Journal of Educational Measurement
Fall 1997, Vol. 34, No. 3, pp. 253–272

Development and Demonstration of Multidimensional IRT-Based Internal Measures of Differential Functioning of Items and Tests

T. C. Oshima
Georgia State University
Nambury S. Raju
Illinois Institute of Technology
Claudia P. Flowers
University of North Carolina at Charlotte

This article defines and demonstrates a framework for studying differential item functioning (DIF) and differential test functioning (DTF) for tests that are intended to be multidimensional. The procedure introduced here is an extension of unidimensional differential functioning of items and tests (DFIT) recently developed by Raju, van der Linden, & Fleer (1995). To demonstrate the usefulness of these new indexes in a multidimensional IRT setting, two-dimensional data were simulated with known item parameters and known DIF and DTF. The DIF and DTF indexes were recovered reasonably well under various distributional differences of θs after multidimensional linking was applied to put the two sets of item parameters on a common scale. Further studies are suggested in the area of DIF/DTF for intentionally multidimensional tests.

Although most currently used models in item response theory (IRT) are based on the unidimensionality assumption, many researchers agree that educational and psychological test data do not always satisfy the unidimensionality assumption (e.g., Ackerman, 1991; Traub, 1983). For example, a test such as a licensure exam may measure several subsets of skills. Another type of multidimensional test may consist of items that require the composite of two or more intentionally defined abilities. For example, a math test may measure a composite skill of mathematics and reading throughout the test, with each item having a different emphasis of the two skills. According to Wang, Wilson, and Adams (1995), the first type of multidimensional test is called the multidimensional between-item test, and the latter type is called the multidimensional within-item test.

Differential item functioning (DIF) or differential test functioning (DTF) for intentionally multidimensional tests should be clearly distinguished from DIF defined for unintentionally multidimensional tests. DIF for intentionally multidimensional tests is defined as the distributional differences on the *additional* trait(s). If

Portions of this article were presented at the 1993 Annual Meeting of the American Educational Research Association, Atlanta, GA. The authors would like to express their appreciation to the editor and two anonymous reviewers for their many valuable comments on a previous version of the manuscript. Requests for information should be sent to either of the first two authors.

the intentionally measured dimensions are defined in the context of a multidimensional IRT model, distributional differences on any of the intentionally measured dimensions should not indicate DIF.

Recently Raju, van der Linden, and Fleer (1992, 1995) introduced an IRT-based framework for assessing differential functioning of items and tests (DFIT). This DFIT framework offers a general procedure for assessing DIF/DTF in tests developed with unidimensional, multidimensional, or polytomous models. The DFIT indexes fall into the class of parametric IRT-based DIF/DTF indexes. Other indexes in this class are Lord's (1980) chi-square, area measures (Raju, 1988; Rudner 1977), and the likelihood ratio test (Thissen, Steinberg, & Wainer, 1988). At the present time, these other procedures cannot handle multidimensional IRT models. Furthermore, these other procedures do not assess DTF.

Raju et al. (1995) offered a detailed description of DFIT only for the unidimensional case. The purpose of this research is to provide an extended description of this new technique for the multidimensional case and to offer a demonstration of this technique using simulated intentionally two-dimensional data with known DIF and DTF.

DFIT for Multidimensional Tests

Differential Test Functioning

According to a multidimensional extension of the two-parameter logistic (M2PL) model (Reckase, 1985; Reckase & McKinley, 1991), the probability of success on item i for an examinee can be written as

$$P_i(\boldsymbol{\theta}) = \frac{1}{1 + e^{-1.7(\mathbf{a}_i' \boldsymbol{\theta} + b_i)}}, \tag{1}$$

where $\mathbf{a}_i$ is an $m \times 1$ vector of item discrimination parameters, b_i (commonly known as d_i)[1] is a scalar parameter related to the difficulty of the item, $\boldsymbol{\theta}$ is an $m \times 1$ vector of ability parameters for the examinee, and m is the number of ability dimensions. Let the test consist of k items and have one set of item parameters for each of two groups (reference group and focal group). Let us also assume that the two sets of item parameters are on a common scale. Now, let $P_{i\text{R}}(\boldsymbol{\theta})$ represent the probability of success on item i for an examinee as if he or she were a member of the reference group; similarly, let $P_{i\text{F}}(\boldsymbol{\theta})$ represent the probability of success for the same examinee on the same item as if he or she were a member of the focal group. In computing $P_{i\text{R}}(\boldsymbol{\theta})$, item parameters ($\mathbf{a}_i$ and b_i) based on the reference group are used in Equation 1. Similarly, for $P_{i\text{F}}(\boldsymbol{\theta})$, item parameters based on the focal group are used. The same vector of $\boldsymbol{\theta}$, however, is used in both computations. If an item were functioning differently in the two groups, then $P_{i\text{R}}(\boldsymbol{\theta})$ and $P_{i\text{F}}(\boldsymbol{\theta})$ would be different for a given examinee.

An examinee's true score, within the IRT context, can be expressed as

$$T = \sum_{i=1}^{k} P_i(\boldsymbol{\theta}). \tag{2}$$

In the present setup, each examinee will have two true scores, one for being a member of the focal group (T_F) and the other for being a member of the reference group (T_R). If T_R and T_F are equal for an examinee, then the examinee's true score is independent of group membership. The greater the difference between T_R and T_F, the greater the DTF. A DTF measure at the examinee level may be defined as $(T_F - T_R)^2$. Letting $D = T_F - T_R$, an overall measure of DTF across examinees may be defined as

$$DTF = E_F(T_F - T_R)^2 = E_F D^2 = \sigma_D^2 + (\mu_{TF} - \mu_{TR})^2 = \sigma_D^2 + \mu_D^2 , \qquad (3)$$

where the expectation (E) can be taken over the reference group or focal group and μ and σ refer to the mean and standard deviation, respectively. We will assume that the expectation is taken over the focal group. The *DTF* given in Equation 3 was used by Stocking and Lord (1983) in the context of scale transformation.

