Applied Psychological Measurement
Volume 33 Number 6
September 2009 441-464
© 2009 SAGE Publications
10.1177/0146621609331959
http://apm.sagepub.com
hosted at
http://online.sagepub.com

The Impact of Multidimensionality on the Detection of Differential Bundle Functioning Using Simultaneous Item Bias Test

Carolyn F. Furlow
Terris Raiford Ross
Phill Gagné
Georgia State University

Douglas, Roussos, and Stout introduced the concept of differential bundle functioning (DBF) for identifying the underlying causes of differential item functioning (DIF). In this study, reference group was simulated to have higher mean ability than the focal group on a nuisance dimension, resulting in DIF for each of the multidimensional items that, when examined together, produced DBF. The empirical power and the Type I error of the Simultaneous Item Bias Test for DBF analysis were examined under various sample sizes, ratios of reference to focal group sizes, correlations between target and nuisance dimensions, magnitudes of DIF/DBF, test lengths, percentages of test items in the bundle, and item discriminations. Power was generally high in cells with larger DIF magnitudes, higher percentages of items in the bundle, larger sample sizes, and with the nuisance dimension having a higher discrimination than the target dimension. Type I error rates approximated the nominal alpha rate for all conditions.

Keywords: *differential item functioning; multidimensionality; differential bundle functioning*

The empirical evidence gathered in the investigation of bias is generally referred to as differential item functioning (DIF). In response to public concern on the bias that exists in some measures of aptitude and/or cognitive ability (e.g., Scholastic Aptitude Test (SAT), intelligence tests, and exams for licensure and promotion), conducting DIF analyses is now standard practice for most large-scale testing companies. The development of statistical and substantive methods of investigating DIF is crucial to the goal of designing fair and valid educational and psychological tests.

Shealy-Stout's Multidimensional Model for DIF (MMD)

During the last decade or note so, researchers have argued that the leading cause of DIF is the inclusion of multidimensional test items. That is, many tests thought to be unidimensional—an important assumption in item response theory (IRT)—are in fact

Authors' Note: Please address correspondence to Carolyn F. Furlow, College of Education, Georgia State University, P.O. Box 3977, Atlanta, GA 30302-3977; e-mail: cfurlow@gsu.edu.

measuring additional latent traits other than the trait of interest (e.g., Oshima & Miller, 1992; Roussos & Stout, 1996a; Russell, 2005; Shealy & Stout, 1993a). These other dimensions represent either intentionally or unintentionally measured traits. According to Shealy and Stout's (1993a) multidimensional model for DIF (MMD), if an additional dimension is unintentionally assessed as part of the construct of interest (e.g., verbal ability necessary for an item designed to measure math ability), then it is termed *nuisance*, but if the test is designed to measure two traits (e.g., math ability and logical reasoning), then the second dimension may be considered auxiliary. Under MMD, DIF that results from auxiliary dimensions is benign (indicating impact), whereas nuisance traits produce adverse DIF (indication of bias). It is also important to note the difference between impact and bias. *Impact* is a reflection of true ability differences between groups on a relevant or intentional construct; *bias* is a term reserved for items that measure an unintended trait for which examinees from one group are systematically advantaged or disadvantaged.

Violations of Unidimensionality Assumption

Many experts agree that most achievement and aptitude tests are actually multidimensional, with perhaps a dominant target dimension and other minor dimensions (e.g., Ackerman, 1992; Camilli & Shepard, 1994; Embretson & Reise, 2000; Lord, 1980; Shealy & Stout, 1993a). Hence, the use of unidimensional IRT models with multidimensional test data violates the unidimensionality assumption and poses a potentially serious threat to item and examinee parameter estimation.

Many studies (e.g., Ansley & Forsyth, 1985; Kirisci, Hsu, & Yu, 2001; Reckase, 1979; Reckase, Ackerman, & Carlson, 1988) have assessed the effects of this violation on measurement equivalence, and the results have been used to both support the continued use of unidimensional IRT and encourage the development of multidimensional IRT (MIRT) models as well. Kirisci et al. (2001) examined the effects of multidimensionality and IRT calibration programs (i.e., XCALIBRE, BILOG, and MULTILOG) on the accuracy of item and ability parameter estimates. They observed an interaction between dimensionality and estimation programs, with BILOG yielding the smallest RMSE for most of the study conditions. XCALIBRE and MULTILOG, however, were associated with less variance in parameter estimates. These findings, combined with the results of previous research, served as a basis for several guidelines offered by Kirisci et al. for practitioners wanting to use unidimensional IRT models to estimate multidimensional test parameters.

Kirisci et al. (2001) made three suggestions. First, it is necessary to assess the multidimensional structure of the data because applying unidimensional IRT models may be permissible if there is only one dominant dimension with several minor dimensions. Second, when there are several dimensions of approximately equal dominance, the magnitude of the correlations between dimensions should be assessed. Citing earlier studies (e.g., Ackerman, 1989; Drasgow & Parsons, 1983; Harrison, 1986), Kirisci et al. suggested proceeding with unidimensional IRT models if the multiple dimensions are highly correlated ($r > .4$). Finally, in the event that dimensions are not highly correlated ($r \leq .4$) and/or the correlations among them vary to a large degree, then MIRT should be applied. Although multidimensionality can seriously affect the performance of unidimensionally based procedures, such as BILOG and LOGIST, these studies provide evidence that correlated dimensions can mediate that influence, making such procedures adequate for item and ability estimation.

Additionally, results from those investigations have corroborated Stout's (1987, 1990) theory and test of *essential unidimensionality* that challenges the traditional unidimensionality assumption of IRT. Stout argued that the presence of exactly one dominant dimension (i.e., essential unidimensionality or $d_{\mathrm{E}} = 1$) even in the presence of other unintentional dimensions is a sufficient and more psychologically appropriate requirement, considering the nature of achievement testing. If unintended dimensions influence item responses to a large degree, however, Stout's test of essential unidimensionality should conclude that $d_{\mathrm{E}} > 1$, and practitioners would be advised against the use of standard IRT data analysis programs (Nandakumar, 1991).

Traditional DIF Analysis

DIF occurs when focal and reference group members of equal ability on the latent trait of interest have different probabilities of answering an item correctly. In the event an item exhibits DIF, a decision must be made about whether to retain the item or delete it from the test. If the item seems to be biased against a subgroup, if the DIF magnitude is strong enough to bias test results, and if a rationale exists for why it may exhibit DIF, then the item should be deleted (Camilli & Shepard, 1994). Without a substantive review of the item to understand the reason it resulted in DIF, however, test developers do not actually know if the source of DIF is because of a construct-relevant or construct-irrelevant dimension being measured by the test (or perhaps just chance). Therefore, it is important to thoughtfully interpret the nature of DIF so that differences between the groups' cognitive skills or opportunities to learn can be appropriately addressed.

DIF analyses are typically conducted in two steps: (a) statistical identification of items that favor particular groups (including effect size measures of practical significance) followed by (b) a substantive review of potentially biased items to locate the sources of DIF. Although many advances have been made in the statistical analysis of DIF items, much remains to be learned about how to pinpoint the reason that DIF occurs. During the substantive analysis of DIF, items are usually reviewed by subject-area experts (e.g., curriculum specialists or item writers) in an attempt to interpret the factors contributing to differential performance between specific subgroups of examinees. Even though this is an important step in the process of eliminating bias and ensuring test fairness, studies indicate that this method of substantive analysis has met with limited success (see Camilli & Shepard, 1994; Engelhard, Hansche, & Rutledge, 1990; Gierl, Rogers, & Klinger, 1999; O'Neill & McPeek, 1993; Roussos & Stout, 1996a; *Standards for Educational and Psychological Testing*, 1999; Sudweeks & Tolman, 1993).

Remarking on the task of predicting DIF items without empirical evidence, Engelhard et al. (1990) state that "the agreement between the judgmental and empirical indices of DIF are very low and usually not better than what would be expected by chance" (p. 358). In many cases, the judgments made by item reviewers tend to disagree either with DIF statistics or with one another. These inconsistencies may be attributable to the many possible hypotheses about why an item displays DIF that arise *after* DIF has been identified statistically. As a result, definitive conclusions about the sources of DIF are rarely drawn. Some researchers believe that this is an inherent problem with single-item DIF analyses and argue that more can be learned from studying groups of items simultaneously rather than one at a time (Boughton, Gierl, & Khaliq, 2000; Douglas, Roussos, & Stout, 1996). Although the current primary method of DIF analysis does not typically provide more conclusive answers regarding DIF sources, a related procedure known as differential bundle functioning (DBF) holds promise for addressing this problem.

