Article

Educational and Psychological
Measurement
2017, Vol. 77(6) 945–970
© The Author(s) 2016
Reprints and permissions:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0013164416657137
journals.sagepub.com/home/epm
SAGE

Using a Multidimensional IRT Framework to Better Understand Differential Item Functioning (DIF): A Tale of Three DIF Detection Procedures

Cindy M. Walker[1] and Sakine Gocer Sahin[2]

Abstract

The theoretical reason for the presence of differential item functioning (DIF) is that data are multidimensional and two groups of examinees differ in their underlying ability distribution for the secondary dimension(s). Therefore, the purpose of this study was to determine how much the secondary ability distributions must differ before DIF is detected. Two-dimensional binary data sets were simulated using a compensatory multidimensional item response theory (MIRT) model, incrementally varying the mean difference on the second dimension between reference and focal group examinees while systematically increasing the correlation between dimensions. Three different DIF detection procedures were used to test for DIF: (1) SIBTEST, (2) Mantel–Haenszel, and (3) logistic regression. Results indicated that even with a very small mean difference on the secondary dimension, smaller than typically considered in previous research, DIF will be detected. Additional analyses indicated that even with the smallest mean difference considered in this study, 0.25, statistically significant differences will almost always be found between reference and focal group examinees on subtest scores consisting of items measuring the secondary dimension.

Keywords

differential item functioning, multidimensional item response theory, validity

[1] University of Wisconsin-Milwaukee, Milwaukee, WI, USA
[2] Hacettepe University, Ankara, Turkey

Corresponding Author:
Cindy M. Walker, The University of Wisconsin-Milwaukee, PO Box 413, Milwaukee, WI 53201, USA.
Email: cmwalker@uwm.edu

The primary goal of administering standardized tests is to differentiate individuals according to their ability to ultimately rank order examinees or draw conclusions about proficiency classifications of examinees. Rank ordering examinees on the basis of their ability level is a unidimensional concept (Ackerman, 1992). Therefore, in both classical test theory (CTT) and traditional item response theory (IRT), it is typically assumed that the ability underlying an individual's test performance is a unidimensional latent trait; however, this assumption is oftentimes violated. In fact, there is typically at least one secondary factor or dimension underlying test performance, in addition to the primary dimension measured by test items, and oftentimes there is more than one additional dimension. Therefore, the reality is that most tests are multidimensional and the secondary factors, or dimensions that influence individuals' performance on particular test items may or may not be related to the dominant dimension (Camilli, 1992). The general cause of differential item functioning (DIF) has been defined to be the presence of multidimensionality in test items; that is, items that function differentially measure at least one dimension in addition to the primary dimension(s) the item is intended to measure (Berk, 1982; Cronbach, 1990; Dorans & Schmitt, 1989; Jensen, 1980; Lord, 1980; Messick, 1989; Roussos & Stout, 1996a; Scheuneman, 1982; Shepard, 1987; Wiley, 1990). DIF occurs when there is an interaction between ability and item characteristics and the occurrence of DIF can affect the conclusions made based on the results of standardized tests that are assumed to be unidimensional (Walker & Beretvas, 2003). The additional, possibly DIF-inducing, dimensions are referred to as secondary dimensions. Roussos and Stout (1996a) categorize each secondary dimension as either an auxiliary dimension, if the secondary dimension is intended to be measured, or as a nuisance dimension, if the secondary dimension is not intended to be measured. DIF that is caused by an auxiliary dimension is referred to as benign DIF; DIF that is caused by a nuisance dimension is be referred to as adverse DIF (Roussos & Stout, 1996a). Ackerman (1992) defines adverse DIF as bias. In both types of DIF, the item characteristic curves for different groups differ due to differences in the underlying ability distribution for some examinees on the secondary dimension(s). According to this perspective, DIF can occur regardless if the secondary dimension is auxiliary or nuisance. However, test validity is impacted by the type of secondary dimension that is the root cause of DIF. An adverse secondary dimension threatens the validity of a test while a benign secondary dimension may not (Walker & Beretvas, 2001). The theoretical reason for the presence of DIF is that an item measures at least one secondary dimension, in addition to the primary dimension being measured by the test, and two groups of examinees differ in their underlying ability distribution for the secondary dimension(s). If two groups do not differ in the secondary underlying ability distributions being measured then items that measure those secondary dimensions will not display DIF. This is because the reason that items function differentially is based on the properties of the items and examinees as well as the interaction between item and examinee characteristics.

One approach in the literature, to aid in our understanding of the most likely cause, or probable sources, of DIF is the multidimensionality-based DIF (MMD) paradigm (Ackerman, 1992; Furlow, Ross, & Gagne, 2009; Gierl, Bisanz, Bisanz, Boughton, & Khaliq, 2001; Roussos & Stout, 1996a). This paradigm is composed of two steps: (1) statistical analysis and (2) substantive analysis. Although great progress has been made in the number of statistical procedures available to test for DIF, progress of the same type has not been made, in terms of the interpretation or the evaluation of substantive processes that may help us to understand *why* DIF occurs, due to the inconsistencies that occur among experts when conducing substantive analyses, post hoc. The MMD paradigm, developed by Roussos and Stout (1996a), which is based on a multidimensional IRT (MIRT) model as recommended by Shealy and Stout (1993), has helped to fill the void between statistical and substantive DIF analytic procedures.

According to Shealy and Stout's MMD, the presence of DIF can be explained by the following formula:

$$E_{\mathrm{R}}(\eta|\theta) - E_{\mathrm{F}}\left(\eta|\theta\right) = (\mu_{\eta_{\mathrm{R}}} - \mu_{\eta_{\mathrm{F}}}\Big) + \theta\left(\rho_{\mathrm{R}}\frac{\sigma_{\eta_{\mathrm{R}}}}{\sigma_{\theta_{\mathrm{R}}}} - \rho_{\mathrm{F}}\frac{\sigma_{\eta_{\mathrm{F}}}}{\sigma_{\theta_{\mathrm{F}}}}\right) + \left(\mu_{\theta_{\mathrm{F}}}\rho_{\mathrm{F}}\frac{\sigma_{\eta_{\mathrm{F}}}}{\sigma_{\theta_{\mathrm{F}}}} - \mu_{\mathrm{R}}\rho_{\mathrm{R}}\frac{\sigma_{\eta_{\mathrm{R}}}}{\sigma_{\theta_{\mathrm{R}}}}\right) \tag{1}$$

Specifically, this formula depicts the expected difference between the secondary abilities, η, of individuals from the two groups, which can vary, when the primary ability level is held constant. It is typically assumed that the common distribution for the (θ, η) trait vector is normally distributed. In Equation (1), $\mu_{\theta\mathrm{G}}$ and $\mu_{\eta\mathrm{G}}$ represent the means for the primary and secondary distributions, respectively, for group G (reference or focal group); while $\sigma_{\theta\mathrm{G}}$ and $\sigma_{\eta\mathrm{G}}$ represent the standard deviations for the primary and secondary distributions, respectively, for group G; and ρ_{G} represents the correlation between the primary and secondary dimensions for group G. When the two groups have the same standard deviations ($\sigma_{\theta_{\mathrm{F}}} = \sigma_{\theta_{\mathrm{R}}}$ and $\sigma_{\eta_{\mathrm{F}}} = \sigma_{\eta_{\mathrm{R}}}$) and the same correlations ($\rho_{\mathrm{R}} = \rho_{\mathrm{F}}$) and when $\sigma_\eta = \sigma_\theta$, Formula (1) becomes

$$E_{\mathrm{R}}(\eta|\theta) - E_{\mathrm{F}}(\eta|\theta) = \left(\mu_{\eta_{\mathrm{R}}} - \mu_{\eta_{\mathrm{F}}}\right) - \rho\left(\mu_{\theta_{\mathrm{R}}} - \mu_{\theta_{\mathrm{F}}}\right) \tag{2}$$

Based on these two formulas, the probable causes for the DIF are explicated by Ackerman (1992) and Roussos and Stout (1996a) in the following manner. If $\mu_{\eta_{\mathrm{R}}} \neq \mu_{\eta_{\mathrm{F}}}$ in Equation (2), then DIF is likely to occur when the term $\rho(\mu_{\theta_{\mathrm{R}}} - \mu_{\theta_{\mathrm{F}}})$ does not have the same sign and magnitude as the term $\mu_{\eta_{\mathrm{R}}} - \mu_{\eta_{\mathrm{F}}}$. When the term $\rho(\mu_{\theta_{\mathrm{R}}} - \mu_{\theta_{\mathrm{F}}})$ has the same sign as the term $\mu_{\eta_{\mathrm{R}}} - \mu_{\eta_{\mathrm{F}}}$ then the secondary dimension is less likely to result in the presence of DIF. However, DIF is more likely to occur when these two terms have opposite signs. Therefore, the presence of DIF depends on both the sign and the magnitude of the difference between the averages for the primary and the secondary dimensions *and* on the size of the correlation between the two dimensions. Interestingly, according to this model, even if $\mu_{\eta_{\mathrm{R}}} \approx \mu_{\eta_{\mathrm{F}}}$ just the existence of a secondary dimension can case DIF to occur if the correlation between the primary and secondary dimensions are not the same for the two groups and there are differences in the underlying means for the two groups on the primary dimension being measured by test items (i.e., $\mu_{\theta R} \neq \mu_{\theta F}$).

