direction of 12 DIF Items for Focal Group

	a₁	a₂	a₃	d	c	MDISC	D	α
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°

Table 5

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

Table 6

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

Table 7

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

Correlation Between Dimensions	Sample Size	Designed study items	SIBTEST				MULTISIB			
			non-DIF	Negligible DIF	Moderate DIF	Large DIF	non-DIF	Negligible DIF	Moderate DIF	Large DIF
0.2	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%

Table 8

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

Correlation Between Dimensions	Sample Size	Designed study items	SIBTEST				MULTISIB			
			non-DIF	Negligible DIF	Moderate DIF	Large DIF	non-DIF	Negligible DIF	Moderate DIF	Large DIF
0.4	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%

Table 9

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

Correlation Between Dimensions	Sample Size	Designed study items	SIBTEST				MULTISIB			
			non-DIF	Negligible DIF	Moderate DIF	Large DIF	non-DIF	Negligible DIF	Moderate DIF	Large DIF
0.6	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%

Table 10

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

Correlation Between Dimensions	Sample Size	Designed study items	SIBTEST				MULTISIB			
			non-DIF	Negligible DIF	Moderate DIF	Large DIF	non-DIF	Negligible DIF	Moderate DIF	Large DIF
0.8	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%

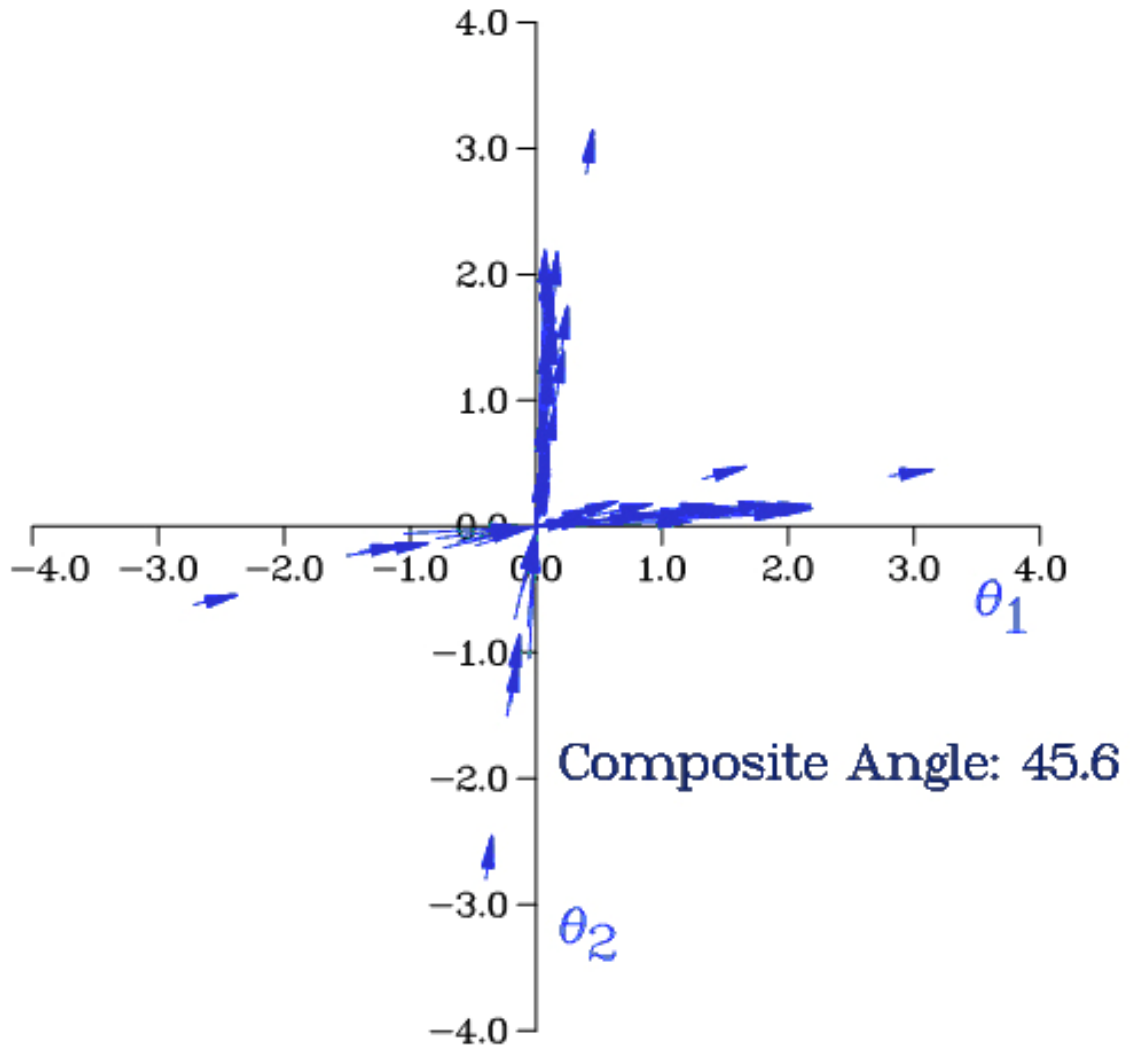


Figure 1. Vector plot of simulated items measuring primary dimension 1 and 2.