DBF

Because substantive DIF analyses following the statistical identification of DIF items have yielded little information in understanding the sources of DIF, methods have been developed that use the results of substantive analyses (item review) to investigate statistically items believed to function differentially. That is, instead of seeking to interpret DIF statistics for substantive meaning, the process (and therefore the logic) is reversed by first forming substantive hypotheses regarding potential DIF items and then testing those items statistically. In particular, Douglas et al. (1996) introduced the concept of item *bundle* DIF and the implications of DBF for identifying the underlying causes of DIF. A bundle is any dimensionally homogenous set of items that is not necessarily adjacent or related to a common text or passage (Douglas et al., 1996).

In DBF analyses, similar items are grouped together based on organizing principles (e.g., content, item type, etc.) believed to affect the performance of different groups of examinees. The basis for DBF analysis is the assertion that tests consist primarily of small bundles of items designed to measure a certain trait, skill, or ability. Similarly, Gierl, Bisanz, Bisanz, Boughton, and Khaliq (2001) assert that “sources of DIF may be more apparent in patterns across multiple items rather than in performance characteristics associated with single items” (p. 27). Also research has shown that methods for statistically identifying bundles or groups of items are more powerful than those which analyze items one at a time (Nandakumar, 1993). Douglas et al. (1996) note that the amount of DIF in a single item might be small enough that it would not be detected statistically but that small amounts can add up to an undesirable amount when they are present in several items. Hence, DBF analysis is preferable to DIF analysis when bundling permits small differences in group performance on individual items to be amplified. Furthermore, the DBF approach of examining potential sources of DIF by identifying suspect item bundles has great implications for improving test design and psychological measurement.

Bundle Formation

Although any number of organizing principles can be used to identify items suspected of measuring multiple abilities, prior studies have used four methods in particular. The methods can be classified as exploratory versus confirmatory, depending on the manner in which they are employed and/or the rationale for using a specific bundling strategy. First, a test’s dimensionality structure can be assessed using test specifications that outline both the content area and cognitive skill categories that the test is designed to measure. A list of test specifications serves as the blueprint to guide item writers when sampling items from the achievement domain. These items are also designed to measure specific cognitive skills and processes. Therefore, a detailed analysis of test specifications may highlight subsets of items that measure a number of different dimensions associated with certain content and skill areas (Gierl, Tan, & Wang, 2005).

For example, Oshima, Raju, Flowers, and Slinde (1998) demonstrated DBF analysis using the cognitive dimensions measured by the reading comprehension portion of the Metropolitan Achievement Test as well as by bundling the items associated with reading passages. Although cognitive classifications did not appear to elicit differential functioning, Oshima et al. did find large DBF in favor of boys for a reading passage titled "The Roadrunner: A Strange Bird." They were able to interpret the potential cause of the differential functioning by reasoning that boys were possibly more familiar with the context of the passage that described characteristics of the roadrunner such as, its diet and speed.

As a second method, the dimensionality structure of a test can be uncovered through the use of subject-area experts who use their experience to identify specific dimensions through a thorough analysis of test content. A content analysis may be conducted during either an item review session with content specialists or a review of the literature for judgments regarding the content of well-known tests such as the SAT (e.g., Douglas et al., 1996; Gierl & Bolt, 2003; O'Neill & McPeek, 1993). For example, Douglas et al. (1996) used a panel of experts to select item bundles from a test deemed to be essentially unidimensional. They argued that this method is especially appropriate in cases in which a test is so dominated by the target dimension that many statistical dimensionality assessment tools would be unable to detect minor unintended dimensions.

Third, cognitive psychology can be used to identify dimensions on which certain groups are hypothesized to differ in ability. Although examples in the literature are scarce, research is underway that uses cognitive theory to predict group differences on specific item types. For example, Gallagher et al. (2000) used high school and college students to investigate gender differences in advanced mathematical problem solving on Scholastic Aptitude Test–Mathematics and Graduate Record Examination–Quantitative, respectively, by dividing math problems into two main types (conventional and unconventional), based on the cognitive processes associated with answering the items correctly. Using "think aloud" problem solving with a group of high-ability high school students, Gallagher et al. found that males tended to be more flexible in their use of problem-solving strategies, whereas females tended to employ conventional problem-solving algorithms to solve advanced mathematical problems.

Finally, unintended dimensions can be identified using statistical dimensionality assessment tools. Using a mixed exploratory and confirmatory approach, Douglas et al. (1996) combined hierarchical agglomerative cluster analysis (HCA) and DIMTEST, a nonparametric statistical dimensionality test based on Stout's (1987) concept of essential unidimensionality, to identify suspect item bundles. First, an exploratory dimensionality analysis was performed so that DIF hypotheses could be developed and then tested with a cross-validation sample. Results revealed a six-item bundle measuring an additional dimension interpreted as "knowledge of some important documents in early American history" (Douglas et al., 1996, p. 477). When tested for DIF, the bundle was found to favor females.

It should be noted that all the methods described above include assumptions and limitations that make them more or less desirable in certain contexts. For example, cognitive classifications listed in a table of test specifications are developed by item writers who try to anticipate steps in the cognitive process that examinees typically follow in arriving at correct answers. Item writers, however, are often content experts, such as teachers and curriculum specialists, who are not usually trained to identify these mental processes. Furthermore, cognitive skill categories are often based on the taxonomy of educational objectives developed by Bloom, Englehart, Furst, Hill, and Krathwohl (1956), which has been shown to be inadequate for classifying or predicting students' cognitive processes on tests of math achievement (Gierl, 1997). As a result, care must be taken in deciding whether to rely on a single organizing principle or whether the research goal is best met by using a combination of strategies.

Regardless of the type of organizing principle used, creating bundles is only a first step in the two-stage Roussos-Stout DIF analysis paradigm (Roussos & Stout, 1996a). The first stage can be described as a substantive analysis of the dimensional structure of a test. It is substantive to the extent that the dimensions are actually interpretable and can be identified as target or nuisance. If nuisance dimensions exist, they are bundled together to represent dimensionality-based DIF hypotheses that, in the second stage of DIF analyses, are tested for statistical significance.

Simultaneous Item Bias Test (SIBTEST)

SIBTEST was developed by Shealy and Stout (1993a, 1993b) as an outgrowth of their MMD. SIBTEST is an IRT-based method that models the relationship between item performance and the latent trait(s) measured by a test. This method can test for significant DIF amplification that occurs when a group of DIF items act together to produce DBF. In the MMD, the latent (e.g., ability) space measures target (θ) and nuisance (η) traits. The SIBTEST method uses a parameter estimate ($\hat{\beta}_{\mathrm{UNI}}$) to indicate the magnitude of DIF in an item. For large samples, $\hat{\beta}_{\mathrm{UNI}}$ has a normal distribution, with a mean of 0 and a standard deviation of 1, under the null hypothesis of no DIF. The statistical hypothesis tested by SIBTEST is

$$H_0 : \beta_{\mathrm{UNI}} = 0 \text{ vs. } H_1 : \beta_{\mathrm{UNI}} \neq 0. \tag{1}$$

Here, β_{UNI} is defined as follows:

$$\beta_{\mathrm{UNI}} = \int [P(\theta, R) - P(\theta, F)] f_F(\theta)\, d\theta \tag{2}$$

where $P(\theta, \mathrm{R})$ and $P(\theta, \mathrm{F})$ are the probabilities of a correct response (conditional on θ) for examinees in the reference and focal groups, respectively. The expression $f_F(\theta)$ is the density function for θ in the focal group. β_{UNI} is integrated over θ and yields a weighted expected score difference between reference and focal group examinees of equal ability on a specific item or bundle. A statistically significant positive value of $\hat{\beta}_{\text{UNI}}$ represents DIF against the focal group, and a statistically significant negative value indicates DIF in favor of the focal group.

SIBTEST requires that test items be divided into a "studied" subtest of items believed to exhibit DIF and a matching or "valid" subtest of items believed to be DIF free. That is, the studied subtest contains the items or bundle believed to measure the target and nuisance dimensions, whereas the matching subtest contains the items believed to measure only the target dimension. Examinee performance on items is compared by placing reference and focal group members into subgroups at each score level on the matching subtest.

