A Comparison of Unidimensional and Three-Dimensional Differential Item Functioning Analysis Using Two-Dimensional Data

Teresa K. Snow and T.C. Oshima

Educational and Psychological Measurement 2009; 69; 732 originally published online Mar 18, 2009;
DOI: 10.1177/0013164409332223

Educational and Psychological Measurement
Volume 69 Number 5
October 2009 732-747

10.1177/0013164409332223
http://epm.sagepub.com
hosted at
http://online.sagepub.com

A Comparison of Unidimensional and Three-Dimensional Differential Item Functioning Analysis Using Two-Dimensional Data

Teresa K. Snow
Georgia Institute of Technology
T. C. Oshima
Georgia State University

Oshima, Raju, and Flowers demonstrated the use of an item response theory–based technique for analyzing differential item function (DIF) and differential test function for dichotomously scored data that are intended to be multidimensional. Their study assumed that the number of intended-to-be measured dimensions was correctly identified. In practice, however, the number of dimensions may be misidentified. Therefore, the purpose of this study was to demonstrate the effects of both underestimation and overestimation of the number of intended-to-be measured dimensions on the multidimensional DIF analysis using simulated two-dimensional data with known DIF items. Results show that overestimation of the number of θ traits had a consequence of decreased power. Underestimation resulted in missing a certain type of nonuniform DIF, as well as confounding the impact with DIF. Recommendations are made on how to conduct a DIF investigation with a multidimensional within-item test.

Keywords: *differential item function; item response theory; multidimensionality*

Developments in item response theory have provided a framework for test developers to detect differential item functioning (DIF) and differential test functioning (DTF). However, most DIF approaches currently in use are limited by the assumption of unidimensionality.

Many educational and psychological tests are multidimensional by design (for an excellent discussion on this topic, see Ackerman, Gierl, & Walker, 2003). These include achievement tests, aptitude tests such as the LSAT (Camilli, Wang, & Fesq, 1995), and psychological assessments such as the NEO Personality Inventory

Authors' Note: We would like to acknowledge the late Dr. Nam Raju for his help and support in modifying the differential functioning of items and tests to accommodate the multidimensional data used in this study.

(Huang, Church, & Katigbak, 1997). Even employee evaluations and licensure exams are often designed for the purpose of measuring a variety of skills.

As a result, in many instances the assumption of unidimensionality may not hold true. For example, an exam may consist of several subtests or groups of items. In the case that each subtest is measuring a distinctly different latent ability, unidimensional (1D) procedures may be separately performed on each subtest by matching on the latent trait. However, if each item is measuring multiple latent abilities, then 1D procedures are not appropriate and so may give erroneous results. Before conclusions are drawn, these multidimensional within-item tests (Wang, Wilson, & Adams, 1997) should undergo thorough multidimensional analyses, and they should be screened for multidimensional DIF.

Previous studies have shown that violations of the unidimensionality assumption may have a considerable effect on item response theory parameter estimation (Ackerman, 1989; Reckase, 1979). Furthermore, various studies (e.g., Ackerman, 1992) have shown that the presence of multidimensionality may cause DIF. For instance, if the test is intended to be 1D, then the distributional difference of other, nuisance dimensions would result in DIF. Therefore, one needs to distinguish which dimensions are intended to be measured and which are considered to be nuisance dimensions.

Assessing the dimensional structure of data is not trivial. Procedures such as Stout's DIMTEST (1987) have traditionally been used to assess essential unidimensionality to determine if 1D DIF procedures are suitable. However, for multidimensional instruments, it is important to correctly identify and match the examinees on all primary dimensions (Mazor, Hambleton, & Clauser, 1998). McDonald (2000) has suggested the use of a factor-analytic approach to analyze distinct clusters, whereas Zhang and Stout (1999) have shown that conditional covariances can provide useful information regarding dimensional structure.

In 1993, Shealy and Stout introduced SIBTEST (simultaneous item bias test) to match participants on an intended primary dimension and so analyze 1D DIF. In 1997, Stout, Li, Nandakumar, and Bolt. extended the SIBTEST methodology to accommodate two-dimensional (2D) data (MULTISIB). Although simulation studies have suggested that this technique is effective in correctly identifying DIF, it cannot accommodate data with more than two primary dimensions.