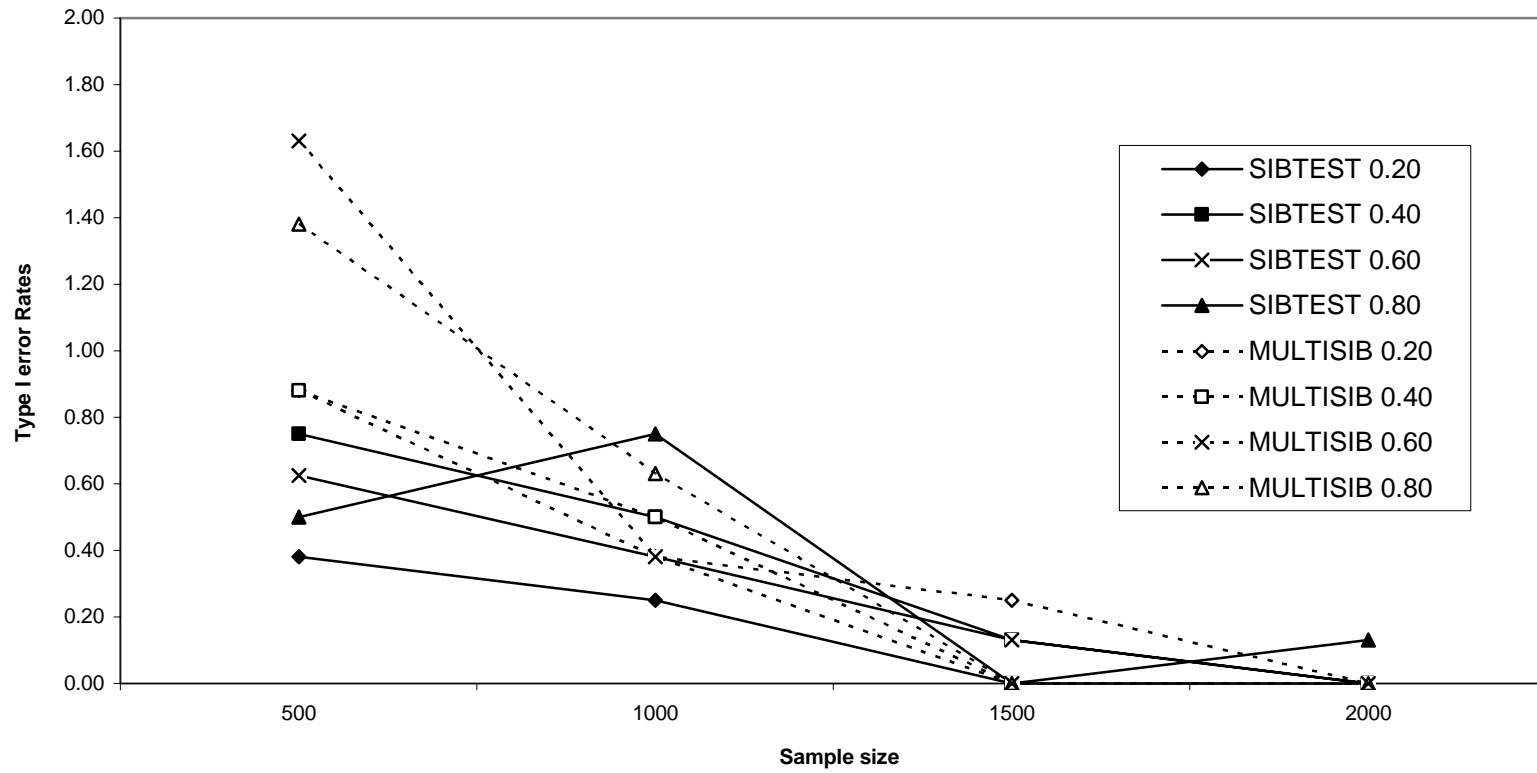


Figure 2. Type I error rates in each correlation condition cross different sample sizes for SIBTEST and MULTISIB.