An unbiased estimate of $\beta_{\text{UNI}}(\widehat{\beta}_{\text{UNI}})$ is obtained using the weighted mean difference between the reference and focal groups on the studied item/bundle across the K matched ability subgroups, or

$$\hat{\beta}_{\text{UNI}} = \sum_{k=0}^{K} p_k \left(\overline{Y^*_{Rk}} - \overline{Y^*_{Fk}} \right), \tag{3}$$

where the proportion of focal group examinees in subgroup k is represented by p_k, and $\overline{Y^*_{Rk}} - \overline{Y^*_{Fk}}$ is the difference in the adjusted means on the studied subtest item or bundle for examinees in the reference and focal groups, respectively, in each subgroup k. Shealy and Stout (1993a) added a regression correction to adjust the means on the studied subtest item or bundle to account for any differences in the target ability distributions of the reference and focal groups.

DIF/DBF Research

The first study involving DBF was conducted by Douglas et al. (1996) who demonstrated the use of two methods for selecting item bundles suspected of exhibiting gender DIF amplification. Method 1 used a panel of judges to identify bundles of items that appeared to measure abilities in addition to the target ability (i.e., logical reasoning) using data from a standardized administration of the logical reasoning subtest of the Law School Admission Test. Using this method, the panel was able to form eight suspect item bundles, thus eight DIF hypotheses, to submit for statistical analysis. Method 2 was a mixed exploratory-confirmatory approach to identifying suspect bundles. This portion of the study involved a statistical IRT dimensionality analysis followed by the use of expert opinion to develop DIF hypotheses based on the number and type of dimensions that were identified statistically.

In both methods, SIBTEST was used to analyze the bundles for differential bundle/test functioning. For the example used with Method 1, although only four of the eight DIF analyses were statistically significant, Douglas et al. (1996) found that for seven of the eight bundles, the direction of DIF hypothesized by the panel of judges agreed with statistical results obtained with SIBTEST. Method 2 was illustrated using a 36-item National Assessment of Educational Progress (NAEP) history examination. The statistical dimensionality analysis augmented by expert opinion yielded a statistically significant six-item bundle, with three of the items not having statistically significant DIF at the 0.05 level when a standard one-at-a-time DIF analysis was conducted. That an unintended dimension could be identified as the source of DBF on the NAEP history exam illustrates the usefulness of these methods in identifying the causes of DIF/DBF in general.

Research has also demonstrated the use of DBF analysis for enhancing the interpretability of DIF results and how using different "organizing principles" for bundling items can improve the substantive review of flagged DIF items (see Gierl et al., 2001; Gierl & Khaliq, 2001). Although the literature on DBF in applied contexts appears to be growing, to date, there has been only one simulation study that has examined the performance of SIBTEST in detecting DBF. A recent simulation study by Russell (2005) assessed the Type I error rate and empirical power of SIBTEST for dichotomously scored items.

Russell (2005) used multidimensionality to produce differential item/bundle functioning in simulated item responses and used SIBTEST for DBF detection under a variety of conditions using an alpha of .01 and 50 replications. The variables in the Type I error portion of the study were dimensionality (items in the studied bundle were generated to be either unidimensional or multidimensional), test length (10 items and 20 items), sample size ($N = 500$ and $N = 1{,}000$ in both the reference and the focal group), and target ability differences against the focal group, also known as impact (standardized differences of 0.0, 0.5, and 1.0). In conditions with no target ability differences, the average Type I error rate was 0.015 for conditions with the unidimensional bundle and 0.03 for conditions with the multidimensional bundle. In conditions with target ability differences, however, the Type I error rates greatly increased. The average Type I error rate in conditions with target ability differences with a unidimensional bundle was 0.523 and 0.238 with a multidimensional bundle. Russell concluded that when subgroups differ greatly on the target ability, SIBTEST is more likely to identify the impact mistakenly as DIF/DBF.

The portion of Russell's (2005) study on statistical power included 24 unique combinations of total sample size (1,000 and 2,000), test length (10 items vs. 20 items), target ability differences or impact (0.0, 0.5, and 1.0), and nuisance ability differences (0.5 and 1.0). SIBTEST demonstrated an overall empirical power of 0.725, with the proportion of true positives in a single condition ranging from 0.22 to 1.00. Greater power to detect DBF was observed for longer tests and for larger sample sizes. For conditions involving target ability differences, SIBTEST performed best when impact was minimal. Finally, nuisance ability differences exhibited a large influence on power for both techniques. DBF power for SIBTEST reached as high as 1.00 for various conditions involving large nuisance ability differences.

The literature also includes studies that specifically examine the performance of SIBTEST for detection of DIF at the item and test level rather than the bundle level. One of the earliest such investigations was conducted by Nandakumar (1993) who used SIBTEST to study simultaneous DIF amplification and cancellation. By varying sample size, test length, percentage of DIF items, and direction of DIF with simulated data, Nandakumar's (1993) study helped to establish SIBTEST as an effective procedure for studying DIF at the item and test level. In the study, SIBTEST was compared with the Mantel-Haenszel technique and found to perform similarly across conditions in the assessment of DIF at the item level. At the test level, SIBTEST successfully estimated the cumulative effect of DIF. That is, whether DIF was amplified to produce differential test functioning or cancelled out because of bidirectional DIF in individual items, SIBTEST was capable of assessing both.

Primacy of the Target Dimension and DIF Detection

One variable previously shown to affect DIF detection is related to item discrimination and whether a multidimensional item has a higher discrimination for the target dimension or for the nuisance dimension (which can be quantified as angular item direction, Reckase & McKinley, 1991). Because a test that unintentionally includes multidimensional items is certainly not designed to measure a nuisance dimension, it is reasonable not to expect those items to be a better measure of the nuisance dimension than the target dimension. In practice, however, this may be more commonplace than previously thought. In mathematics, story or word problems are item types that typically exhibit this behavior. For example, consider the following set of multiple choice items (Reckase, 1985, p. 411) designed to measure mathematical ability:

9. $|-5| + |6| + (-5) + 6 = ?$
   A. −22
   B. −10
   C. 2
   D. 10
   E. 12

20. A serving of a certain cereal, with milk, provides 35% of the potassium required daily by the average adult. If a serving of this cereal with milk contains 112 milligrams of potassium, how many milligrams of potassium does the average adult require each day?
    A. 35
    B. 39
    C. 147
    D. 320
    E. 392

Although Item 9 seems to measure only math ability, Item 20 appears to be measuring both math and reading ability. It is possible that Item 20 has a higher-discrimination value for the reading dimension than for the math dimension.

Few studies have examined the impact of the primacy of the target dimension (e.g., whether the target dimension has a higher discrimination than that of the nuisance dimension) on the detection of DIF, and none have investigated its effect on DBF analysis. One study by Oshima and Miller (1992) observed increases in power (with four DIF detection methods) when multidimensional items embedded with DIF measured the nuisance dimension more than the target dimension (i.e., had a high angular item direction). The results of that study highlight the usefulness of item dimensionality assessment for enhancing DIF detection.

In summary, DBF seems very promising for understanding why items are biased. Although a number of studies have demonstrated methods for creating bundles and useful applications of DBF analysis, there has been little simulation work examining how SIBTEST performs in identifying DBF. The purpose of this study was to assess the performance of SIBTEST for DBF analysis under various conditions with simulated data. Therefore, this study examined how Type I error and power rates of SIBTEST are affected by multiple levels of various manipulated conditions that applied researchers often encounter.

Method

A Monte Carlo simulation study was conducted to assess the performance of the SIBTEST procedure for detecting DBF under various conditions often encountered by applied researchers. The independent variables under investigation were test length, total sample size, the sample size ratio of reference to focal group members, the correlation between the target and nuisance dimensions, the magnitude of DIF/DBF, the percent of items in the test that were part of the bundle, and the primacy of the target dimension. The impact of these variables on power and Type I error performance was the primary focus of the investigation.

The item parameters used to generate unidimensional data across all of the conditions (see the appendix) were selected from Raju, van der Linden, and Fleer's (1995) study of DIF. These parameters were reported to have item characteristics typically found in real testing situations. The same item parameters were used for both the focal and reference group. Item parameters within the multidimensional bundle were held constant for all items with a difficulty parameter value of 0 and a guessing parameter value of 0.2. The discrimination values used are described below. For conditions with no DIF/DBF and no impact, both the focal group and the reference group had a mean ability of 0 on the target ability dimension, with a standard deviation of 1. In conditions with DIF/DBF present, simulees in the reference group had a mean ability of 0 on the nuisance dimension, with a standard deviation of 1; mean differences were applied to the focal group so that they would have a lower ability on this nuisance dimension. This process will be elaborated on below.