The MMD model, depicted in Equations (1) and (2), also helps us understand when DIF will not occur. Specifically, DIF will not occur when an item is not sensitive to a secondary dimension. DIF will also not occur if the two groups do not differ in their underlying ability distributions on the secondary dimension, (i.e., $\mu_{\eta_R} \approx \mu_{\eta_F}$), as long as either the correlation between the primary and secondary dimension is zero (which is quite unlikely in practice) or $\mu_{\theta_R} \approx \mu_{\theta_F}$. If the underlying mean on the primary dimension is not equivalent for the two groups, then DIF will not occur, as long as the two groups have been equated, in terms of their distribution on the primary dimension.

In addition to helping researchers understand the theoretical rationale for, and underlying causes of DIF, the MMD framework can help us to understand the relationship between MIRT and DIF. In fact, Ackerman (1991) listed four situations that might result in the occurrence of DIF when considering Shealy-Stout's MMD: (1) The primary trait means differ, provided there is a correlation between primary and nuisance traits. (2) The nuisance trait means differ. (3) The ratio of the variance of primary trait to that of nuisance trait is not the same for both groups. (4) The correlation of the primary and nuisance trait is not the same for both groups. Several simulation studies have been conducted that utilize this framework and aid in our understanding of MIRT and DIF. For example, a study conducted by Furlow et al. (2009) considered differences in the mean on the secondary dimension, as well as in the correlation between dimensions when testing bundles of items for DBF using SIBTEST. The results of this study indicated that the power of SIBTEST was generally higher when items measuring the secondary dimension were more highly discriminating than those measuring the primary dimension and as the secondary ability mean difference increased. Similarly, a study conducted by Oshima and Miller (1992) considered underlying mean differences in the secondary dimension, stating that "The mean difference on the nuisance trait is probably the most important source of potential bias" (p. 238). The results of this study indicated that IRT-based DIF detection procedures were able to differentiate between bias, defined as differences caused by secondary ability distribution differences, and impact, defined as between-group differences on the primary trait when testing multidimensional items.

While some previous research has helped to determine the impact of different distributional characteristics on the power of different DIF detection, no one study has compared the impact for different DIF detection methods. Moreover, previous studies have not yet determined at what point DIF will occur given different correlation between the primary and secondary dimensions and differences in secondary ability distributions. In fact, most of the previous research (Monahan & Ankenmann, 2005; Pei & Li, 2010; Woods, 2008) pertaining to DIF utilizes a unidimensional IRT model to generate the data, which is not in perfect alignment with the MMD framework for understanding why DIF occurs, and there are no studies in the literature that examine the impact of multidimensionality on DIF detection when manipulating the underlying secondary ability distributions of both reference group and focal group examinees simultaneously. Finally, no previous studies have considered the relationship between DIF and score differences, which might be caused by the presence of DIF, which explicitly links the presence of DIF to the achievement gap. Since the presence of DIF is a function of the interdimensional correlations, as well as the mean differences in the underlying ability distributions, the purpose of this study was to examine these factors to determine how much the underlying secondary ability distributions must differ before DIF is detected, resulting in measurement invariance for the two groups which, in turn, can result in erroneous decisions being made about examinees rank ordering or proficiency classifications (Walker & Beretvas, 2003).

Method

The purpose of this study was to determine the magnitude of dimensionality that will result in statistically significant DIF, since theoretically the reason for the occurrence of DIF is multidimensionality. For this study a 30-item test, 25 of which primarily measured θ_1 and 5 of which primarily measured θ_2, was generated using fixed item parameters. The item parameters were selected to span the θ_1, θ_2 space as shown in Table 1. The same item parameters were used for both the focal group and the reference group.

In MIRT, items can be represented by item vectors in a Cartesian coordinate system. Each item vector is on a line that crosses the origin. The direction of the vector is defined as the vector's angle with positive θ_1 axis. The direction of item *i* is calculated with the following equation:

$$\alpha_i = \text{arc cos} \frac{\alpha_{i1}}{\sqrt{\alpha_{i1}^2 + \alpha_{i2}^2}} \tag{3}$$

Items that are closer to the θ_1 axis primarily measure the θ_1 ability while items that are closer to θ_2 axis primarily measure the θ_2 ability. Items have an angle of 45° with both ability axes equally measure both of the abilities (Ackerman, 1994; Ackerman, Gierl, & Walker, 2003). Accordingly, in this study, the item parameters in Table 1 were chosen such that the angles for the first five items primarily measure θ_2, with angular distances ranging between 70° and 85°; while the angles for the other 25 items primarily measure θ_1, with angular distances ranging between 5° and 20°. This is because a_2 is dominant for the first 5 items, whereas for the other 25 items a_1 is dominant. The discrimination parameters for all items range between 0.087 and 1.679. From a practical perspective, the test that was simulated might be thought of as a multidimensional math test with complex items. When the correlation between dimensions is low, the scale could reflect a math test that has five open-ended items that require an examinee to explain the reasoning or their approach to the solution which would require a high ability to communicate in writing. When the correlation between dimensions is high, the scale could reflect a math test with five items that

Table 1. The Item Parameters Used in This Study.

	Item	a_1	a_2	Angle
Primarily measure θ_2	1	0.434	1.194	70.00
	2	0.303	1.038	73.75
	3	0.365	1.649	77.50
	4	0.203	1.317	81.25
	5	0.087	0.997	85.00
Primarily measure θ_1	6	1.265	0.111	5.00
	7	1.076	0.106	5.63
	8	1.679	0.184	6.25
	9	1.323	0.16	6.88
	10	0.992	0.131	7.50
	11	0.953	0.136	8.13
	12	0.924	0.142	8.75
	13	1.199	0.198	9.38
	14	1.116	0.197	10.00
	15	0.8	0.15	10.63
	16	1.319	0.262	11.25
	17	1.079	0.227	11.88
	18	0.818	0.181	12.50
	19	0.907	0.211	13.13
	20	0.926	0.227	13.75
	21	1.481	0.38	14.38
	22	1.035	0.277	15.00
	23	0.776	0.217	15.63
	24	0.796	0.232	16.25
	25	1.192	0.362	16.88
	26	0.965	0.304	17.50
	27	1.101	0.36	18.13
	28	0.929	0.315	18.75
	29	1.575	0.554	19.38
	30	0.987	0.359	20.00

require quite a bit more reading comprehension skills than the remaining items on the test.

These item parameters were used to generate two-dimensional binary data sets using SAS/IML. The data sets were generated using the compensatory multidimensional two-parameter logistic model (Reckase, 2009) presented in Equation (4).

$$P(X_{ij}=1|\theta_j,a_i,d_i=\frac{e^{\left(\sum_{l=1}^{m}a_i\grave{e}_j+d_i\right)}}{1+e^{\left(\sum_{l=1}^{m}a_i\grave{e}_j+d_i\right)}}\tag{4}$$

where X_{ij} represents the score (0,1) on item i person j, a_i slope parameters associated with item i, d_i intercept term of item i, and θ_j ($\theta_j = \{\theta_1, . . ., \theta_M\}$ is the vector of ability parameters represents a vector of multiple represents a scalar difficulty. Sample size was fixed such that $n_f = n_r$ = 1,000 examinees, and the test length was fixed at 30 items.

As previously stated, according to Equation (2), when the θ distribution of the focal and the reference groups is constant, the expected differences for the conditional distributions are dependent on the average of the primary and the secondary distributions—that is to say, on the average of the distributions of θ and η. In Nandakumar's (1993) simulation study, two levels of η were chosen, 0.5 and 1.0, to represent moderate-to-large degrees of DIF. In a study conducted by Russell (2005), the average difference between focal and reference groups on the primary trait varied among $d_\theta = 0$ (no mean difference), $d_\theta = 0.5$ (moderate differences), and $d_\theta = 1.0$ (large differences). In the study by Furlow et al. (2009), the average of the underlying ability distributions for the secondary dimensions that were considered were (0.0, 0.25, 0.50, 0.75, 1.0) to disadvantage only the focal group. This study expands on these design factors by using conditions similar to Furlow et al. but manipulating the mean for both the reference and focal groups. In all cases the underlying distribution on the primary dimension was simulated as standard normal distribution. Five different means (0.0, −0.25, −0.50, −0.75, −1.0) were considered for the underlying ability distribution for the focal group on the secondary dimension; while five different means (0.0, 0.25, 0.50, 0.75, 1.0) were considered for the underlying ability distributions for the reference group on the secondary dimension. These conditions were crossed to gradually increase and decrease the difference between the means of the distributions of the focal and the reference groups on the secondary dimension. This results in the simulation of DIF due to multidimensionality.