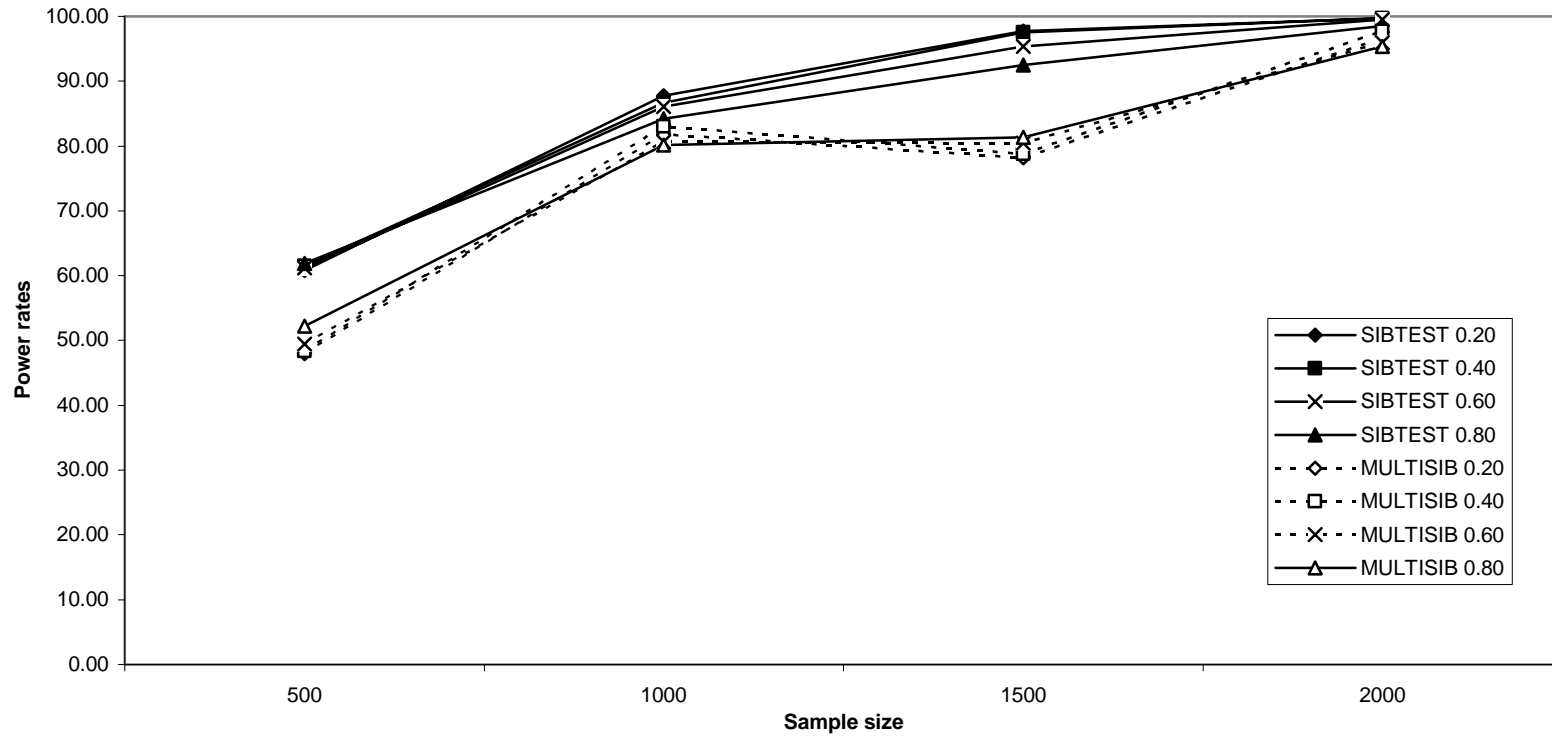


Figure 3. Power rates in each correlation condition cross different sample sizes for SIBTEST and MULTISIB.