Independent Variables

Two levels of test length were used to represent a short test (20 items) and a test of moderate length (40 items). To examine the effect on DBF detection of the percentage of items that contain DIF, two levels of percentage of items containing DIF were used, 10% and 25%; all items that had DIF were included in the bundle, and no items outside of the bundle contained DIF. This resulted in bundles of either 2 or 5 items from the 20-item test and bundles of either 4 or 10 items from the 40-item test; an additional value for number of DIF items in the bundle (5 items) was used for a 40-item test in a separate portion of the design (which will be elaborated on below).

Three levels of overall sample size (1,000, 2,000, and 5,000) as well as two different sample size ratios of reference group to focal group members (50/50 and 90/10) were used in this study. Although sample size is one of the most commonly studied conditions simulated in DIF research, it is an important factor to include here as a control for absolute sample size when sample size ratios are also being manipulated. Instances in which the ratio of reference to focal group members in the population might be expected to be approximately equal would include studies of gender DIF in which there are relatively equal numbers of boys and girls (who constitute the reference and focal groups, respectively). An example of the second instance (unequal proportions) might be when a large number of White students take an exam compared with a smaller number of Black or Hispanic students.

Another condition of interest in this study was the impact of different correlations between the target and nuisance dimensions. Because prior research (e.g., Kirisci, Hsu, & Yu, 2001) has suggested the use of unidimensional IRT methods when the correlation between dimensions is moderate-to-large (e.g., $r > .40$), this study examined the effect of correlations between dimensions on Type I error and power rates of SIBTEST. Three levels of correlation between the target and nuisance dimensions were used in this study: .316, .632, and .837, which correspond to squared correlations of .1, .4, and .7, respectively. This range of values represents interdimension correlations that are considered to be low, medium, and high, respectively.

This study also examined the effect of varying the primacy of the target and nuisance dimensions. For one level of this variable, items were simulated to have a higher discrimination on the target trait than on the nuisance trait ($a_1 = 1.00$ and $a_2 = 0.49$). In the other level of this variable, items had higher discrimination on the nuisance dimension than on the target dimension ($a_1 = 0.30$ and $a_2 = 0.49$), creating a situation in which the items are more effective at discriminating on the nuisance dimension than on the target ability dimension. Beyond switching the target dimension from primary to secondary, the choice of these specific pairs of discrimination values was arbitrary, but they incidentally create an angular item direction of 26.1° and 58.5°, respectively. All items in a bundle had the same pair of discrimination values, which meant that for a given cell, either all of the items with DIF were better at discriminating on the target dimension or all of the items with DIF were better at discriminating on the nuisance dimension.

Another variable manipulated in this study was DIF magnitude, which was expressed as a mean difference in the nuisance ability dimension between the focal group and the reference group. Shealy and Stout's (1993a) model for DIF expresses the magnitude of DIF as the difference between the conditional expectation of η, given θ, for the reference group and focal group,

$$E[\eta_R|\theta] - E[\eta_F|\theta] = d_\eta - \rho d_\theta, \tag{4}$$

where ρ is the correlation between the two dimensions (for the simple case in which the correlation is the same for both the reference and focal groups), d_θ is the difference between the means of the focal and reference group on θ (impact), and d_η is the mean of the focal group on η subtracted from the mean of the reference group on η. For the current study, the mean on the target ability was the same for both groups (indicating no impact), resulting in a d_θ value of 0. Positive values of d_η indicate DIF against the focal group, and negative values indicate DIF against the reference group. In Nandakumar's (1993) simulation study, two levels of d_η were chosen, 0.5 and 1.0, to represent moderate-to-large degrees of DIF. Four levels of DIF against the focal group ($d_\eta = \{0.25, 0.5, 0.75, 1.0\}$) were included in this study to reflect small to large amounts of DIF. The process of incorporating DIF will be elaborated on below.

The procedure contains three studies in one design. The main portion of the design was a power study, conducted to investigate the impact of several manipulated variables on the ability of SIBTEST to detect DBF. This power study included 576 conditions, fully crossing each of the independent variables (2 Test Lengths × 2 Percentages of Items With DIF × 3 Total Sample Sizes × 2 Sample Size Ratios × 3 Correlations Between the Target and Nuisance Dimensions × 4 Magnitudes of DIF × 2 Dimension Primacies = 576).

In an effort to control for the number of items with DIF, an additional 144 cells were incorporated in the design. The simulated test length was 40 items, 5 of those items were simulated to have DIF, and the remaining conditions from the first part of the design were crossed (3 Total Sample Sizes × 2 Sample Size Ratios × 3 Correlations Between the Target and Nuisance Dimensions × 4 Magnitudes of DIF × 2 Dimension Primacies = 144). In this way, the two levels of test length could be compared through a common value for the *number* of items with DIF (5 items with DIF also occurs when 25% of 20 items have DIF), instead of just having common values for the percentage of items with DIF.

A Type I error study was also conducted to examine false detection of DBF with SIBTEST for multidimensional items. For the Type I error study, there were a total of 144 conditions (2 Test Lengths × 2 Percentages of Items With DIF × 3 Total Sample Sizes × 2 Sample Size Ratios × 3 Correlations Between Target and Nuisance Traits × 2 Dimension Primacies).

Data Generation

Unidimensional and multidimensional data for this design were generated according to the conditions described above. IRTGEN (Whitaker, Fitzpatrick, Williams, & Dodd, 2003), a SAS macro program, was used to simulate unidimensional items. Multidimensional items for the reference and focal groups were simulated using the SAS/IML program GENMIRT (Kromrey, Parshall, Chason, & Yi, 1999).

For items measuring only the target ability (θ), dichotomous data were generated using the three-parameter logistic model (3PL),

$$P(u_i = 1|\theta) = c_i + (1 - c_i)\frac{e^{a_i(\theta - b_i)}}{1 + e^{a_i(\theta - b_i)}}, \tag{5}$$

where $P(u_i = 1|\theta)$ is the probability that an examinee correctly answers item i, c_i is the lower asymptote, b_i is the item difficulty, a_i is the item discrimination, and e represents the base of the natural logarithm. A two-dimensional 3PL model was used to simulate items measuring the target and nuisance dimensions,

$$P(u_i = 1|\theta) = c_i + (1 - c_i)\frac{e^{(a_i'\theta + d_i)}}{1 + e^{(a_i'\theta + d_i)}}, \qquad (6)$$

where $P(u_i = 1|\theta)$ is the probability of a correct response for item i given the $\boldsymbol{\theta}$ vector of abilities (for the two-dimensional case, θ and η), c_i is the scalar lower asymptote for the item, a_i is the vector of item discriminations, and d_i is the scalar item difficulty. For conditions involving no DIF, reference and focal group abilities on the target and nuisance traits were generated according to a multivariate normal distribution, with a mean of 0 and a standard deviation of 1. In the power studies, DIF was added to the multidimensional items by increasing the mean ability of the reference group above that of the focal group by 0.25, 0.50, 0.75, or 1.00 standard deviations on the nuisance dimension, depending on the condition being tested.

After generating item response data, SAS was used to call the DOS-based version of SIBTEST and to produce output files for the DBF analyses for each of 1,000 replications for each cell. Power and Type I error rates were calculated for each cell by tallying the number of times that SIBTEST detected statistically significant DBF in the bundle and dividing by the number of replications. The alpha level was .05 for all conditions.

Results

Tables 1 through 4 display the results from the power study. Table 5 displays the mean results for each level of each studied factor. Table 6 displays the results from the conditions that had 40 items total, 5 of which had DIF. Table 7 displays the results from the Type I error study. The results begin with a summary of the overall effect of each manipulated variable on power and Type I error and then move on to a description of cells in which power was at least 0.80.