Differential Item Functioning

CDIF. Equation 3 can be rewritten as

$$DTF = E_F(D^2) = E_F\,[\sum_{i=1}^{k} (d_iD)\,] = \sum_{i=1}^{k} E_F(d_iD) = \sum_{i=1}^{k} [\mathrm{Cov}(d_i, D) + \mu_{d_i}\mu_D] , \qquad (4)$$

where $d_i = P_{iF}(\boldsymbol{\theta}) - P_{iR}(\boldsymbol{\theta}), \sum_{i=1}^{k} d_i = D = T_F - T_R$, and $\mathrm{Cov}(d_i, D)$ is the covariance between the difference in item probabilities for item i (d_i) and the difference between the two true scores (D). The expectation is again taken over the focal group. One definition of differential functioning at the item level may be expressed as

$$CDIF_i = E_F(d_iD) = \mathrm{Cov}(d_i, D) + \mu_{d_i}\mu_D \,. \qquad (5)$$

The notation *CDIF* stands for compensatory DIF, and it will be distinguished from noncompensatory DIF (NCDIF), to be defined later. Combining Equations 4 and 5, we obtain

$$DTF = \sum_{i=1}^{k} CDIF_i \,. \qquad (6)$$

Equation 6 shows that the definition of $CDIF_i$ is additive in the sense that differential functioning at the test level is simply the sum of compensatory differential functioning at the item level. Therefore, a positive *CDIF* for one item may partially or fully cancel a negative *CDIF* for another item in terms of their contribution to *DTF*. Furthermore, the covariance term in Equation 5 reflects the correlated DIF between items; that is, since for item i,

$$\mathrm{Cov}(d_i, D) = \sigma_{d_i}^2 + \sum \mathrm{Cov}(d_i, d_j), i \neq j,$$

CDIF for item i includes correlated DIF between item i and any other item in the test.

Rewriting Equation 3, one obtains

$$DTF = E_F[\sum_{i=1}^{k}(P_{iF} - P_{iR})]^2 = E_F[(P_{1F} - P_{1R}) + (P_{2F} - P_{2R}) + \cdots + (P_{kF} - P_{kR})]^2 . \qquad (7)$$

Equation 7 shows the compensating nature of the proposed *DTF*. For example, if $P_{6F} - P_{6R} = -.2$ and $P_{7F} - P_{7R} = +.2$ for a given examinee, then the DIF in Item 6 cancels out with the DIF in Item 7, and the two items together contribute zero to the examinee's *DTF* score. The proposed *DTF*, therefore, takes into account compensating DIF across items at the examinee level. In addition, Equation 6 shows the nature of compensating DIF across items at the group level. From a practitioner's point of view, this is a useful feature because it enables the practitioner not only to assess which items have compensating DIF or which items to delete due to DIF, but also to estimate the net effect of such an action on DTF. When items with significant CDIF are deleted from the final test, the revised *DTF* can be computed for the retained items using Equation 3.

Several researchers have suggested investigating differential functioning beyond the item level using procedures such as SIBTEST (Shealy & Stout, 1993) and the random effects model for Mantel-Haenszel differential item functioning (Longford, 1995; Longford, Holland, & Thayer, 1993). Within the DFIT framework, one starts with a definition of DTF and then decomposes DTF into differential functioning at the item level (CDIF). Therefore, it is not surprising that the definition of *CDIF* for a given item, shown in Equation 5 (especially the covariance term), includes information about correlated DIF between the item in question and other items in the test. In practice, it is possible that two items with significant DIF may be quite similar, because the stems for the two items are very similarly phrased or because the two items tap very similar content. In such cases, DIF in the two items may have a nonzero correlation, which, in turn, will influence differential functioning at the test level.

NCDIF. The purpose of this section is to define a noncompensatory DIF (NCDIF). If we assume that all items in the test, other than item i, are completely free of DIF, then it must be true that $d_j = 0$ for all $j \neq i$. Then, Equation 5 can be rewritten as

$$NCDIF_i = \sigma^2_{d_i} + \mu^2_{d_i}, \tag{8}$$

which does not include information about DIF from other items in the test. In the unidimensional case, Raju et al. (1995) showed how this definition of NCDIF relates to Lord's (1980) chi-square and the exact unsigned area measure (Raju, 1988). The CDIF and NCDIF terminology is introduced here for the explicit purpose of distinguishing between compensatory DIF and noncompensatory DIF. Among other things, we hope that this distinction clearly articulates the fact that many of the currently popular DIF indexes implicitly assume that items other than the one item under consideration are DIF free.

Significance Tests for DTF and NCDIF

Prior to describing some statistical tests for assessing the significance of DTF, CDIF, and NCDIF, it should be noted that, up to this point, definitions of these terms have been expressed in terms of person parameters ($\boldsymbol{\theta}$) and item parameters ($\mathbf{a}$ and b). However, only estimates of $\boldsymbol{\theta}$, $\mathbf{a}$, and b (denoted as $\hat{\boldsymbol{\theta}}$, $\hat{\mathbf{a}}$, and $\hat{b}$, respectively) are typically available for the focal and reference groups. Therefore, the proposed DTF, CDIF, and NCDIF indexes are to be computed using estimated person and item parameters in practice. Estimates of *DTF*, *CDIF*, and *NCDIF* (denoted, respectively, as $D\hat{T}F$, $CD\hat{I}F$, and $NC\hat{D}IF$) indexes will be computed with the help of $\hat{D}$ and $\hat{d}_i$, which are estimates of D and d_i, respectively, for an examinee. That is,

$$\hat{D} = \sum_{i=1}^{k} \hat{d}_i \,, \tag{9}$$

$$\hat{d}_i = \hat{P}_{iF} - \hat{P}_{iR} \,, \tag{10}$$

$$D\hat{T}F = \text{Mean of } \hat{D}^2 = \hat{\sigma}^2_{\hat{D}} + \hat{\mu}^2_{\hat{D}} \,, \tag{11}$$

$$CD\hat{I}F_i = \hat{\text{Cov}}(\hat{d}_i, \hat{D}) + \hat{\mu}_{\hat{d}_i}\hat{\mu}_{\hat{D}} \,, \tag{12}$$

$$NC\hat{D}IF_i = \hat{\sigma}^2_{\hat{d}_i} + \hat{\mu}^2_{\hat{d}_i} \,, \tag{13}$$

where $\hat{P}_{iF}$ and $\hat{P}_{iR}$ are item probabilities, computed with estimated person parameters ($\hat{\boldsymbol{\theta}}$) and estimated item parameters ($\hat{\mathbf{a}}$ and $\hat{\mathbf{d}}$). The symbols $\hat{\sigma}^2$, $\hat{\mu}$, and $\hat{\text{Cov}}$ represent estimates of σ^2, μ, and Cov, respectively.