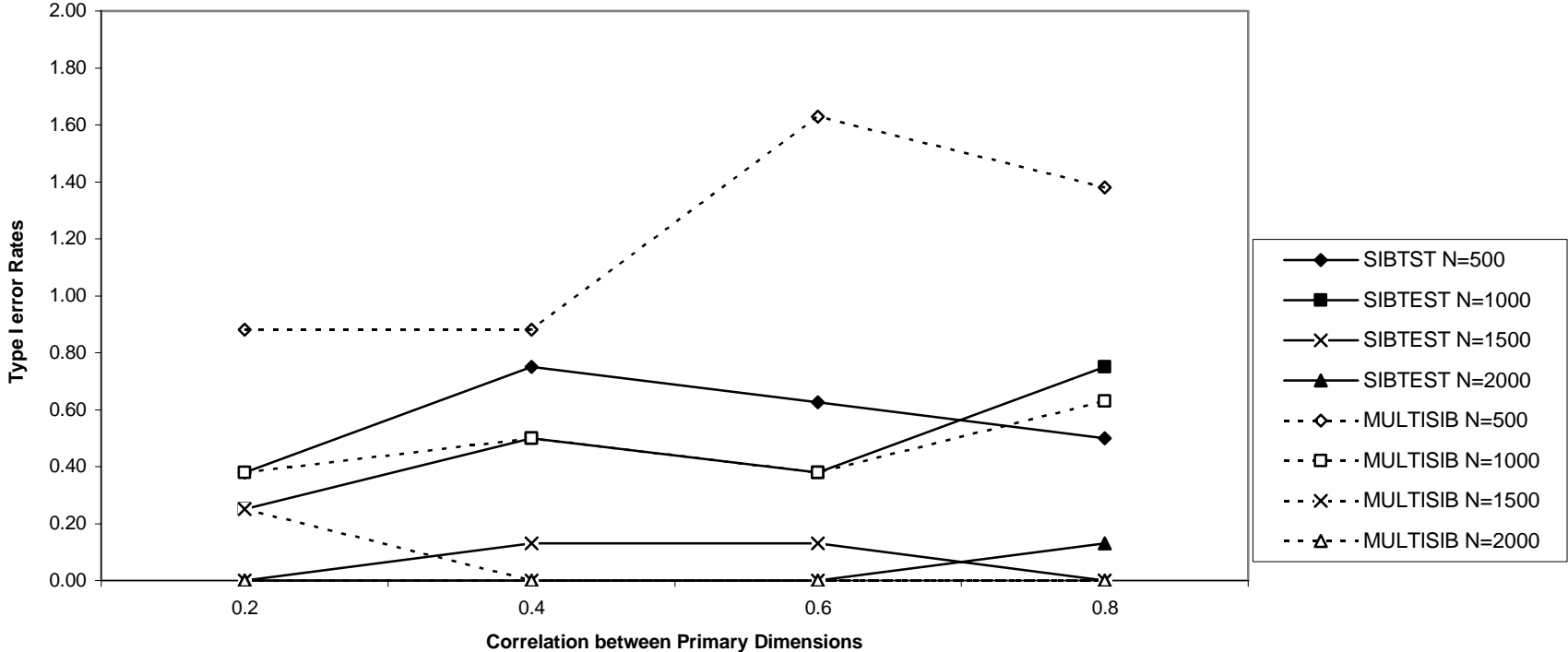


Figure 4. Type I error rates with each sample size condition cross different correlation for SIBTEST and MULTISIB.