Another factor that is known to influence the occurrence of DIF is the correlation between the primary and the secondary dimensions. Therefore, this study also considered the impact of different correlations between dimensions, in terms of DIF detection. Four different correlations between dimensions were considered to reflect no correlation $\rho = 0.0$, a low correlation $\rho = 0.25$, a medium correlation $\rho = 0.5$, and a high correlation $\rho = 0.75$. To summarize, five different underlying ability distributions were considered for the focal group; five different distributions were considered for the reference group, and four different correlational conditions were considered. Each of these factors in the design was fully crossed, resulting in a 100-cell (5 * 5 * 4) design. All combinations of focal and reference group means on the secondary dimensions used in that study. For each condition, 100 replications were conducted.

Three different DIF detection procedures were used to analyze the data: (1) SIBTEST, (2) logistic regression, and (3) Mantel–Haenzsel. For all three procedures, Type I error rates were calculated for non-DIF items and power rates were calculated for DIF items. Then a random-effects analysis of variance (ANOVA) was conducted to determine the impact of the factors in the study, collapsing across the five items that were simulated to primarily measure the second dimension. A summary of the conditions studied in this research are illustrated in Table 2.

Table 2. Research Design.

Factors considered Fully crossed design			
Correlation between θ_1 and θ_2	Mean of θ_2 for focal group	Mean of θ_2 for reference group	DIF Detection Procedure
$\rho = 0.00$	0.00	0.00	SIBTEST—Exploratory
$\rho = 0.25$	−0.25	0.25	SIBTEST—Confirmatory
$\rho = 0.50$	−0.50	0.50	Mantel–Haenszel
$\rho = 0.75$	−0.75	0.75	Logistic regression
	−1.00	1.00	

In addition, to determine the substantive practical implications of this study, in terms of whether multidimensionality that results in DIF could result in achievement gap differences, *t* tests were conducted to determine if there were statistically significant differences between reference and focal group examinees, for both total test scores and subtest scores on the five items that were simulated to primarily measure the second dimension.

Results

Type I Error and Power Rates

When using SIBTEST to test for DIF, one can conduct exploratory studies at the item level and/or confirmatory studies at the bundle level. Therefore, both types of analyses were considered, given that five items were simulated to primarily measure θ_2. When exploratory item level analyses were conducted, the conditioning subtest used consisted of only those items that were simulated to primarily measure θ_1. This is comparable to what would happen in practice, if one suspected that a bundle of items were measuring a benign secondary dimension that might be affecting performance.

Table 3 depicts the Type I error rates and Table 4 depicts the power rates for all of the factors explored in this study when using exploratory and confirmatory (bundle level) SIBTEST, Mantel–Haenszel, and logistic regression at the item level. The results are collapsed across the five items simulated to primarily measure θ_2. Type I error rates occur when there are no distributional differences between the reference and focal group. Therefore, as Table 3 illustrates, the Type I error rates for SIBTEST are slightly above the nominal 0.05 level for both the exploratory and confirmatory conditions, at 0.06 and 0.08, respectively, only when the correlation between the two dimensions is 0. As the correlation between dimensions increases, the Type I error rate decreases and reflects the nominal 0.05 level with a correlation between dimensions as small as 0.25. Interestingly, while the Type I

Table 3. Type I Error When Using Exploratory and Confirmatory (at the Bundle Level) SIBTEST, Mantel–Haenszel, and Logistic Regression to Test for DIF.

Correlation between dimensions	SIBTEST	Confirmatory SIBTEST	Mantel–Haenszel	Logistic regression
ρ = 0.00	0.06	0.08	0.06	0.06
ρ = 0.25	0.05	0.03	0.04	0.05
ρ = 0.50	0.06	0.06	0.05	0.06
ρ = 0.75	0.05	0.03	0.04	0.05

error rate is higher when the correlation between dimensions is 0 for the confirmatory bundle DIF analyses than for the exploratory item level analyses, it is slightly lower for the confirmatory analyses with a nonzero correlation between dimensions. Type I error rates results for Mantel–Haenszel are almost identical to the results obtained when using SIBTEST to test for DIF at the exploratory item level. The Type I error rate is slightly inflated when the correlation between dimensions is zero and is slightly below the nominal 0.05 level with a nonzero correlation between dimensions. Type I error rates for logistic regression are almost identical to the results obtained when using SIBTEST or Mantel–Haenszel to test for DIF at the exploratory item level. However, there are slight differences in that the Type I error rate does not decrease as the correlation between dimensions increases, as was observed with the other two DIF detection procedures.

As seen in Table 4, when exploratory item level DIF analyses are conducted using SIBTEST power reaches more than acceptable levels when the difference between the mean of the second dimensions for the reference and focal group is at least 0.5. However, for confirmatory bundle analyses power is more than acceptable when the difference between the mean of the second dimensions for the reference and focal group is only 0.25, likely due to amplification (Nandakumar, 1993). The correlation between dimensions had little to no impact on the power rates obtained when using SIBTEST in either an exploratory or confirmatory manner.

Power rates results for Mantel–Haenszel are almost identical to the results of exploratory SIBTEST. One exception is that SIBTEST is slightly more powerful when the correlation between dimensions is zero and the difference between the mean on the second dimension for the reference and focal group is only 0.25. Moreover, the power is more than acceptable when the difference between the mean of the second dimensions for the reference and focal group is at least 0.5. When compared to the other two DIF detection procedures, the power rates obtained when using logistic regression to test for DIF are always slightly lower than those obtained from SIBTEST or Mantel–Haenszel when the difference between the mean of the second dimensions for the reference and focal group is less than 0.5.

Table 4. Power Rates When Using Exploratory and Confirmatory (at the Bundle Level) SIBTEST, Mantel-Haenszel, and Logistic Regression to Test for DIF.

		θ_2 of reference group															
		0.25				0.50				0.75				1.00			
Correlation between dimensions	θ_2 of focal group	SIBTEST	Confirmatory SIBTEST	Mantel–Haenszel	LogiStic Regression	SIBTEST	Confirmatory SIBTEST	Mantel-–aenszel	Logistic regression	SIBTEST	Confirmatory SIBTEST	Mantel–Haenszel	Logistic regression	SIBTEST	Confirmatory SIBTEST	Mantel–Haenszel	Logistic regression
$\rho = 0.00$	0.00	0.65	0.99	0.61	0.54	0.99	—	0.99	0.54	—	—	—	—	—	—	—	—
	–0.25	0.99	—	0.99	0.98	—	—	—	0.98	—	—	—	—	—	—	—	—
	–0.50	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
	–0.75	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
	–1.00	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
$\rho = 0.25$	0.00	0.64	0.97	0.60	0.54	0.99	—	0.98	0.54	—	—	—	—	—	—	—	—
	–0.25	0.99	—	0.99	0.97	—	—	—	0.97	—	—	—	—	—	—	—	—
	–0.50	—	—	—	0.99	—	—	—	0.99	—	—	—	—	—	—	—	—
	–0.75	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
	–1.00	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
$\rho = 0.50$	0.00	0.57	0.99	0.54	0.47	0.99	—	0.99	0.47	—	—	—	—	—	—	—	—
	–0.25	0.98	—	0.97	0.94	—	—	—	0.94	—	—	—	—	—	—	—	—
	–0.50	—	—	—	0.99	—	—	—	0.99	—	—	—	—	—	—	—	—
	–0.75	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
	–1.00	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
$\rho = 0.75$	0.00	0.58	0.97	0.57	0.48	0.97	—	0.97	0.48	—	—	—	—	—	—	—	—
	–0.25	0.98	—	0.98	0.95	—	—	—	0.95	—	—	—	—	—	—	—	—
	–0.50	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
	–0.75	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
	–1.00	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—

Impact of Experimental Design Factors Considered

In order to determine the impact of the design factors considered in this study four random effects ANOVA models were fit to the data, one for each of the three exploratory single-item DIF detection procedures considered and a fourth for using SIBTEST in a confirmatory manner to test the bundle of items. The statistic obtained from the DIF detection procedure was used as the dependent variable, and the correlation between dimensions, mean of the reference group and mean of the focal group were used as the independent variables and modeled as random factors. Once again, these analyses pertained to the three different DIF detection procedures that were done at the item level in an exploratory manner and the additional analyses that were conducted in a confirmatory manner using SIBTEST to evaluate the five-item bundle.