According to the above definitions, estimates of *DTF*, *CDIF*, and *NCDIF* have three kinds of errors: (a) estimation error resulting from the use of $\hat{\boldsymbol{\theta}}$, $\hat{\mathbf{a}}$, and $\hat{d}$ in place of $\boldsymbol{\theta}$, $\mathbf{a}$, and b, respectively; (b) equating/linking error; and (c) the typical sampling error resulting from the use of only a sample from a population of examinees. We hope that future research will be successful in proposing significance tests that fully account for the errors associated with the estimation and linking of person and item parameters. It should be noted, however, that while the use of significance tests that do not directly account for the estimation and linking errors is not ideal, it is a common practice in the unidimensional case to compute standard errors for the person and item parameters, without reflecting the estimation and linking errors, and then use them in DIF analysis (Lord, 1980; Raju, 1990).

Chi-square test for D$\hat{\text{T}}$F. The chi-square test to be described below assumes that $\hat{D}$ is normally distributed with a mean of $\mu_{\hat{D}}$ and a finite standard deviation of $\sigma_{\hat{D}}$. This assumption certainly needs to be verified in future investigations. The z-score for examinee s can be written as

$$z_s = \frac{\hat{D}_s - \mu_{\hat{D}}}{\sigma_{\hat{D}}} \,. \tag{14}$$

Since z_s^2 has a chi-square distribution with one degree of freedom (provided z_s is standard normal), the sum of z_s^2 across N_F examinees in the focal group has a chi-square distribution with N_F degrees of freedom. Algebraically, this can be expressed as

$$\chi^2_{N_F} = \sum_{s=1}^{N_F} z_s^2 = \frac{\sum_{s=1}^{N_F} (\hat{D}_s - \mu_{\hat{D}})^2}{\sigma^2_{\hat{D}}} \,. \tag{15}$$

In the present context, the null hypothesis is

$$E(D\hat{T}F) = \mu_{\hat{D}^2} = 0 , \tag{16}$$

which implies that $\mu_{\hat{D}}$ must also be zero. It should be noted that $\mu_{\hat{D}} = 0$ is a necessary but not sufficient condition for the validity of Equation 16. Substituting $\mu_{\hat{D}} = 0$ into Equation 15 yields

$$\chi^2_{N_F} = \frac{\sum_{s=1}^{N_F} \hat{D}_s^2}{\sigma^2_{\hat{D}}} , \tag{17}$$

which, according to the definition of $D\hat{T}F$ for N_F examinees (Equation 3), can be expressed as

$$\chi^2_{N_F} = \frac{N_F(D\hat{T}F)}{\sigma^2_{\hat{D}}} . \tag{18}$$

Substituting the sample-based estimate of the variance of $\hat{D}$, Equation 18 can be rewritten as

$$\chi^2_{N_F-1} = \frac{N_F(D\hat{T}F)}{\hat{\sigma}^2_{\hat{D}}} . \tag{19}$$

This chi-square test (with $N_F - 1$ degrees of freedom) may prove useful in practice in determining whether an observed (or sample-based) DTF index is significantly different from zero. Another statistical test which may prove useful in the present context is $\sqrt{N_F}(\hat{\mu}_{\hat{D}} - \mu_{\hat{D}}) / \hat{\sigma}_{\hat{D}}$, which, according to the previously stated assumptions for the chi-square test, has a t distribution with $N_F - 1$ degrees of freedom. Since the t and chi-square tests are likely to lead to very similar conclusions when N_F is large, Raju et al. (1995) recommend the chi-square test because of its explicit relationship to DTF.

When an observed $D\hat{T}F$ index is statistically significant, one may begin the search for items that may be causing the significant chi-square. After such items are identified and removed from the test, the $D\hat{T}F$ index and its chi-square should be recomputed with the remaining items. Since the value for $\hat{\text{Cov}}(\hat{d}_i, \hat{D})$ depends on, among other things, the number of items that are still in the test, it is recommended that a single item be identified for removal at a time and that the process be continued until the chi-square associated with the revised $D\hat{T}F$ index becomes nonsignificant. Since item $CD\hat{I}F$ indexes add up to the total test $D\hat{T}F$ index, when a given $D\hat{T}F$ index is statistically significant, items with large, positive $CD\hat{I}F$ indexes should be deleted, one at a time, until the $D\hat{T}F$ index based on the remaining items becomes statistically nonsignificant. All such deleted items will then be labeled "DIF" or characterized as having significant $CD\hat{I}F$ indexes. No separate significance test, therefore, is proposed for the $CD\hat{I}F$ index.

Chi-square test for NC$\hat{\text{D}}$IF. In light of the significance test defined above for $D\hat{T}F$, a chi-square significance test, given that $\hat{d}_i$ is normally distributed with a finite variance, may be similarly defined for the (sample-based) $NC\hat{D}IF$ index for item i as

$$\chi^2_{N_F - 1} = \frac{N_F\,(NC\hat{D}IF_i)}{\hat{\sigma}^2_{\hat{d}_i}}. \tag{20}$$

The degrees of freedom for this chi-square test are also equal to $N_F - 1$.

An exploratory Monte Carlo examination of the chi-square test for the $NC\hat{D}IF$ index showed that this index was overly sensitive for large sample sizes (Fleer, 1993). In the no-DIF condition (i.e., identical true item parameters in the focal and reference groups), the percentage of items identified as DIF at the .01 level of significance was substantially greater than 1%. Therefore, after several replications with the no-DIF condition, Fleer found that a cutoff score of .006 for the $NC\hat{D}IF$ index resulted in the false identification of approximately 1% of the items as DIF. Other Monte Carlo studies on DFIT (Fleer, Raju, & van der Linden, 1995; Oshima, Raju, Flowers, & Monaco, 1995) used .006 as a cutoff criterion and showed favorable results. Although a comprehensive review of the adequacy of the criterion is in order, for this article the cutoff of .006 for the $NC\hat{D}IF$ index will be used in identifying false positives and false negatives.

Multidimensional Linking

In the unidimensional case, prior to a DIF analysis, the estimated item parameters for the reference group (for example) are transformed to a scale underlying the estimated item parameters for the focal group because the item parameters from two subpopulations are only invariant up to a linear transformation (Lord, 1980). It is also necessary to put the items on a common scale in the multidimensional case.