Power Main Effects

Sample size and sample size ratio. Total sample size influenced the DBF detection power of SIBTEST. As might be expected, higher power rates were associated with larger sample sizes. For each total sample size, power was substantially higher for cells in which the reference and focal group sizes were equal than for cells that had sharply unequal group sizes (i.e., 90:10). With the 50/50 sample size ratio, mean power was 0.798 ($SD = 0.258$) across cells with $N = 1{,}000$, 0.893 ($SD = 0.188$) across cells in which $N = 2{,}000$, and 0.969 ($SD = 0.078$) across cells in which $N = 5{,}000$. With the 90/10 sample size ratio, mean power was 0.566 ($SD = 0.290$), 0.727 ($SD = 0.286$), and 0.876 ($SD = 0.205$), respectively, across cells with $N = 1{,}000$, $N = 2{,}000$, and $N = 5{,}000$. For

Table 1
Empirical Power for $a_1 = 1.00$ and 10% of Items Having Differential Item Functioning

				Test Length							
				20 Items				40 Items			
				Nuisance Ability Difference, d_η							
	N_{tot}	N_{ref}	N_{foc}	0.25	0.5	0.75	1.00	0.25	0.5	0.75	1.00
$\rho_{\theta\eta} = .316$	1,000	900	100	.140	.285	.519	.676	.155	.371	.611	**.816**
		500	500	.272	.614	**.919**	**.980**	.309	.756	**.967**	**.998**
	2,000	1,800	200	.187	.483	.785	**.928**	.228	.586	**.865**	**.972**
		1,000	1,000	.362	**.884**	**.994**	**1.000**	.496	**.951**	**1.000**	**1.000**
	5,000	4,500	500	.374	**.812**	**.987**	**1.000**	.480	**.928**	**.998**	**1.000**
		2,500	2,500	.704	**.995**	**1.000**	**1.000**	**.834**	**1.000**	**1.000**	**1.000**
$\rho_{\theta\eta} = .632$	1,000	900	100	.116	.250	.447	.620	.170	.322	.556	.750
		500	500	.233	.564	**.852**	**.975**	.298	.678	**.941**	**.997**
	2,000	1,800	200	.178	.447	.712	**.910**	.224	.538	**.818**	**.964**
		1,000	1,000	.372	**.800**	**.994**	**1.000**	.437	**.911**	**1.000**	**1.000**
	5,000	4,500	500	.328	.774	**.972**	**1.000**	.399	**.884**	**.995**	**.999**
		2,500	2,500	.665	**.995**	**1.000**	**1.000**	.774	**.999**	**1.000**	**1.000**
$\rho_{\theta\eta} = .837$	1,000	900	100	.117	.255	.418	.623	.146	.288	.521	.740
		500	500	.218	.539	**.833**	**.966**	.271	.645	**.930**	**.994**
	2,000	1,800	200	.176	.376	.710	**.894**	.217	.495	**.811**	**.956**
		1,000	1,000	.347	**.801**	**.979**	**1.000**	.427	**.890**	**.997**	**1.000**
	5,000	4,500	500	.301	.739	**.971**	**1.000**	.365	**.848**	**.992**	**.999**
		2,500	2,500	.631	**.989**	**1.000**	**1.000**	.744	**.998**	**1.000**	**1.000**

purposes of further discussion, sample size and sample size ratio will be discussed using the harmonic mean of the groups' sample sizes ($\tilde{n}$).

Magnitude of DIF/DBF. DIF bundles were correctly identified by SIBTEST more often for higher magnitudes of DIF/DBF. The average power for detecting a DIF/DBF magnitude of 0.25 was 0.499 ($SD = 0.262$). The average power was 0.815 ($SD = 0.219$), 0.932 ($SD = 0.130$), and 0.975 ($SD = 0.069$), respectively, for detecting DIF/DBF magnitudes of 0.50, 0.75, and 1.00.

Percentage of items with DIF. With 25% of the items containing DIF, power was appreciably higher than when 10% of the items contained DIF. Mean power across the conditions with 25% of the items containing DIF was 0.871 ($SD = 0.218$); mean power in the cells in which 10% of the items contained DIF was 0.775 ($SD = 0.288$). The effect was more pronounced in the cells in which the DIF items had a higher discrimination on the nuisance dimension than on the target dimension.

Dimension primacy. Overall, primacy of the target dimension was influential in DBF detection. Comparing the values in Tables 1 and 3 with those in Tables 2 and 4, it is

Table 2
Empirical Power for $a_1 = 0.30$ and 10% of Items Having Differential Item Functioning

				Test Length							
				20 Items				40 Items			
				Nuisance Ability Difference, d_η							
	N_{tot}	N_{ref}	N_{foc}	0.25	0.5	0.75	1.00	0.25	0.5	0.75	1.00
$\rho_{\theta\eta} = .316$	1,000	900	100	.189	.458	.740	**.941**	.265	.616	**.896**	**.989**
		500	500	.393	**.878**	**.996**	**1.000**	.541	**.955**	**1.000**	**1.000**
	2,000	1,800	200	.290	.724	**.951**	**.995**	.424	**.897**	**.998**	**1.000**
		1,000	1,000	.637	**.995**	**1.000**	**1.000**	**.836**	**.999**	**1.000**	**1.000**
	5,000	4,500	500	.587	**.981**	**1.000**	**1.000**	.743	**.998**	**1.000**	**1.000**
		2,500	2,500	**.930**	**1.000**	**1.000**	**1.000**	**.992**	**1.000**	**1.000**	**1.000**
$\rho_{\theta\eta} = .632$	1,000	900	100	.193	.453	.730	**.910**	.235	.562	**.873**	**.977**
		500	500	.356	**.841**	**.994**	**.999**	.499	**.966**	**1.000**	**1.000**
	2,000	1,800	200	.296	.718	**.954**	**.996**	.389	**.852**	**.993**	**1.000**
		1,000	1,000	.581	**.981**	**1.000**	**1.000**	.783	**1.000**	**1.000**	**1.000**
	5,000	4,500	500	.514	**.981**	**1.000**	**1.000**	.730	**.998**	**1.000**	**1.000**
		2,500	2,500	**.909**	**1.000**	**1.000**	**1.000**	**.983**	**1.000**	**1.000**	**1.000**
$\rho_{\theta\eta} = .837$	1,000	900	100	.170	.428	.705	**.894**	.221	.557	**.875**	**.971**
		500	500	.366	**.831**	**.988**	**1.000**	.468	**.947**	**.998**	**1.000**
	2,000	1,800	200	.262	.678	**.936**	**.993**	.364	**.857**	**.988**	**1.000**
		1,000	1,000	.586	**.988**	**1.000**	**1.000**	.747	**.998**	**1.000**	**1.000**
	5,000	4,500	500	.539	**.964**	**1.000**	**1.000**	.701	**.996**	**1.000**	**1.000**
		2,500	2,500	**.886**	**1.000**	**1.000**	**1.000**	**.981**	**1.000**	**1.000**	**1.000**

evident that power to detect DBF across all conditions with SIBTEST was greater when the nuisance dimension had a higher discrimination than the target dimension. Observed power when $a_1 = 1.00$ averaged 0.739 ($SD = 0.288$), whereas the cells with $a_1 = 0.30$ had an average power of 0.871 ($SD = 0.218$).

Correlation between target and nuisance dimensions. Power increased marginally as dimensions became less correlated, with the average power in cells with $\rho_{\theta\eta} = .837$ being 0.791 ($SD = 0.272$), which increased only to 0.822 ($SD = 0.253$) when $\rho_{\theta\eta} = .316$.

Test length . Power was generally larger with a test length of 40 items than with a test length of 20 items. Across the conditions with 40 total items, mean power was 0.827 ($SD = 0.251$), and across the conditions with 20 items, the mean power was 0.783 ($SD = 0.273$). There was an apparent interaction with DIF magnitude in that the power increases seen from 20 items to 40 items were larger when DIF magnitudes were smaller because perhaps of the fact that at and above $d_\eta = 0.75$, most of the power values were quite high, regardless of simulated test length.