The results obtained when using SIBTEST in an exploratory manner to conduct item level analyses indicated that the three-way interaction term was significant ($F_{48,\ 12} = 10.99, p < .001$). However, this term explained only a very small proportion of the variability in the data (partial $\eta^2 = 0.01$). On the other hand, two of the two-way interaction terms, those that included the correlational factor of the design were statistically significant and also explained a large proportion of the variability in the data, while the third two-way interaction term was not statistically significant ($F_{16,\ 48} = 1.07, p = .41$). Specifically, the correlation by reference group mean interaction term explained approximately 68% of the variability in the data and was statistically significant ($F_{12,\ 48} = 8.38, p < .001$). Similarly, the correlation by focal group mean interaction terms explained approximately 59% of the variability in the data and was also statistically significant ($F_{12,\ 48} = 5.81, p < .001$). These interaction effects are depicted in Figures 1 and 2, respectively. As the figures illustrate, changes in the reference group and focal group marginal mean had the greatest impact on the beta statistics. As the absolute value of these marginal means increased so did the beta statistics, as expected. In fact, even when the absolute value of the marginal mean of the reference or focal group was only 0.25, the absolute value of the average beta statistic was greater than 0.1, which would be considered a large effect size using conventional guidelines. These guidelines state that large DIF occurs when $|\hat{\beta}| \geq 0.088$ (Roussos & Stout, 1996b). However, even the largest marginal mean beta statistic obtained for these conditions, 0.30, which was obtained when the reference or focal group mean was 1.0 or -1.0 and the correlation between dimensions was 0.0, would likely not be a cause for concern using more contemporary effect size guidelines for DIF (Walker, Zhang, Banks, & Cappaert, 2011). These guidelines state that only when the proportion of DIF reaches 0.15 will there likely be ability differences found between reference and focal group examinees and a beta statistic of 0.3 for a one item bundle only results in a proportion of DIF of 0.06. The impact of the correlational factor of the design was far less than that for the reference and focal group means. As the correlation between dimensions increased, the beta statistics tended to decrease, but not to a very great extent.

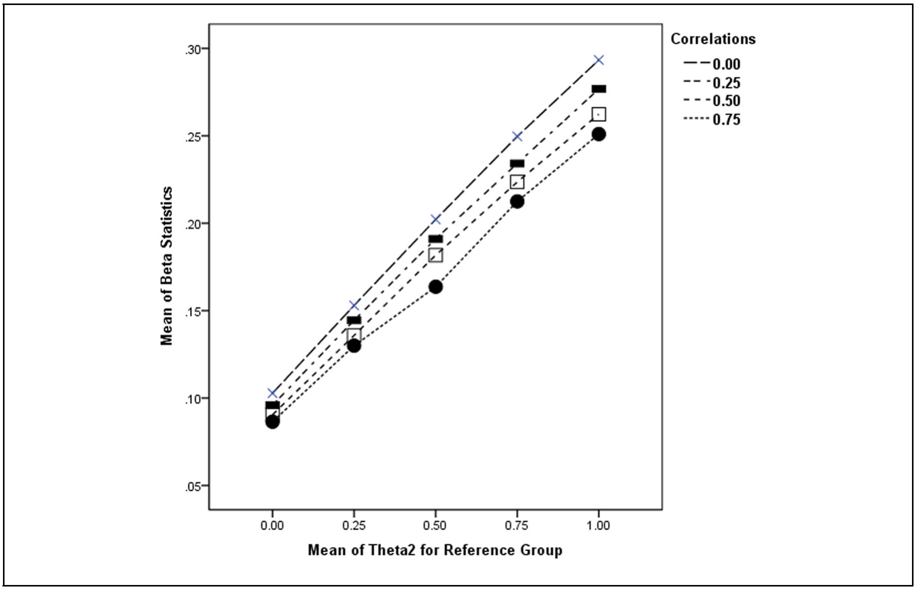

Figure 1. Interaction effect of correlation between dimensions and reference group mean when using SIBTEST to test for DIF in an exploratory manner.

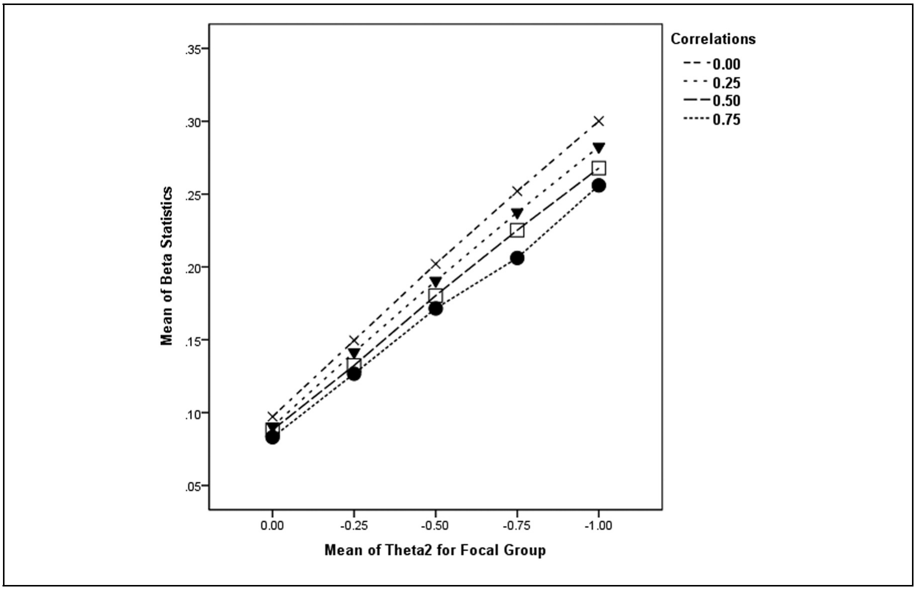

Figure 2. Interaction effect of correlation between dimensions and focal group mean when using SIBTEST to test for DIF in an exploratory manner.

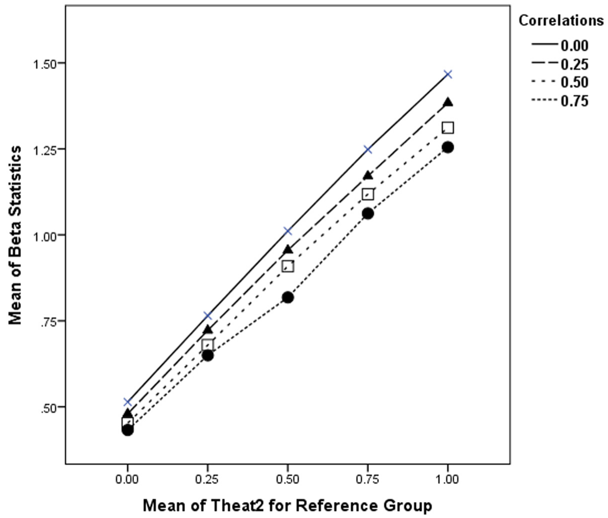

Figure 3. Interaction effect of correlation between dimensions and reference group mean when using SIBTEST to test for DIF in a confirmatory manner.

The results obtained when using SIBTEST in a confirmatory manner were similar to those obtained when using SIBTEST in an exploratory manner. Specifically, the three-way interaction term was significant ($F_{48,\ 49{,}900} = 74.67$, $p < .001$) but only explained a very small proportion of the variability in the data (partial $\eta^2 = 0.07$). Once again, two of the two-way interaction terms, those that included the correlational factor of the design were statistically significant and also explained a large proportion of the variability in the data, while the third two-way interaction term was not statistically significant ($F_{16,\ 48} = 1.06$, $p = .41$). Specifically, the correlation by reference group mean interaction term explained approximately 68% of the variability in the data and was statistically significant ($F_{12,\ 48} = 8.37$, $p < .001$). Similarly, the correlation by focal group mean interaction terms explained approximately 59% of the variability in the data and was also statistically significant ($F_{12,\ 48} = 5.81$, $p < .001$). These interaction effects are depicted in Figures 3 and 4, respectively, and are very comparable to what was obtained for the exploratory DIF analyses. Unlike what was found in the exploratory DIF analyses, the largest marginal mean obtained for the confirmatory SIBTEST analyses was much larger than that obtained for the exploratory analyses. Specifically, the largest marginal mean beta statistic obtained for these conditions was 1.5, which was much larger than the comparable marginal mean beta statistic for the exploratory analyses of 0.3. This marginal mean beta statistic was also obtained when the reference or focal group mean was 1.0 or −1.0 and the correlation between dimensions was 0.0. More important, this results in a proportion of DIF of

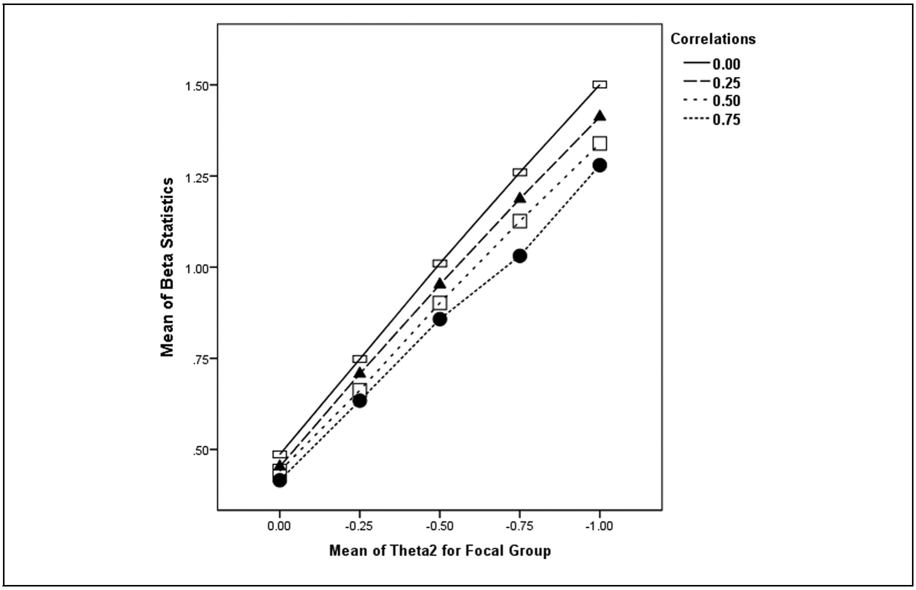

Figure 4. Interaction effect of correlation between dimensions and focal group mean when using SIBTEST to test for DIF in a confirmatory manner.

approximately 30% which would be a cause for concern according to the more contemporary DBF effect guidelines proposed by Walker et al. (2011).