For the multidimensional IRT models with the exponent expressed as $\mathbf{a}'\boldsymbol{\theta} + b$, the probability of correct response is not altered by the following transformations:

$$\mathbf{a}^* = (\mathbf{A}^{-1})'\mathbf{a} \tag{21}$$

$$b^* = b - \mathbf{a}'\mathbf{A}^{-1}\boldsymbol{\beta} \tag{22}$$

$$\boldsymbol{\theta}^* = \mathbf{A}\boldsymbol{\theta} + \boldsymbol{\beta} \tag{23}$$

where $\mathbf{A}$ is an $m \times m$ multiplicative linking matrix and $\boldsymbol{\beta}$ is an $m \times 1$ additive linking vector for the m-dimensional IRT models. The multiplicative linking matrix adjusts variance and covariance differences of ability dimensions for the two groups, and the additive linking vector adjusts the location differences. The equations above are slightly modified from the multidimensional linking originally introduced by Davey (1991).

The multidimensional linking procedure introduced here is an extension of the test characteristic function (TCF) method (Stocking & Lord, 1983). The $\mathbf{A}$ and $\boldsymbol{\beta}$ are sought to minimize the difference between two test characteristic functions on certain matching points of $\boldsymbol{\theta}$. The function to be minimized (F_1) is

$$F_1 = \frac{1}{L}\sum_{i=1}^{L}(T_F - T_R)^2 \tag{24}$$

for L equally spaced $\boldsymbol{\theta}$ points in the m-dimensional space. For example, the two-dimensional test characteristic surfaces can be evaluated at L (e.g., $7 \times 7 = 49$) grid points evenly spaced on the square defined by the corners (-4, -4), (-4, 4), (4, -4), and (4, 4). A detailed explanation of this multidimensional linking is given elsewhere (Oshima, Davey, & Lee, 1996).

A further modification of this TCF method was made in accordance with the definition of DTF previously described. Instead of equally spaced $\boldsymbol{\theta}$ points, $\boldsymbol{\theta}$ points of the entire focal group were used as matching points. The resulting function to be minimized (F_2) is

$$F_2 = \frac{1}{N_F} \sum_{i=1}^{N_F} (T_F - T_R)^2 . \tag{25}$$

Notice that the function to be minimized is precisely what DTF is. Therefore, what is left from linking defines DTF. For this reason, when the presence of DIF is suspected or unknown, it is crucial to employ iterative linking (Candell & Drasgow, 1988) so that potentially DIF items are excluded from the calculation of the linking matrix and vector.

It is important to distinguish the DIF procedure used in the process of iterative linking from the final DIF procedure itself. In iterative linking, the role of the intermediate DIF procedure is to identify "large" DIF items so that linking can be conducted with no or a minimum number of DIF items. Only after the desirable linking is achieved, the DIF/DTF indexes are ready to be interpreted. This point is especially relevant for DTF. In the DFIT framework, the initial DTF (i.e., before iterative linking) is likely to be nonsignificant, provided that the minimization of F_2 has been successful. This DTF, however, is not very interpretable, because linking coefficients are most likely to be contaminated by the possible presence of DIF items.

Raju and his colleagues recommended the use of *NCDIF* in selecting items to be used for linking. Their rationale is that *NCDIF* is similar to other existing IRT-based DIF indexes for which iterative linking is commonly exercised. Although it is also possible to use *CDIF* in lieu of *NCDIF*, the use of *CDIF* for iterative linking has not yet been investigated.

Method

Design

Although the DFIT procedure described earlier can be applied to m-dimensional data in general, the simplest case ($m = 2$) is demonstrated in this article. Using a compensatory multidimensional two-parameter logistic (M2PL) model (Equation 1), 40-item, two-dimensional data sets were generated. Factors of interest in this study were (a) uniform versus nonuniform DIF, (b) unidirectional versus balanced-bidirectional DIF, and (c) $\boldsymbol{\theta}$ distributional differences for the reference group and the focal group. Other factors such as the number of DIF items and the magnitude of DIF were held constant. In the two-dimensional structure in which both dimensions are intended to be measured, items measured both θ_1 and θ_2 throughout the test to various degrees (i.e., a multidimensional within-item test).

The first factor of interest had two levels: uniform DIF and nonuniform DIF. The uniform DIF condition was defined as a difference in the reference group and focal group b parameters in Equation 1. The nonuniform DIF condition was defined as a difference in the $\mathbf{a}$ parameter vector (with elements a_1 and a_2), with or without a difference in the b parameter. It should be noted that there are various combinations to create nonuniform DIF situations. Please refer to Swaminathan and Rogers (1990) for additional information about uniform and nonuniform DIF in the unidimensional case.

The second factor consisted of unidirectional and balanced-bidirectional DIF conditions. In the unidirectional condition, all DIF items favored the reference group. On the other hand, in the balanced-bidirectional condition, one half of the DIF items favored the focal group, and the other half of the DIF items favored the reference group *to the same degree*. As previously noted (see Equation 7), two items with balanced but opposite DIF will cancel each other out in terms of their contribution to DTF at the *examinee level*; these items will therefore be considered free of DIF as far as the CDIF indexes are concerned. The same two items, however, may be considered to have significant DIF within the context of NCDIF definition. Items representing both unidirectional and balanced-bidirectional DIF conditions were included in the current study to assess the sensitivity of the proposed statistical tests to these types of DIF.

The last factor of interest concerned distributional differences of $\boldsymbol{\theta}$s between the two groups. In the multidimensional context, a distributional difference can arise from differences in the variance-covariance structure of $\boldsymbol{\theta}$s and/or the location of $\boldsymbol{\theta}$s. There are many possible differences one can generate. However, for the current study, only four different cases were used. The first case (Case A) was the situation where both groups had $\boldsymbol{\theta}$s drawn from a bivariate normal distribution with zero means, unit variances, and $\rho = 0$. The second case (Case B) is the same as Case A but with $\rho = .5$. The third case (Case C) is the same as Case B, but with a location difference of .5 on θ_2. In other words, this is the situation where, for both groups, $\rho = .5$, but the focal group had a lower mean on θ_2. Recall that θ_2 is an intended-to-be-measured trait. Therefore, this distributional difference alone should not result in DIF. The last case (Case D) reflects a correlation difference between the two groups. The reference group had $\rho = .5$, and the focal group had $\rho = .0$, both with zero means and unit variances.

The constants for the other factors, such as the number of DIF items and the magnitude of DIF, were selected to model practical situations. The number of DIF items was four on the 40-item test. The constants .3 and .5 were used to create the difference on an element of the $\mathbf{a}$ parameter vector and the b parameter, respectively. These constants were added or subtracted, depending on the condition, to or from the item parameters for either the focal group or the reference group.

The .3 difference on an element of the $\mathbf{a}$ parameter vector was selected to coincide with other DIF studies. For example, Kim and Cohen (1992) had an a difference (in a unidimensional IRT setup) of .16 or .32.