Table 3
Empirical Power for $a_1 = 1.00$ and 25% of Items Having Differential Item Functioning

				Test Length							
				20 Items				40 Items			
				Nuisance Ability Difference, d_η							
	N_{tot}	N_{ref}	N_{foc}	0.25	0.5	0.75	1.00	0.25	0.5	0.75	1.00
$\rho_{\theta\eta} = .316$	1,000	900	100	.156	.416	.661	**.849**	.194	.423	.728	**.884**
		500	500	.335	.771	**.970**	**.998**	.381	**.851**	**.987**	**1.000**
	2,000	1,800	200	.251	.607	**.907**	**.991**	.284	.700	**.942**	**.996**
		1,000	1,000	.509	**.966**	**1.000**	**1.000**	.584	**.977**	**1.000**	**1.000**
	5,000	4,500	500	.502	**.927**	**.997**	**1.000**	.524	**.966**	**1.000**	**1.000**
		2,500	2,500	**.842**	**1.000**	**1.000**	**1.000**	**.895**	**1.000**	**1.000**	**1.000**
$\rho_{\theta\eta} = .632$	1,000	900	100	.171	.367	.587	**.806**	.168	.380	.590	**.834**
		500	500	.314	.723	**.957**	**.998**	.297	.776	**.981**	**.999**
	2,000	1,800	200	.226	.571	**.860**	**.976**	.235	.646	**.910**	**.992**
		1,000	1,000	.502	**.934**	**1.000**	**1.000**	.494	**.975**	**1.000**	**1.000**
	5,000	4,500	500	.412	**.893**	**1.000**	**1.000**	.480	**.941**	**.997**	**1.000**
		2,500	2,500	.796	**1.000**	**1.000**	**1.000**	**.867**	**1.000**	**1.000**	**1.000**
$\rho_{\theta\eta} = .837$	1,000	900	100	.157	.336	.573	.782	.138	.373	.585	**.840**
		500	500	.273	.683	**.951**	**.998**	.314	.744	**.966**	**.998**
	2,000	1,800	200	.223	.549	**.823**	**.974**	.217	.596	**.864**	**.980**
		1,000	1,000	.473	**.937**	**.999**	**1.000**	.487	**.962**	**1.000**	**1.000**
	5,000	4,500	500	.380	**.880**	**.992**	**1.000**	.427	**.927**	**.997**	**1.000**
		2,500	2,500	**.776**	**.999**	**1.000**	**1.000**	**.840**	**1.000**	**1.000**	**1.000**

Attaining Sufficient Power

The results presented so far describe the effects of the variables investigated in the study. Now, we turn to presenting the results from an applied angle by explaining the conditions in which power for detecting DIF/DBF was at least 0.80. These results can also be found in Tables 1 through 4 in which the values that are at least 0.80 are listed in boldface.

Among the 144 cells with $d_\eta = 0.25$, 25 cells had power of at least 0.80. The bulk of these (20/25) occurred when the nuisance dimension was the primary dimension measured by the items, with 14 of those 20 having power >0.90. In only two cells ($\rho = .316$ and $\rho = .837$, with 40 total items, 25% of the items having DIF, and both ns = 2,500), was the power 1.00, when $d_\eta = 0.25$.

Nearly two thirds (94/144) of the cells with $d_\eta = 0.50$ had power of at least 0.80, again with the majority of those cells (59/94) having the nuisance dimension as the primary dimension. Seventy-four cells had an average power greater than 0.90, and 30 cells averaged a power of 1.00, only 6 of which occurred with $a_1 = 1.00$. Among the cells with $a_1 = 0.30$ and 25% of the items containing DIF, only 4 cells did not have an average

Table 4
Empirical Power for $a_1 = 0.30$ and 25% of Items Having Differential Item Functioning

				Test Length							
				20 Items				40 Items			
				Nuisance Ability Difference, d_η							
	N_{tot}	N_{ref}	N_{foc}	0.25	0.5	0.75	1.00	0.25	0.5	0.75	1.00
$\rho_{\theta\eta} = .316$	1,000	900	100	.278	.695	**.943**	**.998**	.391	**.815**	**.981**	**1.000**
		500	500	.575	**.979**	**1.000**	**1.000**	.734	**.997**	**1.000**	**1.000**
	2,000	1,800	200	.468	**.928**	**.999**	**1.000**	.555	**.986**	**1.000**	**1.000**
		1,000	1,000	**.849**	**1.000**	**1.000**	**1.000**	**.941**	**1.000**	**1.000**	**1.000**
	5,000	4,500	500	**.800**	**1.000**	**1.000**	**1.000**	**.913**	**1.000**	**1.000**	**1.000**
		2,500	2,500	**.998**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
$\rho_{\theta\eta} = .632$	1,000	900	100	.265	.666	**.939**	**.991**	.351	**.801**	**.958**	**1.000**
		500	500	.520	**.977**	**.999**	**1.000**	.674	**.992**	**1.000**	**1.000**
	2,000	1,800	200	.418	**.893**	**.999**	**1.000**	.518	**.968**	**1.000**	**1.000**
		1,000	1,000	.791	**1.000**	**1.000**	**1.000**	**.900**	**1.000**	**1.000**	**1.000**
	5,000	4,500	500	.762	**1.000**	**1.000**	**1.000**	**.887**	**1.000**	**1.000**	**1.000**
		2,500	2,500	**.990**	**1.000**	**1.000**	**1.000**	**.998**	**1.000**	**1.000**	**1.000**
$\rho_{\theta\eta} = .837$	1,000	900	100	.260	.630	**.905**	**.994**	.298	.721	**.946**	**.999**
		500	500	.512	**.958**	**.998**	**1.000**	.644	**.994**	**1.000**	**1.000**
	2,000	1,800	200	.400	**.875**	**.995**	**1.000**	.489	**.949**	**.999**	**1.000**
		1,000	1,000	.773	**1.000**	**1.000**	**1.000**	**.876**	**1.000**	**1.000**	**1.000**
	5,000	4,500	500	.740	**1.000**	**1.000**	**1.000**	**.848**	**1.000**	**1.000**	**1.000**
		2,500	2,500	**.991**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**

power of at least 0.80, all of which had $\tilde{n} = 180$ (which was the lowest harmonic mean in this study).

Across the conditions with $d_\eta = 0.75$, all cells with $\tilde{n} \geq 500$ had average power values greater than 0.80, and only 3 cells with $\tilde{n} = 360$ had values less than 0.80 (all 3 had $a_1 = 1.00$ and 20 items total, with 10% of the items containing DIF). The remaining 12 cells in which power $<.80$ all had $\tilde{n} = 180$ and $a_1 = 1.00$. All the cells that had $a_1 = 0.30$ and 25% of the items with DIF had an average power more than 0.80; in fact, the average power for each of those 36 cells exceeded 0.90, with two thirds of them reaching 1.00. With $a_1 = 0.30$ and 10% of the items containing DIF, power exceeded 0.80 in all but 3 cells (minimum power $= 0.701$), and the power was 1.00 in 20 cells.

Of the 144 cells in which $d_\eta = 1.00$, only 6 had an average power less than 0.80, with all 6 having $\tilde{n} = 180$ and $a_1 = 1.00$; five of the six cells that did not reach 0.80 had 10% of the items containing DIF, and the sixth cell had 25% of 20 items containing DIF, with $\rho = .837$ (power $= 0.782$). The lowest average power value with $a_1 = 0.30$ and 25% of the items containing DIF was 0.991, with only 4 of the cells averaging less than 1.00. With $a_1 = 0.30$ and 10% of the items containing DIF, only 10 cells had power less than 1.00, and only 1 had power less than 0.90.

Table 5
Mean and Standard Deviation of Power for Each Level of Each Studied Factor

	Power	
	M	*SD*
Test length		
20 items	0.783	0.273
40 items	0.827	0.251
Sample size with a 90/10 ratio		
1000	0.565	0.290
2000	0.727	0.286
5000	0.876	0.205
Sample size with a 50/50 ratio		
1000	0.798	0.258
2000	0.893	0.188
5000	0.969	0.078
Percentage of items in bundle		
10	0.775	0.288
25	0.871	0.218
Dimensions' correlation		
.316	0.822	0.253
.632	0.802	0.265
.837	0.791	0.272
Differential item functioning magnitude		
0.25	0.499	0.262
0.5	0.815	0.219
0.75	0.932	0.130
1	0.975	0.069
Target dimension discrimination		
0.3	0.739	0.288
1	0.871	0.218

Controlling for the Number of Items With DIF

Table 6 contains the power results for the portion of the design in which 144 cells were run with 40 total items, 5 of which had DIF. These cells were included in the design as a way to hold constant the number of items with DIF while varying test length. There are two general comparisons involving these data that are relevant: A comparison of the results in the $a_1 = 1.00$ column of Table 6 with the 20 items column of Table 3 and a comparison of the results in the $a_1 = 0.3$ column of Table 6 with the results in the 20 items column of Table 4. Although there is slight variation between most of the corresponding cells, only 17 of the 144 cells had an absolute difference greater than 0.02. The comparisons clearly illustrate that increasing the total number of items, while holding constant the number of items with DIF, has essentially no effect on the power to detect DIF/DBF when 5 items contain DIF. The apparent main effect of test length found in the first part of the design seems therefore to be the result of the increase in the number of items with DIF that occurred as happenstance when the total number of items was changed while holding constant