The results obtained when using Mantel–Haenszel to test for DIF differed from those obtained when using SIBTEST in either an exploratory or confirmatory manner. For this analytic procedure, the correlation had little impact on the results. Once again, the three-way interaction term was significant ($F_{48,\ 49{,}900} = 4.98$, $p < .001$) but only explained a very small proportion of the variability in the data (partial $\eta^2 = 0.01$). However, for this analytic procedure only the reference by focal group interaction term was statistically significant ($F_{48,\ 16} = 143.78$, $p < .001$) and explained a large proportion of the variability in the data (partial $\eta^2 = 0.98$). The two interaction terms that included the correlation were non-significant for both the reference group mean and the focal group mean ($F_{48,\ 12} = 0.38$, $p = .97$ and $F_{48,\ 12} = 0.30$, $p = .99$), respectively. Figure 5 depicts the interaction between the reference and focal group mean. As the figure illustrates when only the focal group mean differs from zero the statistics increase, in a fairly linear manner. However, as both the reference and focal group means differ more and more from zero, and it seems as if a 0.75 combined mean difference from zero is the tipping point, these statistics reach their maximum values obtained, between 40 and 60. This graph is reflective of the symmetry of the design.

The results obtained when using logistic regression to test for DIF differed from both the results obtained when using SIBTEST and those obtained when using

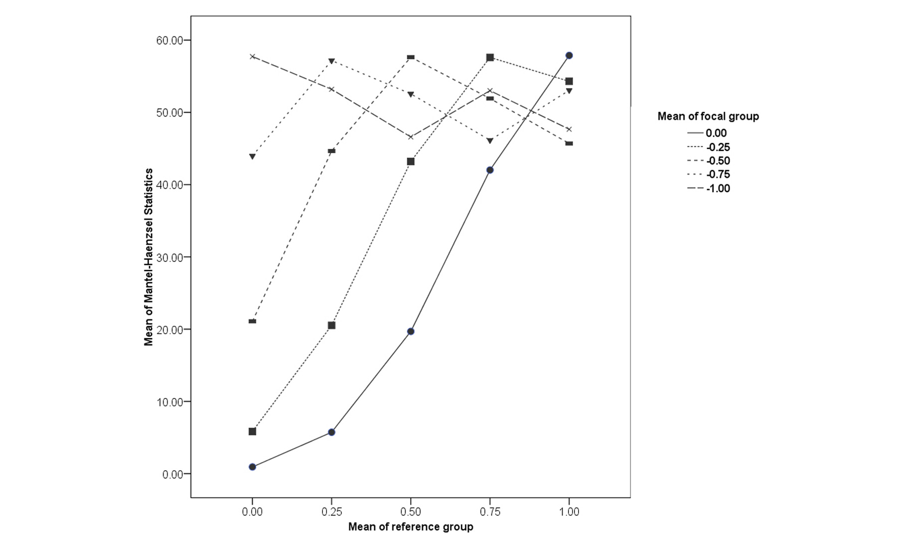

Figure 5. Interaction effect of reference group mean and focal group mean when using Mantel–Haenszel to test for DIF.

Mantel–Haenszel to test for DIF. Once again the three-way interaction term was significant ($F_{48,\ 49,900} = 6.09$, $p < .001$) but only explained a very small proportion of the variability in the data (partial $\eta^2 = 0.01$). However, for this analytic procedure all of the two-way interaction terms were significant and explained a large proportion of the variability in the data. The two-way interaction term for the reference and focal group mean explained the greatest variability in the data ($F_{48,\ 16} = 93.35$, $p < .001$) and explained a large proportion of the variability in the data, while the third two-way interaction term was not statistically significant ($F_{16,\ 48} = 1.07$, $p = .41$). However, the correlation by reference group mean interaction term also explained a large proportion of variability (partial $\eta^2 = 0.55$) and was statistically significant ($F_{12,\ 48} = 4.97$, $p < .001$). Similarly, the correlation by focal group mean interaction terms explained approximately 51% of the variability in the data and was also statistically significant ($F_{12,\ 48} = 4.19$, $p < .001$). Figures 6 through 8 illustrate these interaction effects, with Figures 6 and 7 illustrating the interaction effects that included the correlational design factor and Figure 8 illustrating the reference group mean by focal group mean interaction effect. Figures 6 and 7 are somewhat comparable to what was observed when using SIBTEST to detect DIF. Specifically, changes in the reference group and focal group marginal mean had the greatest impact on the statistics and as the absolute value of these marginal means increased so did the statistics, as expected. Moreover, the statistics decreased as the correlation between dimensions increased; however, the impact of correlational changes was not as dramatic as the

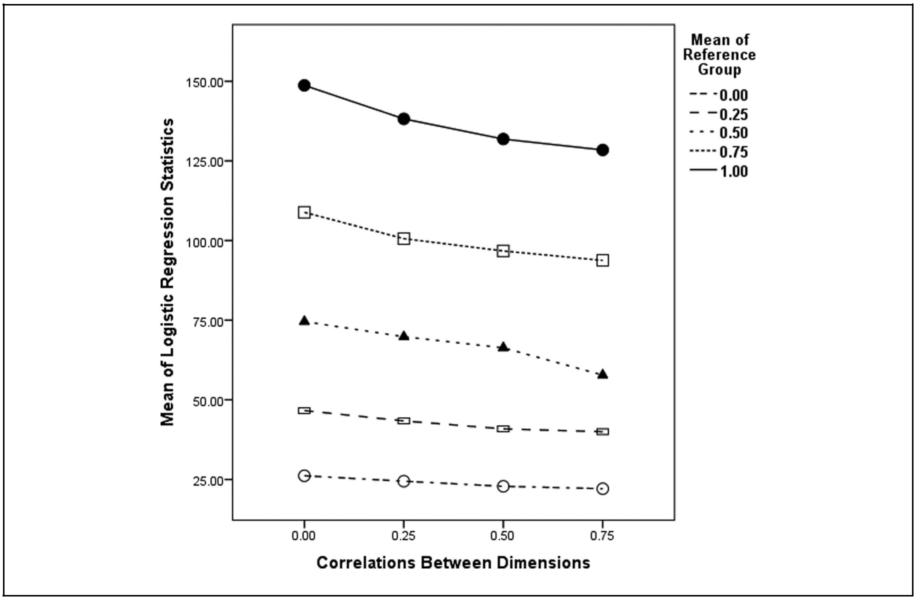

Figure 6. Interaction effect of reference group mean and correlation between dimensions when using logistic regression to test for DIF.

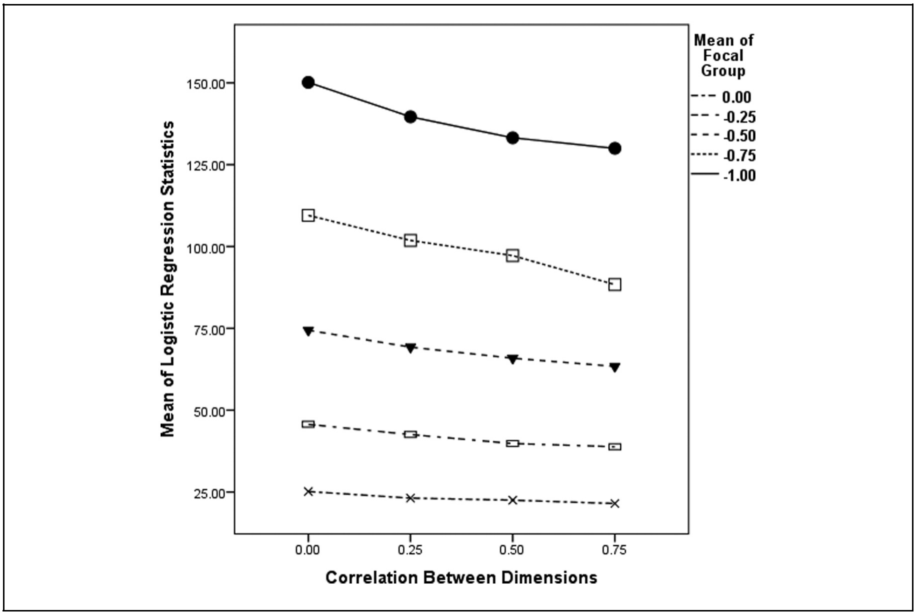

Figure 7. Interaction effect of focal group mean and correlation between dimensions when using logistic regression to test for DIF.