In 1995, Raju, van der Linden, and Fleer introduced an item response theory–based method for assessing the differential functioning of items and tests (DFIT). This procedure provided a distinct advantage over other methods—such as the Mantel–Haenszel technique, Lord's chi-square, and Thissen's likelihood ratio test—in that it allowed for the computation of a DTF index, as well as item DIF indices. This can be helpful when evaluating the overall effects of DIF on the test instrument, given that some items may have a cancellation effect. Therefore, even though DIF may be present, the DTF index may be nonsignificant. At the item level, the DIFT framework offers two kinds of indices: compensatory differential

item function (CDIF) and noncompensatory differential item function (NCDIF). CDIF is unique such that the sum of CDIF values equals to DTF. NCDIF, however, is similar to other commonly used DIF indices, and it assumes that all the items (except the studied item) are DIF free. Raju et al. (1995) offer a detailed explanation of these indices; an instructional module for DFIT is also available (Oshima & Morris, 2008).

In 1997, Oshima, Raju, and Flowers extended the 1D model developed by Raju et al. (1995) to apply to the multidimensional case using dichotomous data. Their DIF technique was developed for use with data that are meant to be multidimensional. Using a multidimensional two-parameter logistic model, Oshima and colleagues simulated 2D data with known DTF and DIF. They then used a recovery procedure to examine DTF and DIF indices under various distributional differences of the two intended-to-be measured abilities (θs) by manipulating both the variance–covariance structure and location parameters. After appropriate linking, the multidimensional DFIT procedure did identify DIF correctly in various situations, including when the distributions of intended-to-be measured θs were different for the reference and focal groups. However, their study used the 2D model to simulate the data and the 2D (correct) solution to calibrate the data. In practice, the number of dimensions is not always apparent. Difficulty in identifying the number of dimensions may lead a researcher to identify fewer or more dimensions than what are actually present when calibrating multidimensional data.

Therefore, the purpose of this study was to examine the effects of both underestimating and overestimating the latent trait dimensions on the results obtained from the DIF analyses using simulated 2D data.

DFIT Procedure

If the probability of success on item i for examinee s with ability level θ is $P_i(\theta_s)$, then the examinee's expected proportion correct on a test with n items can be represented by the following equation:

$$Ts = \sum_{i=1}^{n} P_i(\theta_s) \tag{1}$$

If the sample is divided into two groups, with R representing the reference group (usually the majority group) and F representing the focal group (usually the minority group or group of interest), two sets of item parameters will be generated. If the item parameters are truly invariant, then the probability of success on the items should be the same at a given ability level (θ) for both groups, using the estimated item parameters, after parameter estimates have been placed on a common scale. As a result, T_s should be equal, regardless of group membership. To determine

whether this is the case, Raju et al. (1995) suggested generating two scores for each individual based on predicted ability (θ): one as a member of the focal group (T_{sF}) and one as a member of the reference group (T_{sR}). After equating, if $T_{sR} = T_{sF}$, then an examinee's expected proportion correct is independent of group membership. By summing these values across all examinees, DTF can be calculated as follows:

$$\mathrm{DTF} = \varepsilon_F (T_{sR} - T_{sF})^2, \tag{2}$$

where the expectation is taken over the focal group. This equation can also be written as

$$\mathrm{DTF} = \varepsilon_F \left[\left(\sum_{i=1}^{n} d_{is} \right)^2 \right], \tag{3}$$

where d_{is} represents the difference in probabilities of success on item i between the reference and focal groups for a given θ level for examinee s. DTF relates to CDIF as follows (for a detailed overview of DTF, see Oshima et al., 1997):

$$\mathrm{DTF} = \sum_{i=1}^{n} \mathrm{CDIF}_i \tag{4}$$

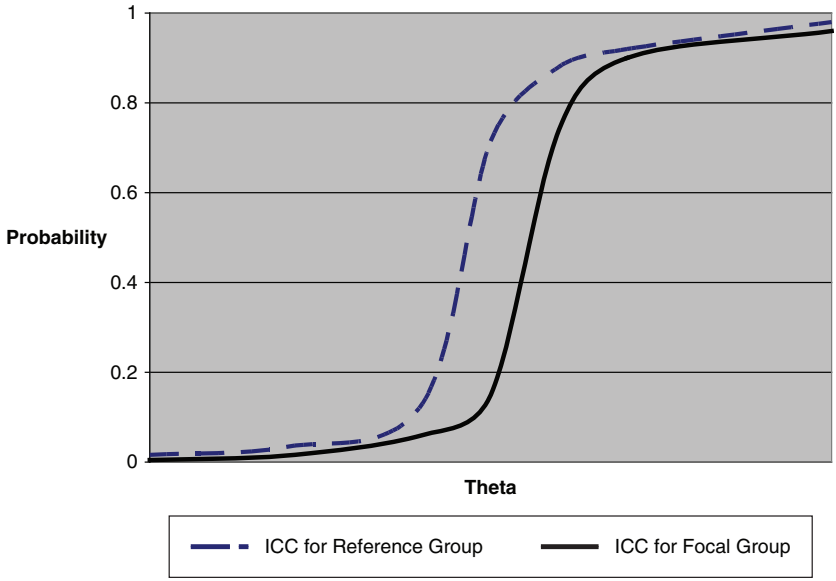
Whereas DTF represents the cumulative effect of CDIF across all items, NCDIF represents a positive value indicating the magnitude of item DIF present, defined by the following equation:

$$\mathrm{NCDIF}_i = \varepsilon_F [P_{iF}(\theta_s) - P_{iR}(\theta_s)]^2 = \varepsilon_F d_i^2(\theta_s). \tag{5}$$