Table 6
Empirical Power for 40 Items With 5 Items Having Differential Item Functioning

				a_1							
				0.3				1.0			
				Nuisance Ability Difference, d_η							
	N_{tot}	N_{ref}	N_{foc}	0.25	0.50	0.75	1.00	0.25	0.50	0.75	1.00
$\rho_{\theta\eta} = .316$	1,000	900	100	0.285	0.668	**0.938**	**0.998**	0.171	0.387	0.648	**0.844**
		500	500	0.623	**0.990**	**1.000**	**1.000**	0.322	0.764	**0.967**	**0.999**
	2,000	1,800	200	0.477	**0.923**	**0.999**	**1.000**	0.249	0.626	**0.904**	**0.988**
		1,000	1,000	**0.865**	**1.000**	**1.000**	**1.000**	0.555	**0.958**	**1.000**	**1.000**
	5,000	4,500	500	**0.817**	**1.000**	**1.000**	**1.000**	0.480	**0.928**	**1.000**	**1.000**
		2,500	2,500	**0.994**	**1.000**	**1.000**	**1.000**	**0.843**	**1.000**	**1.000**	**1.000**
$\rho_{\theta\eta} = .632$	1,000	900	100	0.257	0.647	**0.909**	**0.988**	0.141	0.355	0.604	0.795
		500	500	0.557	**0.976**	**1.000**	**1.000**	0.300	0.713	**0.953**	**0.995**
	2,000	1,800	200	0.407	**0.900**	**1.000**	**1.000**	0.214	0.563	**0.859**	**0.975**
		1,000	1,000	0.797	**0.999**	**1.000**	**1.000**	0.470	**0.935**	**0.999**	**1.000**
	5,000	4,500	500	0.760	**0.997**	**1.000**	**1.000**	0.416	**0.905**	**0.998**	**1.000**
		2,500	2,500	**0.987**	**1.000**	**1.000**	**1.000**	**0.811**	**1.000**	**1.000**	**1.000**
$\rho_{\theta\eta} = .837$	1,000	900	100	0.206	0.605	**0.895**	**0.989**	0.153	0.305	0.556	0.749
		500	500	0.560	**0.960**	**1.000**	**1.000**	0.282	0.683	**0.929**	**0.997**
	2,000	1,800	200	0.399	**0.885**	**0.995**	**1.000**	0.215	0.532	**0.846**	**0.969**
		1,000	1000	0.786	**1.000**	**1.000**	**1.000**	0.436	**0.916**	**1.000**	**1.000**
	5,000	4,500	500	0.743	**0.999**	**1.000**	**1.000**	0.395	**0.855**	**0.995**	**0.999**
		2,500	2,500	**0.989**	**1.000**	**1.000**	**1.000**	0.755	**1.000**	**1.000**	**1.000**

the percentage of items that had DIF. The actual effect seen with the increase in total number of items was the effect of increasing the number of items containing DIF.

Type I Error

As illustrated in Table 7, SIBTEST adhered closely to the nominal alpha level of .05. Across all study conditions, the average Type I error rate was 0.047. The error rates ranged from 0.04 to 0.06, with no meaningful pattern apparent for the slight differences in the Type I error rates across the cells of the study.

Discussion

As with many DIF studies involving sample size (see Narayanan & Swaminathan, 1994; Roussos & Stout, 1996b; Shealy & Stout, 1993b), the results of this study indicate an increase in the power to detect DIF/DBF as sample size increases. By varying the ratio of the sample sizes of the focal and reference groups, it was clear that the harmonic mean

Table 7
Type I Error Rates for the Multidimensional Item Bundle

				Test Length							
				20 Items				40 Items			
				a_1							
				0.3		1.0		0.3		1.0	
				Percentage of Items in Bundle							
	N_{tot}	N_{ref}	N_{foc}	10	25	10	25	10	25	10	25
$\rho_{\theta\eta} = .316$	1,000	900	100	.058	.065	.048	.055	.054	.047	.052	.052
		500	500	.052	.042	.051	.060	.050	.053	.054	.068
	2,000	1,800	200	.050	.065	.056	.046	.048	.056	.051	.048
		1,000	1,000	.054	.060	.043	.037	.056	.053	.044	.061
	5,000	4,500	500	.026	.043	.055	.044	.053	.042	.053	.045
		2,500	2,500	.041	.044	.053	.037	.043	.040	.039	.043
$\rho_{\theta\eta} = .632$	1,000	900	100	.037	.046	.052	.047	.061	.055	.040	.064
		500	500	.050	.049	.050	.047	.044	.058	.041	.055
	2,000	1,800	200	.050	.052	.033	.045	.061	.059	.058	.060
		1,000	1,000	.053	.044	.055	.051	.065	.054	.051	.048
	5,000	4,500	500	.043	.053	.051	.039	.054	.039	.055	.042
		2,500	2,500	.053	.045	.045	.057	.038	.061	.054	.051
$\rho_{\theta\eta} = .837$	1,000	900	100	.045	.063	.039	.056	.056	.051	.050	.053
		500	500	.051	.049	.044	.057	.050	.053	.057	.055
	2,000	1,800	200	.045	.057	.056	.063	.049	.047	.065	.048
		1,000	1,000	.060	.054	.044	.042	.045	.046	.051	.045
	5,000	4,500	500	.043	.051	.051	.036	.054	.060	.042	.057
		2,500	2,500	.056	.054	.049	.059	.056	.063	.047	.062

was more important than the total number of examinees, with statistical power consistently increasing as the harmonic mean increased. Applied researchers should therefore go to some effort to keep both group sizes as large as possible rather than relying on one very large group. Consider, for example, statistical power in the cells in which both groups had 500 cases. Increasing the reference group size to 4,500 generally resulted in a marked increase in power, but increasing both group sizes to 1,000 from 500 (for a total increase of only 1,000) consistently resulted in an even greater increase in power.

The impact of increasing d_η on DIF detection was not surprising, as higher amounts of DIF should be detected at greater frequency than smaller amounts. The highest power rates were observed for DIF magnitudes of 0.50 or greater, combined with high sample sizes and low correlation between the target and nuisance dimension, although $\rho_{\theta\eta}$ had a relatively weak effect on power across the range used in this study. These results are consistent with those obtained by Oshima and Miller (1992) regarding power and those of Lee (2005) regarding the modest effect of $\rho_{\theta\eta}$.

Power also increased as factors contributing to DIF amplification increased. As the number of items with DIF increased, SIBTEST had greater power to detect DIF in item bundles, even at the smaller magnitudes of DIF (it is, however, worth repeating that the power performance of SIBTEST at $d_\eta = 0.25$ was not impressive). Substantial gains in power occurred when the target dimension switched from being the primary dimension ($a_1 = 1.00$, $a_2 = 0.49$) to the secondary dimension ($a_1 = 0.30$, $a_2 = 0.49$) measured by the items containing DIF. Those who are developing tests might therefore have greater difficulty detecting DIF in items that strongly discriminate on the target dimension; this is potentially problematic because it is generally considered preferable for items to have high values of a on the target dimension. Future research should more thoroughly and deliberately manipulate angular item direction to learn more about the effect of relative differences in the a-parameters on DIF detection.

Type I error was assessed using bundles of multidimensional items containing no DIF. SIBTEST consistently adhered closely to the nominal alpha level of .05. That inflated Type I error was not observed in this study illustrates SIBTEST's ability to distinguish between benign multidimensionality and DIF. These findings support those from Oshima and Miller (1992) and those found by Russell (2005) in comparable conditions with no impact.

This study analyzed DBF created by having all items in the bundle contain DIF, with all DIF favoring the reference group, while no items outside of the bundle contained DIF. Future research should investigate conditions in which not every item in the bundle contains DIF and conditions in which the bundles do not include all the items that contain DIF. Future research should also be conducted on the power to detect DIF when one or more items containing DIF that are within a bundle favor the focal group along with items that favor the reference group, thereby investigating various degrees of DIF cancellation. Another suggestion is to examine the interaction between item difficulty and item discrimination and its effect on DBF detection with SIBTEST. This study held difficulty constant across items within a bundle. It might be the case that DIF becomes easier/harder to detect at different levels of item difficulty.