Data Generation

Item parameters were generated to create a test which measures θ_1 and θ_2 throughout the test. The item direction parameter[2] (α) ranges from 0° to 90° and defines the degree to which each item measures θ_1 and θ_2. In this test, the item directions of 0°, 30°, 45°, 60°, and 90° were embedded systematically throughout the test. Item parameters are listed in Table 1.

Ability parameters (θ_1 and θ_2) were simulated from a random normal distribution with a mean of 0 and a standard deviation of 1. Then, a set of correlated $\boldsymbol{\theta}$ was generated for some conditions. Correlated θ_1 and θ_2 were simulated by first generating two independent, normally distributed pseudorandom variables z_1 and z_2 and then transforming them to θ_1 and θ_2 by weighted linear transformations. The weights were the elements of $\mathbf{T}'$, a matrix which satisfies $\mathbf{R} = \mathbf{T}'\mathbf{T}$, where $\mathbf{R}$ is the target correlation matrix. The sample size for each group was 1,000. Additional details on simulating multidimensional test data can be found in Oshima and Miller (1992).

For organizational clarity, DIF items were shifted to the end of the test (Items 37–40). In addition, for these four items, multidimensional discrimination[3] (*MDISC*) and multidimensional item difficulty[4] (*MID*) parameters for the reference group were replaced with the average values of the respective parameters (*MDISC* = 1.13, *MID* = 0) to avoid any unnecessary effect of discrimination and/or difficulty of the item on the detection of DIF. It has been shown, for example, that DIF items with higher discrimination parameters were more likely to be identified as DIF items (Oshima & Miller, 1992). Table 2 presents item parameters for the four DIF items.

Two conditions were considered under uniform DIF (Conditions 1 and 2). In Condition 1 (uniform and unidirectional DIF), the item directions for the four DIF items were 0°, 30°, 60°, and 90°. In this condition, the *b* parameter for all four DIF items was lowered by .5 for the focal group, thus making these items harder for the focal group. In Condition 2 (uniform and balanced-bidirectional DIF), item directions were either 30° or 60°. Items 37 and 38 favored the reference group, whereas Items 39 and 40 favored the focal group to the same degree. A close examination of this condition reveals that the reference group item parameters for Items 37 and 38 were identical to the focal group item parameters for Items 39 and 40; similarly, the reference group item parameters for Items 39 and 40 were identical to the focal group item parameters for Items 37 and 38. That is, at the item level, Items 37–40 had DIF, but that DIF had no effect on the DTF index.

Two conditions were also considered under nonuniform DIF (Conditions 3 and 4). For nonuniform unidirectional DIF (Condition 3), each of the four items had a different pattern of nonuniform DIF, while the item directions were held constant at 45° for the reference group. Only in Item 40, both the a_1 and a_2 parameters and the *b* parameter were lowered for the focal group. Finally, in Condition 4 (nonuniform and balanced-bidirectional DIF), two types of nonuniform DIF were considered: difference in **a** only, and differences in **a** and *b*. Again, as in Condition 2, items were arranged so that there was differential functioning only at the item level, but not at the test levelL

Analysis

Item parameters for the generated data were calibrated using NOHARM (Fraser, 1988). In all the conditions in this study, an exploratory analysis was used in which

Table 1. True Item Parameters Before DIF 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
Mean		0.69	0.63	0.02
SD		0.73	0.58	0.92

Table 2

Item Parameters for Generating DIF Conditions for the Last Four Items

(a) Uniform DIF

Condition		Reference Group				Focal Group			
	Item	*α*	*a*	*a*	*b*	*α*	*a*	*a*	*b*
Unidirectional **(Condition 1)**									
	37	0	1.13	.00	.0	0	1.13	.00	-.5
	38	30	.98	.57	.0	30	.98	.57	-.5
	39	60	.57	.98	.0	60	.57	.98	-.5
	40	90	.00	1.13	.0	90	.00	1.13	-.5
Balanced-Bidirectional **(Condition 2)**									
	37	30	.98	.57	.0	30	.98	.57	-.5
	38	60	.57	.98	.0	60	.57	.98	-.5
	39	30	.98	.57	-.5	30	.98	.57	.0
	40	60	.57	.98	-.5	60	.57	.98	.0

(b) Non-Uniform DIF

Condition		Reference Group				Focal Group			
	Item	*α*	*a*	*a*	*b*	*α*	*a*	*a*	*b*
Unidirectional **(Condition 3)**									
	37	45	.80	.80	.0	58	.50	.80	.0
	38	45	.80	.80	.0	45	.50	.50	.0
	39	45	.80	.80	.0	69	.50	1.30	.0
	40	45	.80	.80	.0	45	.50	.50	-.5
Balanced-Bidirectional **(Condition 4)**									
	37	45	.80	.80	.0	45	.50	.50	.0
	38	45	.80	.80	.0	45	.50	.50	-.5
	39	45	.50	.50	.0	45	.80	.80	.0
	40	45	.50	.50	-.5	45	.80	.80	.0

the user does not specify the pattern matrices of **F** (composed of **a**s for the 40 items) except that a_2 of Item 1 is set to be zero. This restriction on Item 1 is imposed to solve the rotational indeterminacy. In the exploratory analysis, the correlation matrix of $\boldsymbol{\theta}$ (called the **P** matrix) is set to be an identity matrix. Regardless of the true ability distributions of examinees, **P** is always an identity matrix. However, the correlation of $\boldsymbol{\theta}$ is reflected in the **a** estimates. In fact, $\boldsymbol{\theta}$ can be transformed into any correlation of $\boldsymbol{\theta}$ by the transformation equation (Equation 23), provided the **a** estimates and *b* estimates are also transformed (Equations 21 and 22), respectively. For this reason, it does not seem crucial to recover the original $\boldsymbol{\theta}$ distributions in DIF studies as long as an appropriate linking is performed. In theory, any distributional difference of intended-to-be-measured traits (location, variance, and covariance) between two groups should be corrected by the linking process.

The linking analysis was conducted using a computer program called IPLINK (Lee & Oshima, 1996). This computer program implemented the linking procedure described earlier. A two-stage iterative linking procedure was used. That is, items showing a fairly large *NCDIF* (> .006) were eliminated first, and linking was performed again using the remaining items. Using the second-stage linking coefficients, all item parameters for the reference group were transformed.