Therefore, NCDIF is the average squared difference of the probability of success on an item between the reference and focal groups, with the expectation taken over the focal group. If NCDIF = 0, then there is no DIF present, and the item response function is identical among all groups (Raju & Ellis, 2000).

Although frequently used to assess DIF and DTF, this model assumes that the test being evaluated is 1D, meaning that only one ability measure or latent trait is accounting for an examinee's performance. In 1997, Oshima et al. extended Raju and colleagues' item response theory–based technique (1995) to apply to multidimensional data. The equations used to calculate DIF and DTF using multidimensional data are, in essence, the same as those described above, which are used to calculate values for 1D data. The primary difference lies in the ability estimation parameters. In the 1D situation, only one latent trait is presumably accounting for an examinee's performance. Therefore, based on the 1D two-parameter model (Equation 6), probability of success for an examinee s can be described by one item discrimination parameter, a, and one parameter related to item difficulty, b:

Figure 1
Unidimensional Item Characteristic Curve (ICC)



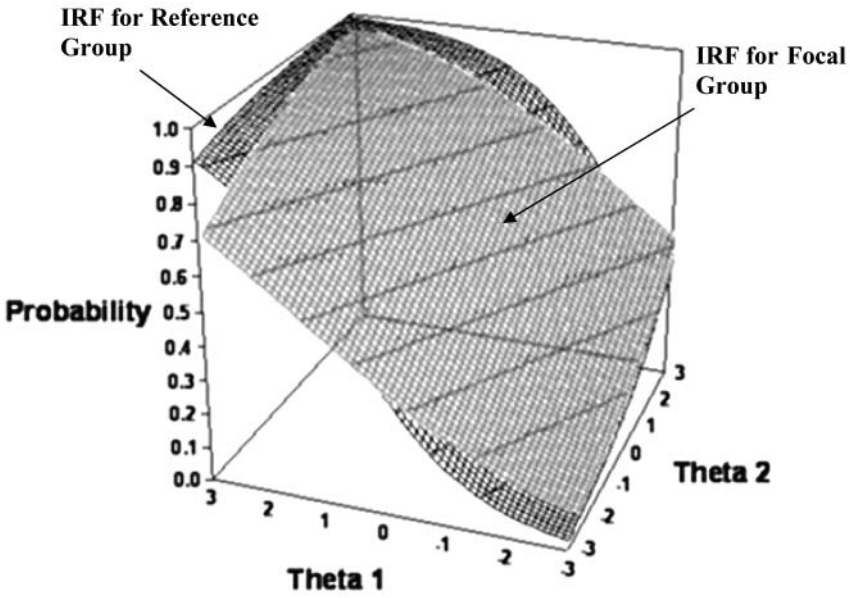
$$P_i(\theta_s) = \frac{1}{1 + e^{-1.7(a_i\theta_s + b_i)}} \tag{6}$$

As shown in Figure 1, the relationship between these item parameters and the ability dimension (θ) can be represented graphically by the item response function, also known as the item characteristic curve for 1D item response theory.

If the data are multidimensional, multiple traits may be accounting for an examinee's performance. In this situation, each dimension can affect the item parameters. If so, the *a* parameter becomes an array of vectors that describe the contributions made by each trait on the ability estimate of interest. Similarly, individual θ estimates are obtained for each dimension detected in the model. For example, if the data are 2D, θ_1 could represent the primary construct of interest, whereas θ_2 might denote a secondary dimension. Figure 2 provides a graphical example of two item response functions in the 2D space.