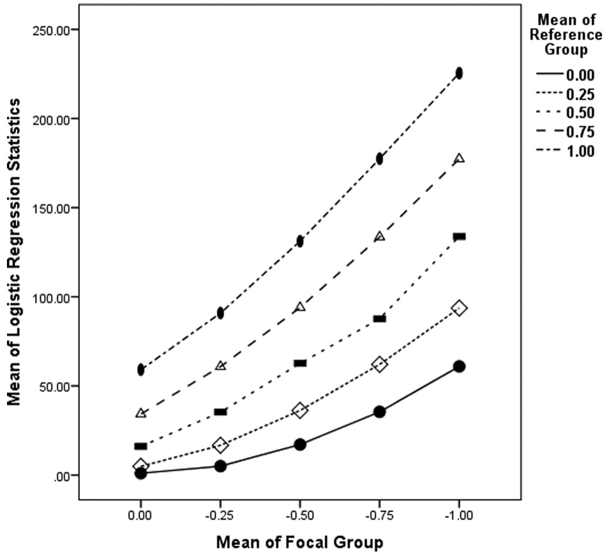

Figure 8. Interaction effect of reference group mean and focal group mean when using logistic regression to test for DIF.

impact of mean differences. Figure 8, which illustrates the interaction effect of changes in reference and focal group means, shows a much clearer pattern than what was observed when using Mantel–Haenszel to detect DIF. In fact, there appears to be an exponential relationship, such that as the mean of the reference group increases and the focal group decreases, so do the test statistics, in an exponential manner.

Implications for Practitioners

In order to determine the practical implications of these results, *t* tests were conducted between reference and focal group examinees to determine if statistically significant differences existed on total scores, as well as on subscale scores for the five items that primarily measured the secondary dimension. These results can help us to understand the relationship between DIF and the achievement gap, which is important because DIF analyses were first undertaken as a way to address the achievement gap (Angoff, 1993). Statistically significant differences between reference and focal group examinees can be thought of as achievement gap differences that are *not* caused by differences in the primary dimension measured by test items, but rather in the secondary dimension measured by test items, be they nuisance or benign. Table 5 depicts the number of times statistically significant differences were found between reference and focal group examinees. As the table illustrates, statistically significant differences

Table 5. Number of Times Statistically Significant Differences Were Found in Observed Test Scores Between Reference and Focal Group Examinees.

			θ_2 of reference group									
			0		0.25		0.5		0.75		1	
			Subtest	Whole test	Subtest	Whole test	Subtest	Whole test	Subtest	Whole test	Subtest	Whole test
θ_2 of focal group	$\rho = 0.00$	0	3	8	100	47	100	98	100	100	100	100
		–0.25	97	51	100	100	100	100	100	100	100	100
		–0.5	100	98	100	100	100	100	100	100	100	100
		–0.75	100	100	100	100	100	100	100	100	100	100
		–1	100	100	100	100	100	100	100	100	100	100
	$\rho = 0.25$	0	5	5	97	45	100	94	100	100	100	100
		–0.25	99	36	100	90	100	100	100	100	100	100
		–0.5	100	98	100	100	100	100	100	100	100	100
		–0.75	100	100	100	100	100	100	100	100	100	100
		–1	100	100	100	100	100	100	100	100	100	100
	$\rho = 0.50$	0	6	9	99	49	100	94	100	100	100	100
		–0.25	98	42	100	95	100	99	100	100	100	100
		–0.5	100	93	100	100	100	100	100	100	100	100
		–0.75	100	100	100	100	100	100	100	100	100	100
		–1	100	100	100	100	100	100	100	100	100	100
	$\rho = 0.75$	0	6	6	99	31	100	96	100	99	100	100
		–0.25	96	35	100	92	100	100	100	100	100	100
		–0.5	100	94	100	100	100	100	100	100	100	100
		–0.75	100	100	100	100	100	100	100	100	100	100
		–1	100	100	100	100	100	100	100	100	100	100

were found between reference and focal group examinees on the five items that primarily measured the second dimension, even with ability distribution differences as small as 0.25. This is a much smaller difference than has typically been considered in previous research. Ability distribution differences this small will only result in statistically significant findings between reference and focal group examinees on overall test scores about half of the time. Interestingly, the correlation between dimensions had little impact on the decisions that would be made, in terms of the achievement gap. However they did result in lower test statistics, as Figure 9 illustrates. As the correlation between dimensions increased the test statistic obtained decreased, but not enough to result in non-significant findings.

Discussion

In this study, DIF was examined within the framework of multidimensionality. Therefore, the mean of the secondary dimension measured was differentiated for both the focal and reference group. In addition, the effect of the correlation between dimensions was investigated, in terms of its impact on DIF detection. The obtained results were evaluated in accordance with the Type I error and power rates statistics.

In general, the results indicated that when the difference between the mean of the reference and focal groups was at least 0.5, adequate power was obtained for all DIF detection procedures considered in this study. This finding is reflective of findings from previous studies (Furlow et al., 2009; Oshima & Miller, 1992; Russell, 2005). However, unlike other studies, the conditions explored in this study systematically increased the mean on the secondary dimension for reference group examinees while, at the same time, systematically decreasing the mean on the secondary dimension for focal group examinees. Therefore, the results of this study demonstrated that comparable results will be obtained when the sum of the absolute value of the mean differences on the secondary dimension is equivalent to the mean difference for either only reference group or focal group examinees.

Specifically, for all three DIF detection procedures considered, the average statistics obtained in all conditions in which the mean deviates from zero are equivalent are similar. For example, almost identical results are obtained when the mean of the second dimension for the reference group is 0.25 and that for the focal group is -0.75, resulting in a total difference of 1, and when the mean of the second dimension for the reference group is 0.0 and that for the focal group is -1, which also results in a total difference of 1. These results are presented in the appendix (Tables A1-A3). This finding is also highlighted by the ANOVA results which showed that the two-way interaction of mean values for focal and reference group examinees explained a great majority of the variability in the results.

This study also demonstrated how much more powerful it is to do bundle analyses, when using SIBTEST, as opposed to exploratory single-item studies. When using SIBTEST in a confirmatory manner to test a five-item bundle, only having a mean difference of 0.25 resulted in more than adequate power. When conducting single