Finally, the DFIT program was used to calculate DIF and DTF indexes. As described earlier, the standard DFIT analysis requires estimated ability parameters. As an alternative to using estimated **θ**s, we propose that simulated **θ**s be used. In other words, **θ**s randomly sampled from a multivariate standard normal distribution can be used. In practice, we can provide a set of **θ**s as part of the DFIT program. This approach is a simplification of calculating *DTF* as an expectation with respect to **θ**s with a multivariate standard normal distribution. Since the theory used in NOHARM also assumes a multivariate standard normal distribution for latent traits, the use of this distribution seems justifiable. Incidentally, the current NOHARM does not provide **θ** estimates. Even when **θ** estimates become available, the use of simulated **θ**s included in the DFIT program may still offer an advantage over estimated **θ**s in terms of convenience for practitioners. In the present study, the same simulated or predetermined **θ**s with mean zero, unit variance, and $\rho = 0$ were used in all conditions. Note that these simulated **θ**s are independent of the **θ**s used for data generation.

Results and Discussion

As noted earlier, the purpose of this research was to describe and demonstrate Raju et al.'s (1995) DFIT indexes in the multidimensional context. It is important to keep in mind that a recovery analysis such as the one used in this study involves the evaluation of not only the DFIT technique but also the performance of the calibration program and the linking procedure. To separate the issue of the performance of the DFIT technique from the performance of the NOHARM calibration program and the multidimensional linking procedure, the DFIT analysis was first conducted with true item parameters (the true condition) and later repeated with estimated item parameters (the estimated condition). The first analysis was considered to be an optimal condition for the DFIT technique, because the NOHARM calibration errors and subsequent scaling errors were removed from influencing the DFIT indexes. On the other hand, the latter analysis, which can be used in practice, reflected the calibration and scaling errors. Reported in Table 3 are the results from all the conditions studied (single replication for each condition), including the true (indicated in boldface) and estimated conditions.

DFIT Analysis With True Item Parameters ("True" Conditions)

In all conditions, the CDIF and NCDIF indexes for the non-DIF items (Items 1–36) were .000, as expected (not shown in Table 3). As shown in Table 3, in the uniform and unidirectional condition (Condition 1), the CDIF indexes were .039, .040, .040, and .040 for the four DIF items (Items 37–40), respectively, and the DTF index was .159, with a statistically significant chi-square. When Items 40, 39, 37, and 38 were eliminated successively, the DTF index was no longer significant. The NCDIF index of .010 was the same for all four DIF items and was statistically significant. In the uniform and balanced-bidirectional condition (Condition 2), the

Table 3 CDIF and NCDIF Indices With True and Estimated Item Parameters Under Four Different θ Distributions (A, B, C, and D).

(a) Uniform DIF

	CDIF					NCDIF				
	True	Estimated				**True**	Estimated			
		Case					Case			
		A	B	C	D		A	B	C	D
Unidirectional (Condition 1)										
						FP Rate				
						0/36	0/36	0/36	0/36	0/36
CDIF for Items 37-40						NCDIF for Items 37-40				
Item 37	**.039**	.034	.020	.023	.036	**.010**	.013	.010	.013	.013
Item 38	**.040**	.029	.016	.017	.031	**.010**	.009	.007	.008	.010
Item 39	**.040**	.034	.020	.018	.034	**.010**	.012	.010	.008	.012
Item 40	**.040**	.020	.014	.013	.020	**.010**	.005	.006	.005	.005
DTF	**.159**	.099	.043	.044	.106					
Deleted Items	**40** **39** **37** **38**	37 38 39	37 39	37 38	39 38 37					
Balanced-Bidirectional (Condition 2)										
						FP Rate				
						0/36	0/36	0/36	0/36	2/36
CDIF for Items 37-40						NCDIF for Items 37-40				
Item 37	**.000**	-.001	.009	.000	-.001	**.010**	.012	.010	.011	.013
Item 38	**.000**	.003	.009	.002	.002	**.010**	.008	.009	.008	.009
Item 39	**.000**	.008	-.007	.001	.002	**.010**	.009	.006	.007	.008
Item 40	**.000**	.011	-.007	.005	.003	**.010**	.016	.011	.015	.016
DTF	**.000**	.035	.017	.012	.029					
Deleted Items	**none**	none	37	none	none					

CDIF index was .000 for all four DIF items, with a DTF of .000. As expected, the NCDIF indexes were all equal to .010 and were statistically significant.

In the nonuniform DIF conditions (Conditions 3 and 4), a similar trend was observed. The effect of a parameter vector differences was evident in Condition 3. The difference in both a_1 and a_2 parameters produced larger CDIF and NCDIF indexes than the difference only in a_1 parameters. When a_1 and a_2 parameters were both different but in opposite directions (Item 39), the CDIF index was about the same as, and the NCDIF index was larger than, when the a_1 and a_2 parameters were both different in the same direction (Item 38). The largest CDIF and NCDIF indexes were observed when a_1, a_2, and b parameters were all different across groups (Item 40). Only Item 40 needed to be eliminated to achieve nonsignificant DTF. The NCDIF indexes for Items 39 and 40 were statistically significant. The result that some nonuniform DIF items were not identified as DIF items should be interpreted with caution. As is the case in any DIF analysis, the detection of

(b) Non-Uniform DIF

	CDIF					NCDIF				
	True	Estimated				**True**	Estimated			
		Case					Case			
		A	B	C	D		A	B	C	D
Unidirectional (Condition 3)										
						FP Rate				
						0/36	0/36	0/36	0/36	0/36
CDIF for Items 37-40						NCDIF for Items 37-40				
Item 37	**.012**	.005	.005	.001	.027	**.003**	.002	.007	.004	.013
Item 38	**.012**	.004	.006	.007	.023	**.005**	.002	.006	.006	.007
Item 39	**.014**	.009	-.001	-.001	.016	**.011**	.008	.006	.006	.010
Item 40	**.024**	.015	.006	.016	.025	**.018**	.014	.020	.015	.022
DTF	**.062**	.026	.010	.023	.094					
Deleted Items	**40**	40	none	40	none					
Balanced-Bidirectional (Condition 4)										
						FP Rate				
						0/36	1/36	0/36	0/36	0/36
CDIF for Items 37-40						NCDIF for Items 37-40				
Item 37	**.000**	-.001	-.002	.008	-.010	**.005**	.004	.008	.010	.005
Item 38	**.000**	-.008	-.004	-.001	-.007	**.018**	.017	.016	.012	.017
Item 39	**.000**	.006	.003	-.005	.014	**.005**	.008	.007	.005	.009
Item 40	**.000**	.010	.006	.001	.015	**.018**	.024	.022	.018	.028
DTF	**.000**	.028	.013	.021	.035					
Deleted Items	**none**	27	none	40	none					

Note.
' Case A (ρ = .0, ρ = .0), Case B (ρ = .5, ρ = .5), Case C (ρ = .5, ρ = .5, and location difference of .5 on θ), and Case D (ρ = .0, ρ = .5)
False positive rate for NCDIF, i.e., the number of items with NCDIF > .006 for Items 1-36.
Items to be deleted to achieve non-significant DTF.