In this case, a multidimensional extension of the compensatory two-parameter logistic model can be used to describe the relationship between the latent traits. In this model, the *b* parameter is replaced by *d*, a scalar parameter related to item difficulty. The model can be expressed as follows:

Figure 2
Two-Dimensional Item Response Functions (IRFs)



$$P(\boldsymbol{\theta}_s) = \frac{1}{1 + e^{-1.7(\mathbf{a}_i'\boldsymbol{\theta}_s + d_i)}}, \tag{7}$$

where $\mathbf{a}_i$ is an $m \times 1$ vector of item discrimination parameters and θ_s is an $m \times 1$ vector of ability parameters for an examinee, with m representing the number of dimensions. Item DIF can then be calculated for items using Equation 5. As in the case for 1D DFIT, an appropriate linking needs to be performed before the item parameter estimates are compared between the two groups. The multidimensional linking necessarily involves adjusting the variance and covariance differences of ability dimensions, as well as the location differences of the reference and focal groups (for details on multidimensional linking, see Oshima, Davey, & Lee, 2000).

Method

To demonstrate the effect of misidentifying the number of underlying dimensions on DIF measurements, a 2D dichotomous data set with known underlying structure

was used. This study is an extension of a previously published study by Oshima et al. (1997) in which the framework for multidimensional DIF and DTF was originally developed. The data were generated using a multidimensional two-parameter logistic model, using the same item parameters as the original study. It contains item responses for a 40-item test, with a sample size of 1,000 per group (focal and reference). The primary difference was that in this study, we not only applied the correct 2D solution to the data but also under- and overestimated the number of dimensions, by applying 1D and three-dimensional (3D) solutions, to determine the magnitude of effect on the results. Furthermore, although only one replication was used in the original study under various conditions, 100 replications were used in the present study.

As in the previous study, four cases of distributional differences of θs between the reference and focal groups were considered in the present study: Case A consisted of reference and focal groups that had θ_s drawn from a bivariate normal distribution with zero means, unit variances, and no correlation between θs ($\rho = 0$). In Case B, ρ was changed to .50. Case C was the same as Case B except that a location difference of .50 was added on θ_2. This is an impact condition because θ_2 is an intended-to-be-measured trait. The distributional difference on the intended-to-be-measured traits should be distinguished from DIF, which specifically relates to the distributional difference on the not-intended-to-be-measured traits. Finally, Case D consisted of a correlational difference between the reference and focal groups. The reference group had $\rho = .50$ and the focal group had $\rho = .00$.

As described in the original study, successful multidimensional linking should take care of the distributional differences. Oshima and colleagues (1997) demonstrated that linking was successful in this respect (although the authors recommended interpreting Case D with caution). The same linking method was used in the present study for consistency. We do not, however, know the effectiveness of the multidimensional linking when the dimensional structure is incorrectly identified.

The item parameters used were presented by Oshima et al. (1997) for 2D data designed to measure θ_1 and θ_2 throughout the test (see Table 1). Item directions define the degree to which each item measures θ_1 and θ_2; degrees of 0, 30, 45, and 90 were systematically embedded throughout the test, thereby creating a multidimensional within-item test.

The number of DIF items and the magnitude of DIF were selected to coincide with the original study as well. However, Oshima et al. (1997) had four conditions representing two types of DIF (uniform or nonuniform) under two types of configurations (unidirectional or balanced–directional). In the unidirectional condition, all DIF items favored the reference group, whereas in the balanced–directional configuration, half the DIF items favored the focal group and the remaining items favored the reference group. For purposes of this study, only the unidirectional configuration was investigated, because NCDIF was of primary interest, rather than CDIF, which allows DIF cancellation at the test level.

Table 1
True Item Parameters Before Differential Item Function Was Embedded

Item	α	a_1	a_2	b
1	0	1.62	0.00	1.65
2	30	0.91	0.52	1.59
3	45	1.01	1.01	–1.60
4	60	0.42	0.73	0.25
5	90	0.00	0.69	0.25
6	0	1.08	0.00	0.14
7	30	1.02	0.59	0.18
8	45	1.19	1.19	–0.47
9	60	0.20	0.35	0.37
10	90	0.00	0.85	0.01
11	0	1.72	0.00	–1.75
12	30	0.88	0.51	0.15
13	45	0.53	0.53	0.39
14	60	0.27	0.47	–0.34
15	90	0.00	0.94	0.13
16	0	1.54	0.00	–1.75
17	30	0.77	0.44	–0.40
18	45	0.89	0.89	–0.03
19	60	0.32	0.56	0.43
20	90	0.00	0.77	0.81
21	0	3.52	0.00	–0.69
22	30	0.63	0.36	0.39
23	45	0.53	0.53	0.04
24	60	0.46	0.79	–1.74
25	90	0.00	0.68	–0.94
26	0	0.39	0.00	0.37
27	30	0.35	0.20	0.36
28	45	0.75	0.75	0.51
29	60	0.76	1.32	–1.64
30	90	0.00	0.80	–0.19
31	0	1.52	0.00	0.17
32	30	1.22	0.71	0.31
33	45	0.55	0.55	0.94
34	60	0.50	0.86	–0.91
35	90	0.00	3.23	0.11
36	0	1.09	0.00	–0.23
37	30	1.26	0.73	1.49
38	45	0.44	0.44	–0.24
39	60	0.31	0.53	0.44
40	90	1.11	1.81	2.31
M		0.69	0.63	0.02
SD		0.73	0.58	0.92