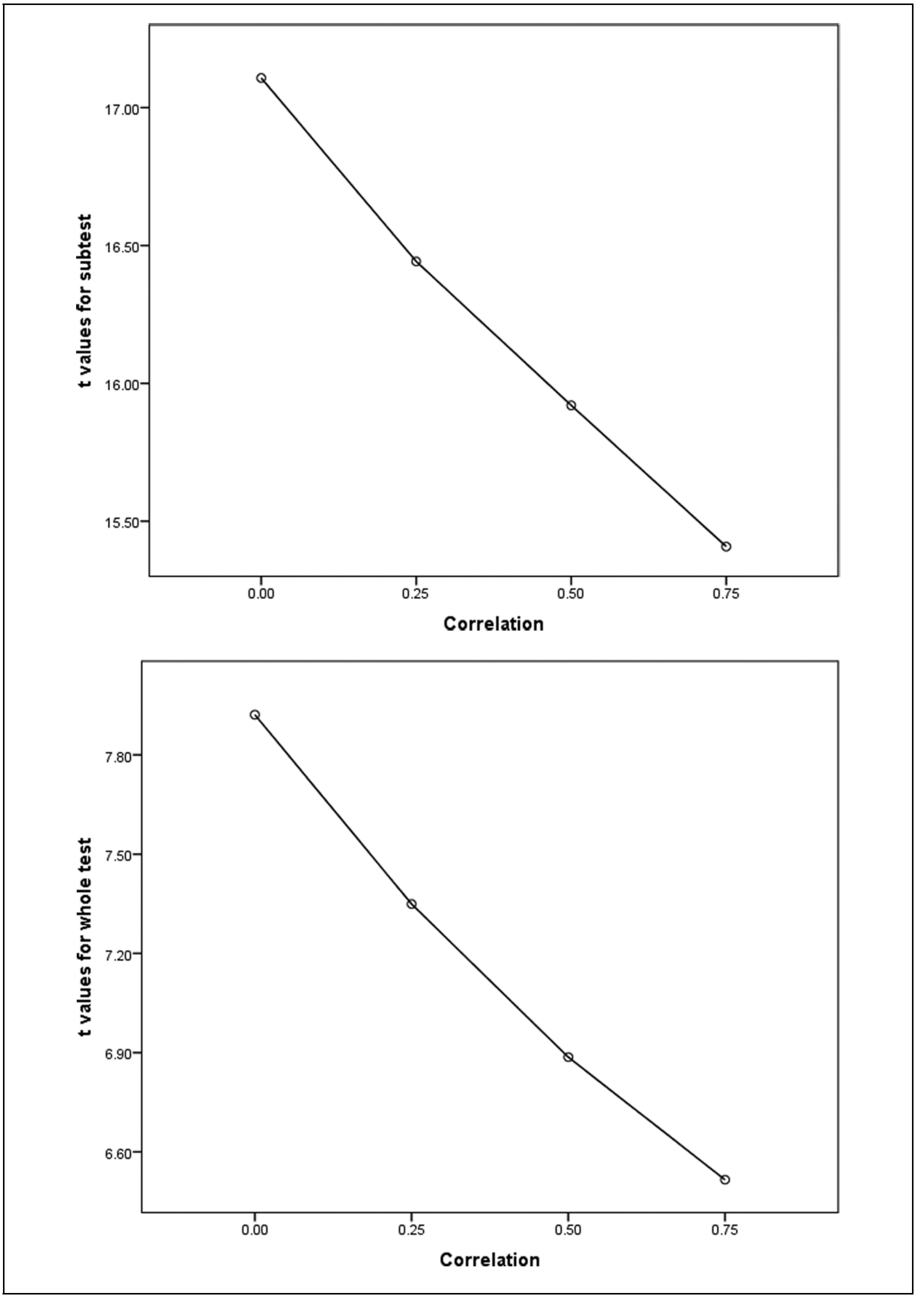

Figure 9. Average *t*-test statistics obtained for different correlational conditions.

item analyses using SIBTEST a mean difference of 0.50 was needed to achieve an acceptable power rate.

Interestingly, increasing the correlation between dimensions had little impact on the power rates, although the Type I error rates decreased a bit as the correlation between dimensions increased for both Mantel–Haenszel and SIBTEST. However, the correlation between dimensions did influence the actual test statistics obtained for SIBTEST and logistic regression, but not to a great extent. This finding is also reflective of previous studies (Oshima & Miller, 1992; Russell, 2005). This is somewhat surprising, knowing that having a higher correlation between dimension does impact the dimensionality of a test.

Kirisci, Hsu, and Yu (2001) reported that a unidimensional IRT model can be used to scale multidimensional data if the correlation between dimensions is higher than 0.4. If the correlation between dimensions is less than 0.4, then a multidimensional model should be used to scale the data. Considering that DIF occurs due to multidimensionality, one might hypothesize that DIF will not occur if the correlation between dimensions is greater than 0.4. However, the results of this study suggest otherwise. In this study, although the magnitude of the DIF statistics obtained increased as the correlation between dimensions decreased, for two of the three DIF detection procedures considered, the power rates were not impacted by increasing the correlation between dimensions. Therefore, DIF occurs even when the correlation between dimensions is relatively high. This finding is similar to the results of previous studies conducted by Lee (2005) and Furlow et al. (2009).

Finally, it was interesting to note that achievement gap differences, defined by statistically significant differences between reference and focal group examinees, were almost always found on total subtest scores, for the five items that were simulated to primarily measure the secondary dimension with distributional differences between the two groups as small as 0.25. While the finding was not as dramatic for overall test score differences, there were still a number of times that achievement gap differences could be accounted for by the interaction between dimensionality and differences on the secondary ability distribution. Future research should explore this phenomenon more thoroughly, utilizing smaller differences between dimensions and varying the number of items measuring the secondary dimension to help us understand the point at which DIF due to multidimensionality can help explain the achievement gap.

Appendix

Table A1. Descriptive Statistics of SIBTEST Results.

	Reference	Focal	Mean		Reference	Focal	Mean		Reference	Focal	Mean		Reference	Focal	Mean		Reference	Focal	Mean		Reference	Focal	Mean
Item 1	0.00	0.00	−0.001	Item 2	0.00	0.00	0.001	Item 3	0.00	0.00	−0.001	Item 4	0.00	0.00	−0.001	Item 5	0.00	0.00	0.001	All items	0.00	0.00	0.000
		−0.25	0.044			−0.25	0.042			−0.25	0.056			−0.25	0.048			−0.25	0.043			−0.25	0.234
		−0.50	0.081			−0.50	0.087			−0.50	0.110			−0.50	0.095			−0.50	0.086			−0.50	0.459
		−0.75	0.120			−0.75	0.128			−0.75	0.161			−0.75	0.140			−0.75	0.127			−0.75	0.674
		−1.00	0.154			−1.00	0.169			−1.00	0.210			−1.00	0.178			−1.00	0.167			−1.00	0.878
	0.25	0.00	0.044		0.25	0.00	0.041		0.25	0.00	0.057		0.25	0.00	0.051		0.25	0.00	0.045		0.25	0.00	0.238
		−0.25	0.086			−0.25	0.085			−0.25	0.112			−0.25	0.101			−0.25	0.090			−0.25	0.475
		−0.50	0.127			−0.50	0.125			−0.50	0.167			−0.50	0.147			−0.50	0.132			−0.50	0.698
		−0.75	0.164			−0.75	0.169			−0.75	0.220			−0.75	0.192			−0.75	0.172			−0.75	0.916
		−1.00	0.198			−1.00	0.210			−1.00	0.266			−1.00	0.229			−1.00	0.210			−1.00	1.112
	0.50	0.00	0.091		0.50	0.00	0.081		0.50	0.00	0.112		0.50	0.00	0.106		0.50	0.00	0.091		0.50	0.00	0.481
		−0.25	0.132			−0.25	0.123			−0.25	0.170			−0.25	0.153			−0.25	0.137			−0.25	0.714
		−0.50	0.174			−0.50	0.166			−0.50	0.224			−0.50	0.201			−0.50	0.177			−0.50	0.944
		−0.75	0.210			−0.75	0.209			−0.75	0.275			−0.75	0.242			−0.75	0.219			−0.75	1.155
		−1.00	0.245			−1.00	0.251			−1.00	0.325			−1.00	0.283			−1.00	0.256			−1.00	1.359
	0.75	0.00	0.133		0.75	0.00	0.118		0.75	0.00	0.164		0.75	0.00	0.155		0.75	0.00	0.133		0.75	0.00	0.704
		−0.25	0.178			−0.25	0.158			−0.25	0.222			−0.25	0.205			−0.25	0.176			−0.25	0.940
		−0.50	0.210			−0.50	0.194			−0.50	0.264			−0.50	0.240			−0.50	0.211			−0.50	1.119
		−0.75	0.256			−0.75	0.246			−0.75	0.331			−0.75	0.296			−0.75	0.263			−0.75	1.391
		−1.00	0.292			−1.00	0.290			−1.00	0.379			−1.00	0.337			−1.00	0.304			−1.00	1.601
	1.00	0.00	0.178		1.00	0.00	0.151		1.00	0.00	0.212		1.00	0.00	0.206		1.00	0.00	0.177		1.00	0.00	0.925
		−0.25	0.220			−0.25	0.194			−0.25	0.270			−0.25	0.255			−0.25	0.220			−0.25	1.159
		−0.50	0.262			−0.50	0.237			−0.50	0.327			−0.50	0.305			−0.50	0.266			−0.50	1.396
		−0.75	0.300			−0.75	0.280			−0.75	0.380			−0.75	0.348			−0.75	0.305			−0.75	1.613
		−1.00	0.336			−1.00	0.323			−1.00	0.430			−1.00	0.388			−1.00	0.345			−1.00	1.822

Table A2. Descriptive Statistics of Mantel–Haenzsel Results.