DIF/DTF depends on the magnitude of DIF. With a different (i.e., larger) magnitude of DIF and/or different item parameter characteristics (i.e., either more difficult or easy items and/or more discriminating items), results are likely to change. In the nonuniform and balanced-bidirectional condition (Condition 4), the DTF index was zero, and two of the four items (Items 38 and 40) had statistically significant NCDIF indexes; the NCDIF indexes for the remaining two items (Items 37 and 39) barely missed being significant.

It should be noted that the simulated (as opposed to estimated) θs were used as ability parameters for calculating the DFIT indexes for the true condition described above. The very same set of θs was also used in the subsequent DFIT analysis with estimated item parameters.

DFIT Analysis With Estimated Item Parameters ("Estimated" Conditions)

Results from estimated item parameters are also shown in Table 3, next to the results for the true parameters. The four different cases of θ distributions are labeled as A, B, C, and D.

In the uniform and unidirectional condition (Condition 1) across the A–D cases, 2 to 3 out of 4 possible CDIF items were identified. There were no false positives (FPs) for *CDIF*. This suggests that the test is slightly conservative. It could be due to the linking method employed in this study. *NCDIF* performed fairly well. The number of FPs was zero for all of Cases A–D. The true positive (TP) rate was 3/4 (75%) in Cases A–D using the .006 criterion. It is difficult to determine what percentage of TP is "good." In the unidimensional case, a study reported a TP rate of 47–62% when 10% of the items were DIF items using the area measures (Oshima & Miller, 1992). All of the last four items showed some degree of NCDIF, although some items did not reach significance.

For the uniform and balanced-bidirectional condition (Condition 2), no CDIF item was identified, as expected, except for Case B. Even in Case B, only one item was erroneously identified as CDIF. *NCDIF* showed similar results for Condition 1, except that 2 out of the 36 non-DIF items were identified as having significant NCDIF in Case D.

For the nonuniform unidirectional condition (Condition 3), Item 40, which was the only CDIF item according to the analysis with true parameters, was slightly underidentified as CDIF. However, again, the FP rate for *CDIF* was zero. For the two items that had significant NCDIF in the true condition (Items 39 and 40), most of them were identified as having significant NCDIF, or, if not, they were close to the cutoff of .006.

Finally, for the nonuniform balanced-bidirectional condition (Condition 4), again, there was a slight (one item at most) overidentification of CDIF. The TP rate for *NCDIF* ranged from 75% to 100%, with no FPs in most cases.

In general, the agreement between the true and estimated conditions was fairly close in all the conditions. It is interesting to note that after the first-stage linking, DTF in *all* conditions was nonsignificant in this study, suggesting that IPLINK successfully minimized DTF. After the second-stage linking, on the other hand, significant DTF emerged when expected. These results confirm our belief that the iterative linking is a necessary step in the DFIT framework.

The most interesting finding is that the agreement between the true and estimated conditions did not deteriorate as the distributional differences of θs were introduced (Cases A–D). As mentioned earlier, linking should take care of distributional differences. In other words, the comparison of Cases A–D is an evaluation of the linking procedure, not an evaluation of DFIT. As shown under the true condition, if there is no linking error (and also no calibration error), *DFIT* performs as expected.

The DFIT procedure did distinguish between (a) a situation in which there is DIF but the θ distributions are the same for the two groups and (b) a situation in which there is no DIF but the groups have different θ distributions. To investigate the latter situation (i.e., the null condition), data were simulated with identical item parameters (thus no DIF) but different distributions ($\rho = .5$ for both groups, but the mean of θ_2 was higher for one group by .5). The results of this investigation showed no FPs for *CDIF* (i.e., nonsignificant DTF) and no FPs for *NCDIF*. In fact, the highest NCDIF index was .001, with most of the items having an NCDIF index of .000.

A cautious interpretation is necessary for Table 3, especially for Case D. The results are based on only one replication for the purpose of demonstration. Our investigations in the area of linking and DFIT suggest that linking in the presence of differences in correlation can be quite difficult as opposed to linking in the presence of mean and/or variance differences. Although Case D showed reasonable accuracy for this particular case, it would be safe to recommend the use of the DFIT procedure when two groups have similar correlations.

Conclusions

This article first described the theoretical framework for investigating DIF and DTF using the DFIT procedure with intentionally m-dimensional data. Then, it was empirically shown that the DIF and DTF embedded in the intentionally two-dimensional test were recovered reasonably well under various distributional differences of $\boldsymbol{\theta}$s after multidimensional linking was applied to put the two sets of item parameters on a common scale. There are two areas for future investigations to enhance the performance of the DFIT procedure in the multidimensional context. The first is the area of multidimensional calibration and multidimensional linking. The second is, of course, the DFIT procedure itself.

Concerning the first area, the recovery of item parameters by NOHARM or other multidimensional calibration programs needs to be investigated. We used a sample size of 1,000 and a test of 40 items. Further research can elaborate on the issue of the stability of NOHARM calibrations as a function of sample size and the number of items in a test. The impact of the type of linking on DFIT needs to be investigated. As described earlier, there is a close relationship between linking and DFIT. A newly developed program, IPLINK, can link any number of dimensions with different types of linking algorithms. An evaluation of IPLINK is currently underway.

Concerning the DFIT procedure itself, the role of estimated ability parameters in DFIT needs to be investigated. A new IRT calibration program, soon to be commercially available, estimates person and item parameters for multidimensional and unidimensional data with either dichotomous or polytomous scoring (E. Muraki, personal communication, April 10, 1996).