In conclusion, there has been a push toward using more theory-driven tests for the presence of DIF (Gierl et al., 2001) rather than examining statistically flagged DIF items in an effort to patch together an explanation for the DIF that happened to be large enough to be statistically significant. The use of a MMD has been supported in this study that basically states that if you can identify a potential unintended dimension being measured by an item or a set of items and allow the possibility that examinee subgroups exist that differ on that unintended trait, then a researcher can submit specific items for DIF/DBF analysis and essentially test a substantive hypothesis about what is causing DIF/DBF on a test. This study begins to address some of the methodological aspects of DBF analyses and illustrates avenues for continued research into the application of DBF analyses.

Appendix

Item Parameters Used to Generate the Unidimensional Item Response Data

			Test Length				
			20 Items		40 Items		
			Number of Items in Bundle				
a	*b*	*c*	2 (10%)	5 (25%)	4 (10%)	10 (25%)	5
0.55	0	0.2	X	X	X	X	X
0.55	0	0.2			X	X	X
0.73	−1.04	0.2	X	X	X	X	X
0.73	−1.04	0.2			X	X	X
0.73	0	0.2	X	X	X	X	X
0.73	0	0.2	X	X	X	X	X
0.73	0	0.2			X	X	X
0.73	0	0.2			X	X	X
0.73	1.04	0.2	X	X	X	X	X
0.73	1.04	0.2			X	X	X
1	−1.96	0.2	X	X	X	X	X
1	−1.96	0.2			X	X	X
1	−1.04	0.2	X	X	X	X	X
1	−1.04	0.2	X	X	X	X	X
1	−1.04	0.2			X	X	X
1	−1.04	0.2			X	X	X
1	0	0.2	X	X	X	X	X
1	0	0.2	X	X	X	X	X
1	0	0.2	X	X	X	X	X
1	0	0.2	X	X	X	X	X
1	0	0.2			X	X	X
1	0	0.2			X	X	X
1	0	0.2			X	X	X
1	0	0.2			X	X	X
1	1.04	0.2	X	X	X	X	X
1	1.04	0.2	X	X	X	X	X
1	1.04	0.2			X	X	X
1	1.04	0.2			X	X	X
1	1.96	0.2	X	X	X	X	X
1	1.96	0.2	X		X	X	X
1	1.96	0.2	X		X		X
1	1.96	0.2	X		X		X
1	1.96	0.2			X		X
1	1.96	0.2			X		X
1	1.96	0.2			X		X
1	1.96	0.2			X		

Note: An “X” indicates that the item parameters were included in those conditions.

References

Ackerman, T. A. (1989). Unidimensional IRT calibration of compensatory and noncompensatory multidimensional items. *Applied Psychological Measurement*, *13*, 113-127.

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.

Ansley, T. N., & Forsyth, R. A. (1985). An examination of the characteristics of unidimensional IRT parameter estimates derived from two-dimensional data. *Applied Psychological Measurement*, *9*, 37-48.

Bloom, B. S., Englehart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (1956). *A taxonomy of educational objectives*, *handbook 1: The cognitive domain*. New York: David McKay.

Boughton, K., Gierl, M. J., & Khaliq, S. N. (2000, May). *Differential bundle functioning on mathematics and science achievement tests: A small step toward understanding differential performance*. Paper presented at the annual meeting of the Canadian Society for Studies in Education. Edmonton, Alberta, Canada.

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

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

Drasgow, F., & Parsons, C. K. (1983). Application of unidimensional item response theory models to multidimensional data. *Applied Psychological Measurement*, *7*, 189-199.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum.

Engelhard, G., Hansche, L., & Rutledge, K. E. (1990). Accuracy of bias review judges in identifying differential item functioning on teacher certification tests. *Applied Measurement in Education*, *3*, 347-360.

Gallagher, A. M., De Lisi, R., Holst, P. C., McGillicuddy-DeLisi, A. V., Morley, M., & Cahalan, C. (2000). Gender differences in advanced mathematical problem-solving. *Journal of Experimental Child Psychology*, *75*, 165-190.

Gierl, M. J. (1997). Comparing cognitive representations of test developers and students on a mathematical test with Bloom's taxonomy. *Journal of Educational Research*, *91*(1), 26-32.

Gierl, M. J., Bisanz, J., Bisanz, G. L., Boughton, K. A., & Khaliq, S. N. (2001). Illustrating the utility of differential bundle functioning analysis to identify and interpret group differences on achievement tests. *Educational Measurement: Issues and Practice*, *20*, 26-36.

Gierl, M. J., & Bolt, D. (2003, April). *Implications of the multidimensionality-based DIF analysis framework for selecting a matching and studied subtest*. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

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*, *38*, 164-187.

Gierl, M. J., Rogers, W. T., & Klinger, D. (1999, April). *Consistency between statistical procedures and content reviews for identifying translation DIF*. Paper presented at the annual meeting of the National Council on Measurement in Education. Montreal, Quebec, Canada.

Gierl, M. J., Tan, X., & Wang, C. (2005). *Identifying content and cognitive dimensions on the SAT* (No. 2005-11). New York: College Board.

Harrison, D. A. (1986). Robustness of IRT parameter estimation to violations of the unidimensionality assumption. *Journal of Educational Statistics*, *11*, 91-115.

Kirisci, L., Hsu, T., & Yu, L. (2001). Robustness of item parameter estimation programs to assumptions of unidimensionality and normality. *Applied Psychological Measurement*, *25*, 146-162.

Kromrey, J. D., Parshall, C. G., Chason, W. M., & Yi, Q. (1999). *Generating item responses based on multidimensional item response theory*. Retrieved July 25, 2005, from http://www2.sas.com/proceedings/sugi24/Posters/p241-24.pdf

Lee, Y. (2005). The impact of a multidimensional item on differential item functioning (DIF). *Dissertation Abstracts International*, *65* (7A), 2490. (UMI No. 3139494).

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

Nandakumar, R. (1991). Traditional versus essential unidimensionality. *Journal of Educational Measurement*, *28*, 99-117.

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

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

O'Neill, K. A., & McPeek, W. M. (1993). Item and test characteristics that are associated with differential item functioning. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 255-276). Hillsdale, NJ: Lawrence Erlbaum.

Oshima, T. C., & Miller, M. D. (1992). Multidimensionality and item bias in item response theory. *Applied Psychological Measurement*, *16*, 237-248. Oshima, T. C., Raju, N. S., Flowers, C. P., & Slinde, J. A. (1998). Differential bundle functioning using the DFIT framework: Procedures for identifying possible sources of differential functioning. *Applied Measurement in Education*, *11*, 353-369.

Raju, N. S., van der Linden, W. J., & Fleer, P. F. (1995). IRT-based internal measures of differential functioning of items and tests. *Applied Psychological Measurement*, *19*, 353-368.

Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of Educational Statistics*, *4*, 207-230.

Reckase, M. D. (1985). The difficulty of items that measure more than one ability. *Applied Psychological Measurement*, *9*, 401-412.

Reckase, M. D., Ackerman, T. A., & Carlson, J. E. (1988). Building a unidimensional test using multidimensional items. *Journal of Educational Measurement*, *215*, 193-203.

Reckase, M. D., & McKinley, R. L. (1991). The discriminating power of items that measure more than one dimension. *Applied Psychological Measurement*, *15*, 361-373.

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

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

Russell, S. S. (2005). Estimates of Type I error and power for indices of differential bundle and test functioning. *Dissertation Abstracts International*, *66* (5B), 2867. (UMI No. 3175804)

Shealy, R., & Stout, W. F. (1993a). A model-based standardization approach that separates true bias/DIF from group differences and detects test bias/DIF 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. 197-239). Hillsdale, NJ: Lawrence Erlbaum.

Standards for educational and psychological testing. (1999). Washington, DC: American Educational Research Association, American Psychological Association, National Council on Measurement in Education.

Stout, W. F. (1987). A nonparametric approach for assessing latent trait unidimensionality. *Psychometrika*, *52*, 589-617.

Stout, W. F. (1990). A new item response theory modeling approach with applications to unidimensional assessment and ability estimation. *Psychometrika*, *55*, 293-326.

Sudweeks, R. R., & Tolman, R. R. (1993). Empirical versus subjective procedures for identifying gender differences in science test items. *Journal of Research in Science Teaching*, *30*, 3-19.

Whitaker, T. A., Fitzpatrick, S. J., Williams, N. J., & Dodd, B. G. (2003). IRTGEN: A SAS macro program to generate known trait scores and item responses for commonly used item response theory models. *Applied Psychological Measurement*, *27*, 299-300.