Table 2
Item Parameters for Generation of Differential Item Function (DIF) Items

		Reference				Focal			
DIF	Item	α	a_1	a_2	d	α	a_1	a_2	d
Uniform	37	0	1.13	0.00	0	0	1.13	0.00	–0.5
	38	30	0.98	0.57	0	30	0.98	0.57	–0.5
	39	60	0.57	0.98	0	60	0.57	0.98	–0.5
	40	90	0.00	1.13	0	90	0.00	1.13	–0.5
Nonuniform	37	45	0.80	0.80	0	30	0.50	0.80	0.0
	38	45	0.80	0.80	0	60	0.50	0.50	0.0
	39	45	0.80	0.80	0	69	0.50	1.30	0.0
	40	45	0.80	0.80	0	45	0.50	0.50	–0.5

All DIF analyses were performed under a uniform DIF condition (variation using only the *d* parameter) and a nonuniform DIF condition (variation using the *a* parameters with or without the *d* variation). In each condition, the last four items in the 40-item test were embedded with DIF. The remaining items (Items 1–36) were non-DIF items. In other words, the item parameters were identical for the reference and focal groups. Table 2 presents a list of the parameters used for DIF items (which are identical to those used for Condition 1 and Condition 3 in the original 1997 study by Oshima et al.). The data were simulated using SAS 9.13 (SAS Institute, Cary, NC).

In both the 1D and multidimensional solutions, item parameters were calibrated using NOHARM (Frasier, 1988). Following calibration, IPLINK software (Lee & Oshima, 1996) was used to transform the item parameters from the reference group to the underlying metric of the item parameters of the focal group. Items were linked using the direct method, which uses a linear transformation to minimize the sum of squared difference between corresponding item parameters. Although the original study performed two-stage linking to select non-DIF items (as linking items), this study presumed that purification of the linking items was successful. Therefore, Items 1 through 36 were used as linking items.

The DFIT program (Raju, 1997) was used for all DIF analyses. In recent years, simulation-based methods described by Oshima, Raju, and Nanda (2006) have been recommended for determining the cutoff score used to identify significant NCDIFF. However, the item parameter replication method, which was incorporated in the 1D DFIT program (DFIT8; Raju, Oshima, & Wolach, 2009), is not yet available for the multidimensional DFIT program. Therefore, in this study, the cutoff values for 1D, 2D, and 3D solutions were determined by obtaining the distribution of non-DIF items. For each solution, the 95th percentile rank score out of 28,800 non-DIF items (Items 1–36) served as the cutoff value (36 items × 100 replications × 4 cases × 2 conditions). This criterion was helpful as a tool to identify the relative rate of false positives and true positives for various situations

Table 3
Noncompensatory Differential Item Function (NCDIF) Estimates for Items 37–40

	Uniform DIF				Nonuniform DIF			
	True	1D	2D	3D	True	1D	2D	3D
Case A								
Item 37	.010	.009	.013	.014	.003	.002	.004	.005
Item 38	.010	.011	.012	.012	.005	.005	.006	.007
Item 39	.010	.010	.012	.012	.011	.000	.014	.015
Item 40	.010	.008	.011	.013	.018	.018	.019	.020
Case B								
Item 37	.010	.010	.012	.013	.003	.002	.003	.004
Item 38	.010	.010	.011	.012	.005	.006	.007	.008
Item 39	.010	.010	.011	.012	.011	.000	.007	.007
Item 40	.010	.009	.012	.013	.018	.019	.019	.020
Case C								
Item 37	.010	.002	.010	.011	.003	.002	.004	.005
Item 38	.010	.007	.009	.010	.005	.006	.007	.008
Item 39	.010	.014	.009	.010	.011	.002	.008	.009
Item 40	.010	.021	.011	.012	.018	.017	.016	.016
Case D								
Item 37	.010	.009	.014	.013	.003	.001	.004	.005
Item 38	.010	.011	.012	.012	.005	.004	.004	.005
Item 39	.010	.010	.011	.013	.011	.001	.016	.017
Item 40	.010	.008	.012	.014	.018	.018	.019	.020

within each solution. The cutoff values were .002, .004, and .007 for the 1D, 2D, and 3D solutions, respectively. Items displaying NCDIF values greater than the cutoff score were considered to show significant DIF at the .05 level. Following the original study, simulated θs (from a bivariate standard normal distribution) were used as the θs from the focal group in calculating NCDIF, given that NOHARM did not provide these estimates.