The major challenge of the application of the DFIT procedure appears to be the issue of hypothesis testing in relation to sample size and alternative approaches to interpret the magnitude of DIF and DTF. While the results from the proposed significance tests and/or empirically determined cutoff levels appear to be promising, there is still a need for further research on the distribution of D, which is assumed to be normal, and on the empirically determined cutoff level (.006) used with the DTF and NCDIF indexes. Currently, a study is being conducted to investigate the adequacy of the .006 cutoff level under various sample sizes, numbers of parameters, and linking procedures. There is also a need for significance tests which take into account the estimation and linking errors associated with $\hat{\boldsymbol{\theta}}$, $\hat{\mathbf{a}}$, and $\hat{b}$. In addition, further study on variables that impact the DFIT indexes is required. The influence of sample size (smaller than the one used in this study), relative amounts of DIF in items (e.g., minimum required for detection), number and mix (unidirectional and bidirectional) of DIF items, different test lengths, and different numbers of ability dimensions should be systematically investigated. This study used only one replication to demonstrate the applicability of the DFIT framework within the two-dimensional IRT context. Future research should include additional replications to assess Type I and Type II error rates.

Obviously, there is need for further research with respect to the issue of DIF with multidimensional data sets. This article has demonstrated a possible starting point for this challenging but potentially very useful area of research. The generalizability of our results is limited to the multidimensional structure and items parameters used in this study.

References

Ackerman, T. A. (1991). The use of unidimensional parameter estimates of multidimensional items in adaptive testing. *Applied Psychological Measurement, 15*, 13–24.

Candell, G. L., & Drasgow, F. (1988). An iterative procedure for linking metrics and assessing item bias in item response theory. *Applied Psychological Measurement, 12*, 253–260.

Davey, T. C. (1991). *Some issues in linking multidimensional item calibrations.* Paper presented at the Office of Naval Research Contractors Meeting on Model-Based Psychological Measurement, Princeton, NJ.

Fleer, P. F. (1993). *A Monte Carlo assessment of a new measure of item and test bias.* Unpublished doctoral dissertation, Illinois Institute of Technology, Chicago.

Fleer, P. F., Raju, N. S., & van der Linden, W. J. (1995, April). *A Monte Carlo assessment of DFIT with dichotomously scored unidimensional tests.* Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco.

Fraser, C. (1988). *NOHARM: A computer program for fitting both unidimensional and multidimensional normal ogive models of latent trait theory.* New South Wales, Australia: Center for Behavioral Studies, University of New England Armidale.

Kim, S., & Cohen, A. S. (1992). Effects of linking methods on detection of DIF. *Journal of Educational Measurement, 29*, 51–66.

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

Longford, N. T. (1995). *Models for uncertainty in educational testing.* New York: Springer-Verlag.

Longford, N. T., Holland, P. W., & Thayer, D. T. (1993). Stability of MH D-DIF statistics across populations. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 171–196). Hillsdale, NJ: Erlbaum.

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

Oshima, T. C., Davey, T. C., & Lee, K. (1996). *Multidimensional linking: Four minimization procedures.* Unpublished manuscript.

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

Oshima, T. C., Raju, N. S., Flowers, C. P., & Monaco, M. (1995, April). *A Monte Carlo assessment of DFIT with dichotomously scored multidimensional tests.* Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco.

Raju, N. S. (1988). The area between two item characteristic curves. *Psychometrika, 53*, 495–502.

Raju, N. S. (1990). Determining the significance of estimated signed and unsigned areas between two item response functions. *Applied Psychological Measurement, 14*, 197–207.

Raju, N. S., van der Linden, W. J., & Fleer, P. F. (1992, April). ***An IRT-based internal measure of test bias with applications for differential item functioning.*** Paper presented at the Annual Meeting of American Educational Research Association, San Francisco.

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

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

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

Rudner, L. M. (1977, April). ***An approach to biased item identification using latent trait measurement theory.*** Paper presented at the Annual Meeting of American Educational Research Association, New York.

Shealy, R. T., & Stout, W. F. (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.

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 7*, 210–210.

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

Thissen, D., Steinberg, L., & Wainer, H. (1988). Use of item response theory in the study of group differences in trace lines. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 147–169). Hillsdale, NJ: Erlbaum.

Traub, R. E. (1983). A priori considerations in choosing an item response model. In R. K. Hambleton (Ed.), *Applications of item response theory* (pp. 57–70). Vancouver: Educational Research Institute of British Columbia.

Wang, W., Wilson, M., & Adams. (1995, April). *Item response modeling for multidimensional between-items and multidimensional within-items.* Paper presented at the International Objective Measurement Conference, Berkeley, CA.

Notes

[1] The exponent of the multidimensional IRT model is commonly expressed as $\mathbf{a}'\boldsymbol{\theta} + d$ or $\mathbf{a}'(\boldsymbol{\theta} - \mathbf{b})$, where $d = -\mathbf{a}'\mathbf{b}$. The first parameterization is used in this article. However, the notation d was replaced with b to avoid the confusion with another d repeatedly used as a DIF measure in the DFIT framework.

[2] In the M2PL model, item directions (α_{i1}) determine the weighted composite of traits measured by an item. The angle can be determined using the direction cosines given by

$$\cos \alpha_{ik} = \frac{a_{ik}}{\sqrt{\sum_{k=1}^{m} a_{ik}^2}},$$

where a_{ik} is the kth element of the vector $\mathbf{a}_i$. In the two-dimensional space, if an item measures only θ_1, then α_{i1} is 0°; if an item measures only θ_2, then α_{i1} is 90°. α_{i1} can be any value from 0° to 90°, depending on the degree to which an item measures the two traits.

[3] Reckase and McKinley (1991) defined *MDISC* as

$$MDISC_i = \sqrt{\sum_{k=1}^{m} a_{ik}^2}.$$

[4] Reckase (1985) defined *MID* as

$$MID_i = -\frac{b_i}{MDISC_i}.$$

Authors

T. C. OSHIMA is Associate Professor, Department of Educational Policy Studies, Georgia State University, University Plaza, Atlanta, GA 30303; oshima@gsu.edu. *Degree:* PhD, University of Florida. *Specialization:* educational measurement.

NAMBURY S. RAJU is Distinguished Professor and Director, Center for Research and Service, Institute of Psychology, Illinois Institute of Technology, Chicago, IL 60616-3793; nsraju@charlie.cns.iit.edu. *Degrees:* BA, Madras University; MS, Purdue University; MA, PhD, Illinois Institute of Technology. *Specializations:* psychometric theory, test development, industrial/organizational psychology.

CLAUDIA P. FLOWERS is Assistant Professor, Department of Educational Administration, Research and Technology, University of North Carolina at Charlotte, 9201 University City Boulevard, Charlotte, NC 28223-0001; cpflower@email.uncc.edu. *Degree:* PhD, Georgia State University. *Specialization:* educational measurement.