	Reference	Focal	Mean		Reference	Focal	Mean		Reference	Focal	Mean		Reference	Focal	Mean		Reference	Focal	Mean
Item 1	0.00	0.00	1.000	Item 2	0.00	0.00	0.957	Item 3	0.00	0.00	0.961	Item 4	0.00	0.00	0.783	Item 5	0.00	0.00	0.939
		–0.25	5.313			–0.25	4.642			–0.25	7.957			–0.25	6.064			–0.25	4.659
		–0.50	16.561			–0.50	16.349			–0.50	28.484			–0.50	21.304			–0.50	15.717
		–0.75	35.833			–0.75	34.416			–0.75	59.327			–0.75	46.584			–0.75	33.918
		–1.00	60.012			–1.00	58.942			–1.00	45.139			–1.00	66.777			–1.00	58.553
	0.25	0.00	5.212		0.25	0.00	4.664		0.25	0.00	8.027		0.25	0.00	6.311		0.25	0.00	4.983
		–0.25	17.516			–0.25	16.491			–0.25	28.981			–0.25	22.755			–0.25	16.875
		–0.50	38.378			–0.50	34.009			–0.50	60.157			–0.50	47.848			–0.50	35.633
		–0.75	62.844			–0.75	59.559			–0.75	41.272			–0.75	64.713			–0.75	59.672
		–1.00	56.418			–1.00	60.150			–1.00	57.647			–1.00	36.607			–1.00	60.642
	0.50	0.00	19.291		0.50	0.00	15.639		0.50	0.00	29.213		0.50	0.00	24.237		0.50	0.00	16.997
		–0.25	39.227			–0.25	33.306			–0.25	62.562			–0.25	50.454			–0.25	37.708
		–0.50	66.287			–0.50	57.957			–0.50	37.097			–0.50	66.053			–0.50	60.836
		–0.75	49.051			–0.75	59.095			–0.75	58.177			–0.75	38.068			–0.75	55.192
		–1.00	42.796			–1.00	42.442			–1.00	45.908			–1.00	58.147			–1.00	39.195
	0.75	0.00	39.530		0.75	0.00	32.712		0.75	0.00	60.961		0.75	0.00	51.138		0.75	0.00	35.508
		–0.25	65.946			–0.25	55.053			–0.25	41.169			–0.25	63.240			–0.25	60.474
		–0.50	47.030			–0.50	56.537			–0.50	58.140			–0.50	45.737			–0.50	55.369
		–0.75	47.372			–0.75	38.956			–0.75	46.028			–0.75	58.075			–0.75	40.327
		–1.00	53.311			–1.00	57.028			–1.00	49.023			–1.00	48.566			–1.00	57.328
	1.00	0.00	66.380		1.00	0.00	53.386		1.00	0.00	47.285		1.00	0.00	60.201		1.00	0.00	61.371
		–0.25	48.931			–0.25	64.017			–0.25	57.496			–0.25	41.673			–0.25	53.855
		–0.50	49.892			–0.50	40.833			–0.50	46.547			–0.50	53.646			–0.50	42.089
		–0.75	53.139			–0.75	54.962			–0.75	48.847			–0.75	50.434			–0.75	57.661
		–1.00	48.428			–1.00	45.402			–1.00	49.134			–1.00	48.740			–1.00	46.610

Table A3. Descriptive Statistics of Logistic Regression Results.

	Reference	Focal	Mean		Reference	Focal	Mean		Reference	Focal	Mean		Reference	Focal	Mean		Reference	Focal	Mean
Item 1	0.00	0.00	1.087	Item 2	0.00	0.00	1.025	Item 3	0.00	0.00	1.078	Item 4	0.00	0.00	0.873	Item 5	0.00	0.00	1.033
		−0.25	4.388			−0.25	3.894			−0.25	6.862			−0.25	5.222			−0.25	4.068
		−0.50	12.769			−0.50	13.062			−0.50	23.809			−0.50	17.758			−0.50	13.092
		−0.75	27.552			−0.75	27.015			−0.75	50.348			−0.75	38.646			−0.75	27.920
		−1.00	46.540			−1.00	46.704			−1.00	88.716			−1.00	64.296			−1.00	49.180
	0.25	0.00	4.306		0.25	0.00	3.864		0.25	0.00	6.906		0.25	0.00	5.473		0.25	0.00	4.337
		−0.25	13.538			−0.25	12.934			−0.25	24.082			−0.25	18.886			−0.25	14.040
		−0.50	29.332			−0.50	26.279			−0.50	52.215			−0.50	39.428			−0.50	29.167
		−0.75	48.983			−0.75	45.871			−0.75	91.322			−0.75	68.632			−0.75	49.302
		−1.00	72.910			−1.00	70.983			−1.00	135.486			−1.00	100.494			−1.00	74.468
	0.50	0.00	14.977		0.50	0.00	12.207		0.50	0.00	24.106		0.50	0.00	20.145		0.50	0.00	14.108
		−0.25	29.772			−0.25	25.509			−0.25	53.423			−0.25	41.592			−0.25	30.933
		−0.50	52.455			−0.50	45.555			−0.50	92.720			−0.50	71.667			−0.50	50.953
		−0.75	76.728			−0.75	70.642			−0.75	139.682			−0.75	104.674			−0.75	78.082
		−1.00	105.847			−1.00	99.824			−1.00	196.359			−1.00	146.488			−1.00	107.339
	0.75	0.00	30.106		0.75	0.00	25.130		0.75	0.00	51.079		0.75	0.00	42.066		0.75	0.00	28.896
		−0.25	53.739			−0.25	42.493			−0.25	91.524			−0.25	73.042			−0.25	49.627
		−0.50	75.439			−0.50	62.077			−0.50	127.880			−0.50	101.142			−0.50	71.905
		−0.75	110.302			−0.75	96.583			−0.75	198.786			−0.75	152.730			−0.75	109.337
		−1.00	145.373			−1.00	132.506			−1.00	262.321			−1.00	200.043			−1.00	146.942
	1.00	0.00	53.254		1.00	0.00	40.957		1.00	0.00	85.925		1.00	0.00	73.433		1.00	0.00	51.042
		−0.25	80.433			−0.25	64.095			−0.25	136.152			−0.25	110.854			−0.25	76.797
		−0.50	113.143			−0.50	92.430			−0.50	194.379			−0.50	158.636			−0.50	110.596
		−0.75	148.560			−0.75	126.087			−0.75	260.097			−0.75	207.079			−0.75	144.907
		−1.00	187.024			−1.00	164.211			−1.00	332.758			−1.00	257.538			−1.00	185.699

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

References

Ackerman, T. A. (1991, April). *A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective*. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

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. (1994). Using multidimensional item response theory to understand what items and tests are measuring. *Applied Measurement in Education*, *7*, 255-278.

Ackerman, T. A., Gierl, M. J., & Walker, C. M. (2003). Using multidimensional item response theory to evaluate educational and psychological tests. *Educational Measurement: Issues and Practice*, *22*(3), 37-51.

Angoff, W. H. (1993). Perspectives on differential item functioning methodology. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 3-23). Hillsdale, NJ: Lawrence Erlbaum.

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

Camilli, G. (1992). A conceptual analysis of differential item functioning in terms of a multidimensional item response model. *Applied Psychological Measurement*, *16*, 129-147.

Cronbach, L. J. (1990). *Essentials of psychological testing* (5th ed.). New York, NY: Harper & Row.

Dorans, N. J., & Schmitt, A. P. (1989, March). *The methods for dimensionality assessment and DIF detection*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

Furlow, C. F., Ross, R. T., & Gagne, P. (2009). The impact of multidimensionality on the detection of differential bundle functioning using simultaneous item bias test. *Applied Psychological Measurement*, *33*, 441-464.

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*(2), 26-36.

Jensen, A. R. (1980). *Bias in mental testing*. New York, NY: Macmillan.

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.

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.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). New York, NY: Macmillan.

Monahan, P. O., & Ankenmann, R. D. (2005). Effect of unequal variances in proficiency distributions on type-I error of the Mantel-Haenszel chi-square test for differential item functioning. *Journal of Educational Measurement*, *42*, 101-131.

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

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

Pei, K. L., & Li. J. (2010). Effects of unequal ability variances on the performance of logistic regression, Mantel-Haenszel, SIBTEST IRT, and IRT likelihood ratio for DIF detection. *Applied Psychological Measurement*, *34*, 453-456.

Reckase, M. (2009). *Multidimensional item response theory*. New York, NY: Springer.

Roussos, L. A., & Stout, W. F. (1996a). A multidimensionality-based DIF analysis paradigm. *Applied Psychological Measurement*, *20*, 353-371.

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)

Scheuneman, J. D. (1982). A posteriori analyses of biased items. In R. A. Berk (Ed.), *Handbook of methods for detecting test bias* (pp. 180-198). Baltimore, MD: Johns Hopkins University Press.

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

Shepard, L. A. (1987). Discussant comments on the NCME symposium: Unexpected differential item performance and its assessment among black, Asian American, and Hispanic students. In A. P. Schmitt & N. J. Dorans (Eds.), *Differential item functioning on the Scholastic Aptitude Test* (RM-87-1). Princeton, NJ: Educational Testing Service.

Walker, C. M., & Beretvas, S. N. (2001). An empirical investigation demonstrating the multidimensional DIF paradigm: A cognitive explanation for DIF. *Journal of Educational Measurement*, *38*, 147-163.

Walker, C. M., & Beretvas, S. N. (2003). Comparing multidimensional and unidimensional proficiency classifications: Multidimensional IRT as a diagnostic aid. *Journal of Educational Measurement*, *40*, 255-275.

Walker, C. M., Zhang, B., Banks, K., & Cappaert, K. (2011). Establishing effect size guidelines for interpreting the results of differential bundle functioning analyses using SIBTEST. *Educational and Psychological Measurement*, *72*, 415-434.

Wiley, D. E. (1990). Test validity and invalidity reconsidered. In R. Snow & D. E. Wiley (Eds.), *Improving inquiry in social science* (pp. 75-107). Hillsdale, NJ: Lawrence Erlbaum.

Woods, C. M. (2008). Likelihood-ratio DIF testing: Effects of nonnormality. *Applied Psychological Measurement*, *32*, 511-526.