All analyses were performed assuming a 1D, 2D, or 3D structure. Because the data in this study were known to be 2D, the 2D model served as a reference for comparison. Therefore, results obtained from the 1D and 3D models were compared to this model.

Results

Table 3 displays a summary of the NCDIF results. The true DIF value is listed next to each item, along with the average estimated item DIF for each solution over 100 replications. The true DIF value was calculated using the item parameters listed in Tables 1 and 2, in conjunction with the simulated θs from a bivariate

Table 4
Differential Item Functioning (DIF) Detection Rates (Expressed as a Percentage) for True DIF Items

	Uniform DIF			Nonuniform DIF		
	1D	2D	3D	1D	2D	3D
Case A						
Item 37	100	97	94	17	42	17
Item 38	100	100	96	83	63	35
Item 39	100	99	94	2	99	93
Item 40	100	98	87	100	100	99
Case B						
Item 37	100	98	88	24	23	6
Item 38	100	98	82	98	80	53
Item 39	100	100	90	0	79	46
Item 40	100	99	89	100	100	100
Case C						
Item 37	40	95	72	41	24	10
Item 38	100	96	76	91	80	52
Item 39	100	98	76	46	75	63
Item 40	100	98	91	100	100	99
Case D						
Item 37	100	100	78	8	37	12
Item 38	99	99	90	74	41	17
Item 39	100	99	91	0	100	93
Item 40	100	100	93	100	100	99

standard normal distribution. In general, NCDIF values are slightly underestimated for the 1D solution and slightly overestimated for the 3D solution. Nevertheless, the estimated NCDIF values were fairly close to the true NCDIF values for all conditions, except a few. The most notable exception is Item 39 for the 1D solution in the nonuniform DIF condition. Recall that this is the nonuniform DIF condition where the a_1 parameter was lower for the focal group (compared to the reference group) and the a_2 parameter was higher for the focal group (compared to the reference group). Perhaps when only one a parameter is estimated in the 1D solution, this type of nonuniform DIF (hereafter, bidirectional nonuniform DIF) cancels out and does not show as DIF.

Table 4 presents detection rates using the empirical cutoffs for DIF items for each solution. For uniform DIF, the true NCDIF values were large (.010; see Table 2) for all DIF items (Items 37 and 38). These values also greatly exceeded the cutoff value for each solution. Therefore, high detection rates were expected for all conditions. Our results show that detection rates were high in the 1D and 2D solutions (with the exception of one condition in the 1D solution), thereby suggesting

Table 5
False-Positive Rates (Expressed as a Percentage) for Non–Differential Item Functioning Items

	Uniform			Nonuniform		
	1D	2D	3D	1D	2D	3D
Case A	0.36	5.58	3.22	0.61	6.83	3.72
Case B	0.31	2.33	4.97	0.31	2.00	5.11
Case C	21.25	2.69	4.94	24.83	3.14	6.28
Case D	2.92	9.03	7.03	2.61	15.56	8.58

good performance. However, detection rates overall were lower for the 3D solution, thus suggesting questionable performance. Perhaps when the number of dimensions is overestimated, the distributional difference on the nuisance traits, which should show up as DIF, may be mistaken as impact; therefore, it does not correctly identify DIF in some cases.

For nonuniform DIF, the true NCDIF values varied in magnitude. Items 37, 38, 39, and 40 had true NCDIF values of .003, .005, .011, and .018, respectively. As a result, we would expect Items 39 and 40 to be identified as DIF because they both exceed the cutoff values. However, whether Items 37 and 38 should be detected as DIF depends on one's definition of how large DIF must be in order to be meaningful. The true NCDIF values for Items 37 and 38 hovered around the cutoff values. Therefore, lower detection rates should not necessarily be viewed as poor performance for those items. Based on the results of the 2D solution reported in Table 3 (used as the benchmark of good performance), the 3D solution once again exhibited lower detection rates. The pattern of the detection rates for various types of nonuniform DIF for the 3D solution was similar to that of the 2D solution. For the 1D solution, the detection rate for Item 39 is remarkably low, again indicating that the 1D solution could miss bidirectional nonuniform DIF. In addition, Item 37 for the 1D solution had occasionally somewhat lower detection rates, compared to those of the 2D solution, thereby suggesting that the 1D solution may have a problem with nonuniform DIF detection when only one of the a parameters is different between the reference and focal groups. In contrast, when both of the a parameters differed uniformly (Item 38), the 1D solution consistently showed detection rates higher than those for the 2D solution.