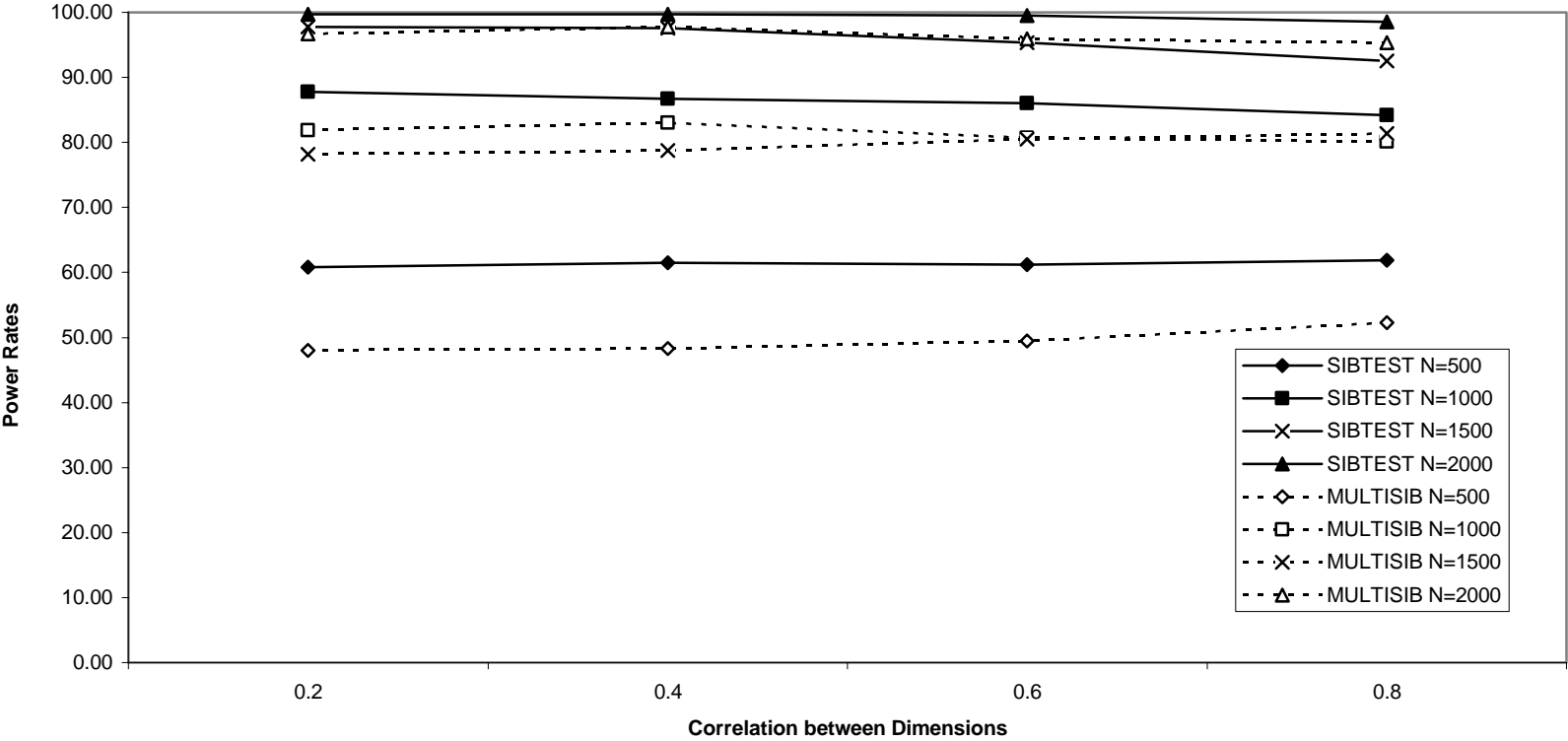


Figure 5. Power rates in each sample size condition cross different correlation for SIBTEST and MULTISIB.