Table 5 presents false-positive rates for non-DIF items (Items 1–36). Each rate is a percentage expressed as the total number of false-positive items, identified out of a possible 3,600 (36 items $\times$ 100 replications). A false-positive rate of 5% is expected given that the level of significance was .05. For the 2D and 3D solutions, the false-positive rates were in the reasonable range, except for Case D. For the 1D solution, though, the false-positive rates were either extremely small (Cases A and

B) or extremely large (Case C). Case C is the only situation with impact, meaning that there was a distributional difference for the secondary trait in the 2D test. One can reasonably speculate that the impact on the secondary dimension shows up as DIF when only one trait is accounted for (the 1D solution) given that DIF is conceptualized as the distributional difference on the nuisance trait. By comparison, in the 2D solution where two traits are accounted for, the distributional difference of the secondary trait (impact, not DIF by definition) did not cause excessive false positives. Note that the 3D solution handled this situation fairly well. For all solutions, the most problematic seemed to be Case D (correlational difference between the reference and focal groups). Even with the 2D solution, Case D produced excessive false-positive rates.

Discussion

The present study investigated DIF detection and the effect of underestimating and overestimating the number of dimensions. Specifically, we investigated the consequences of applying 1D DFIT and 3D DFIT to 2D data. In practice, identifying the correct number of dimensions is not always straightforward (for an example of real data analysis, see Stone & Yeh, 2006). Therefore, it is important to investigate the effect of misspecification of latent space on DIF techniques. As discussed earlier, we looked only at the multidimensional within-item test where two traits are measured throughout the test. It is this type of test for which multidimensional DIF techniques can prove to be most useful.

The results of our study suggest that underestimation and overestimation can be a problem but in different ways. When the 3D solution was applied to the 2D data, the overall power decreased, with a slight tendency for inflated false-positive rates in some conditions. When the 1D solution was applied to the 2D data, overall detection rates and false-positive rates were good in most conditions (and sometimes even better than those for the 2D solution). However, there were a few exceptions, and in those conditions the negative consequences were quite severe. One of the consequences was that a certain type of nonuniform DIF was totally missed—namely, bidirectional nonuniform DIF, where the a_1 parameter was higher for the reference group whereas the a_2 parameter was higher for the focal group, thus creating a similar overall discrimination. Another consequence was that the impact related to one of the intended-to-be-measured dimensions can be mistaken as DIF when the number of dimensions is underestimated. The latter can be a more serious problem in practice, because a good multidimensional item may be thrown away as a DIF item.

With respect to false positives, the rates were highest in all solutions for Case D, which reflected a correlation difference between the reference and focal groups. Although the detection rate did not seem to suffer in Case D, the much higher false-positive rates observed in this study, in all solutions (1D, 2D, and 3D),

deserve further investigation. The correlational difference between two groups suggests a difference in the dimensionality structure. In an extreme case where one group has a correlation coefficient of 1.00, the 2D structure is reduced to the 1D structure. The results of this study indicate that the difference in dimensionality structure between two groups can pose a serious threat to multidimensional DIF analyses. Furthermore, the current multidimensional DIF technique assumes comparable dimensionality structures between the two groups. Therefore, a thorough dimensionality analysis is recommended before the application of the multidimensional DIF technique.

In practice, several software programs are currently available to determine the dimensionality structure of the dichotomous test. Among them are DIMTEST (Stout et al., 1997), Mplus (Muthen & Muthen, 2001), TESTFACT (Wilson et al., 2003), and NOHARM (Frasier, 1988). DIMTEST can be used only to test if the data are essentially 1D. The other three can be used to investigate the number of dimensions. Methods for determining the number of dimensions or factors include examining eigenvalues as well as residual and fit statistics. To illustrate the process, two of our data sets (one with no correlation between θs, the other with a correlation of .50) were chosen to undergo dimensionality testing. Both were rejected for essential unidimensionality using DIMTEST. However, an evaluation of the root mean square residual statistic by NOHARM suggested that all solutions (1D, 2D, and 3D) satisfied that criterion for use. Factor analysis performed with TESTFACT produced eigenvalues of 15.19, 5.13, 1.18, 1.13, and 1.06 for the first data set (no correlation) and 19.23, 2.99, 1.12, 1.02, and 0.96 for the second data set (.50 correlation), both suggesting a 2D structure.

In our study, the data were clearly simulated as 2D. Even so, the above dimensionality investigation does not provide the definitive answer to the number of dimensions; in practice, the task is even more difficult (for detailed information on Mplus, TESTFACT, and NOHARM, see Stone & Yeh, 2006).

Multidimensional DFIT is still under development. For 1D DFIT, a new significance test was developed using the item parameter replication method (Oshima et al., 2006). To date, there is no item parameter replication method for multidimensional DFIT. The results from our study, as related to the power and false positives, need to be interpreted with caution because those rates can change, depending on the choice of the cutoff scores. Until the item parameter replication method is incorporated into the multidimensional DFIT program, DFIT users will have to come up with their own cutoff scores, by either simulating data or perhaps using the cutoff scores presented in this study as a temporary rule of thumb for the 1D, 2D, and 3D solutions.

In summary, under the conditions investigated, we found that using the 1D or 3D solutions for the 2D data presented some problems but in different ways. However, we did confirm that applying the 2D DFIT procedure to the 2D data—that is, correctly identified dimensions are used for the multidimensional DIF analysis—was a promising DIF detection technique for the multidimensional within-item test,

provided that the correlations of traits are similar between the two groups. Therefore, as this study suggests, an essential first step in the DIF analysis for intentionally multidimensional tests is to correctly identify the number of intended-to-be-measured dimensions. Likewise, equivalence of the dimensionality structures between two groups of interest must be established before applying a multidimensional DIF technique. Further study is needed in the areas of detection and the evaluation of equivalence of multidimensional data structures.

References

Ackerman, T., Gierl, M., & Walker, C. M. (2003). Using multidimensional item response theory to evaluate educational and psychological tests. *Educational Measurement: Issues and Practice*, *22*(3), 37-53.

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.

Camilli, G., Wang, M.-M., & Fesq, J. (1995). The effects of dimensionality on equating the law school admission test. *Journal of Educational Measurement, 32*, 79-96.

Frasier, C. (1988). NOHARM [Computer program]. New South Wales, Australia: University of New England, Center for Behavioral Studies.

Huang, C. D., Church, A. T., & Katigbak, M. S. (1997). Identifying cultural differences in items and traits—Differential item functioning in the NEO Personality Inventory. *Journal of Cross-Cultural Psychology, 28*, 192-218.

Lee, K., & Oshima, T. C. (1996). IPLINK: Multidimensional and unidimensional item parameter linking in item response theory. *Applied Psychological Measurement*, *20*, 230.

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

McDonald, R. P. (2000). A basis for multidimensional item response theory. *Applied Psychological Measurement*, *24*, 99-114.

Muthén, L.K. & Muthén, B. (2001). Mplus (version 2.0) [Computer software]. Los Angeles, CA: Muthén & Muthén.

Oshima, T. C., Davey, T. C., & Lee, K. (2000). Multidimensional linking: Four practical approaches. *Journal of Educational Measurement, 37*, 357-373.

Oshima, T. C., & Morris, S. B. (2008). An NCME instructional module on Raju's differential functioning of items and tests (DFIT). *Educational Measurement: Issues and Practice*, *27*, 43-50.

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

Oshima, T. C., Raju, N. S., & Nanda, A. (2006). A new method for assessing statistical significance in the differential functioning of items and tests (DFIT) framework. *Journal of Educational Measurement*, *43*, 1-17.

Raju, N. S. (1997). DFIT [Computer program]. Atlanta: Georgia Institute of Technology.

Raju, N. S., & Ellis, B. B. (2000). Differential item functioning. In F. Drasgow & N. Schmitt (Eds.), *Measuring and analyzing behavior in organizations: Advances in measurement and data analysis* (pp. 156-188). San Francisco: Jossey-Bass.

Raju, N. S., Oshima, T.C., & Walach, A.H. (2009). DFIT8 [Computer software]. St. Paul, MN: Assessment Systems Corporation.

Raju, N. S., van der Linden, W. J., & Fleer, P. F. (1995). An IRT-based internal measure of test bias with applications for differential item functioning. *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.

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.

Stone, C., & Yeh, C. C. (2006). Assessing the dimensionality and factor structure of multiple-choice exams: An empirical comparison of methods using the multistate bar examination. *Educational and Psychological Measurement*, *66*, 193-214.

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

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

Wang, W.-C., Wilson, M. R., & Adams, R. J. (1997). Rasch models for multidimensionality between items and within items. In M. Wilson, G. Engelhard, & K. Draney (Eds.), *Objective measurement: Vol. 4. Theory into practice* (pp. 139-155). Norwood, NJ: Ablex.

Wilson, D., Wood, R., Schilling, S., & Gibbons, R. (2003). TESTFACT (Version 4.0) [Computer software]. Lincolnwood, IL: Scientific Software International.

Zhang, J. & Stout, W. (1999). The theoretical DETECT index of dimensionality and its application to approximate simple structure. *Psychometrika, 64*, 213-249.