A Description and Demonstration of the Polytomous-DFIT Framework

Claudia P. Flowers, T. C. Oshima and Nambury S. Raju

Applied Psychological Measurement 1999; 23; 309
DOI: 10.1177/01466219922031437

A Description and Demonstration of the Polytomous-DFIT Framework

Claudia P. Flowers, University of North Carolina at Charlotte

T. C. Oshima, Georgia State University

Nambury S. Raju, Illinois Institute of Technology

Raju, van der Linden, & Fleer (1995) proposed an item response theory based, parametric differential item functioning (DIF) and differential test functioning (DTF) procedure known as differential functioning of items and tests (DFIT). According to Raju et al., the DFIT framework can be used with unidimensional and multidimensional data that are scored dichotomously and/or polytomously. This study examined the polytomous-DFIT framework. Factors manipulated in the simulation were: (1) length of test (20 and 40 items), (2) focal group distribution, (3) number of DIF items, (4) direction of DIF, and (5) type of DIF. The findings provided promising results and indicated directions for future research. The polytomous DFIT framework was effective in identifying DTF and DIF for the simulated conditions. The DTF index did not perform as consistently as the DIF index. The findings are similar to those of unidimensional and multidimensional DFIT studies. *Index terms: differential functioning of items and tests, differential item functioning, differential test functioning, polytomous data, simulation, unidimensionality.*

Differential test functioning (DTF) and differential item functioning (DIF) research has focused primarily on dichotomously scored items and tests. With the increased use of polytomously scored items and evidence of greater discrepancy in ethnic groups' performance using performance-based assessment (Dunbar, Koretz, & Hoover, 1991; Zwick, Donoghue, & Grima, 1993), there has been increased interest in polytomous DIF/DTF procedures. A new item response theory (IRT) based, parametric procedure proposed by Raju, van der Linden, & Fleer (1995), known as differential functioning of items and tests (DFIT), can be used with unidimensional and multidimensional data derived from dichotomous and/or polytomous scoring.

The DFIT framework has many useful features for test developers. First, it is the only parametric, IRT-based, psychometric measure of differential functioning at both the test and item levels. When IRT is used to develop tests, IRT-based DIF/DTF procedures that use item parameter estimates, such as DFIT, maintain a common framework in test development. Second, DFIT provides an index that does not assume that all items in the test, other than the item under study, are unbiased. Third, during the development phase, DFIT provides a DTF procedure for determining the overall effect of eliminating an item from a test. Fourth, DFIT allows examining DIF/DTF in a mixed test format, such as a combination of polytomous and dichotomous items. Finally, DFIT can be extended to multidimensional data.

Raju et al. (1995) presented the theoretical framework of DFIT and offered an empirical demonstration of DFIT using dichotomous data. Oshima, Raju, & Flowers (1997) extended the DFIT framework to the dichotomous multidimensional case. This paper describes the extension of the DFIT framework to the polytomous unidimensional case, and explains the procedure for detecting

Applied Psychological Measurement, Vol. 23 No. 4, December 1999, 309–326

DIF and DTF in the polytomous-DFIT framework. The performance of the procedure is examined with simulated data.

The Graded Response Model

As with the dichotomous models, many polytomous models exist, e.g., Samejima's (1969) graded response model (GRM), Masters' (1982) partial credit model, the rating scale model (Andrich, 1978), the nominal response model (Bock, 1972), the generalized partial credit model (Muraki, 1992), and the free-response model (Samejima, 1972). The DFIT framework can be used with any polytomous model; Samejima's GRM was used in this study.

Samejima's (1969) GRM assumes an ordered response, i.e., the more steps successfully completed by the examinee, the higher the category score. The examinee is limited to selecting only one category per item. In the GRM, the probability of person s responding above category k to item i is:

$$P^*_{ik}(\theta) = \frac{\exp[Da_i(\theta_s - b_{ik})]}{1 + \exp[Da_i(\theta_s - b_{ik})]} , \tag{1}$$

where

b_{ik} is the boundary or threshold between category k and $k+1$ associated with item i,
a_i is the item slope or discrimination parameter, and
θ_s is the trait parameter.

Equation 1 is referred to as the *boundary response function* (BRF). The BRF is similar to the item response function (IRF) of the two-parameter dichotomous model, except that more than one function is needed per item. The number of functions for each item is one less than the number of response categories in that item. For example, a five-category item requires four BRFs. In the homogeneous case of the GRM, the item discrimination parameter, a_i, is assumed to be constant across all categories in item i. However, it could vary across items in a test. As a result, all BRFs have equal slopes for each category in an item, which ensures that the curves do not cross. For each item, multiple difficulty parameters, b_{ik}, are required. The number of b parameters is equal to the number of BRFs. The BRFs are a cumulative probability of a response above category k. A graphic illustration of the BRFs for a three-category item is provided in Figure 1a.

To calculate the probability of responding in a particular category, the adjacent boundary is subtracted from the cumulative probability. This can be expressed as

$$P_{ik}(\theta) = P^*_{i(k-1)}(\theta) - P^*_{ik}(\theta) . \tag{2}$$

This function is often referred to as the *item category response function* (ICRF). Because the first and last categories lack an adjacent boundary (i.e., no BRF below the first category and no BRF above the last category), Samejima (1969) defined $P^*_{i0}(\theta)$ and $P^*_{im}(\theta)$ as

$$P^*_{i0}(\theta) = 1 \tag{3}$$

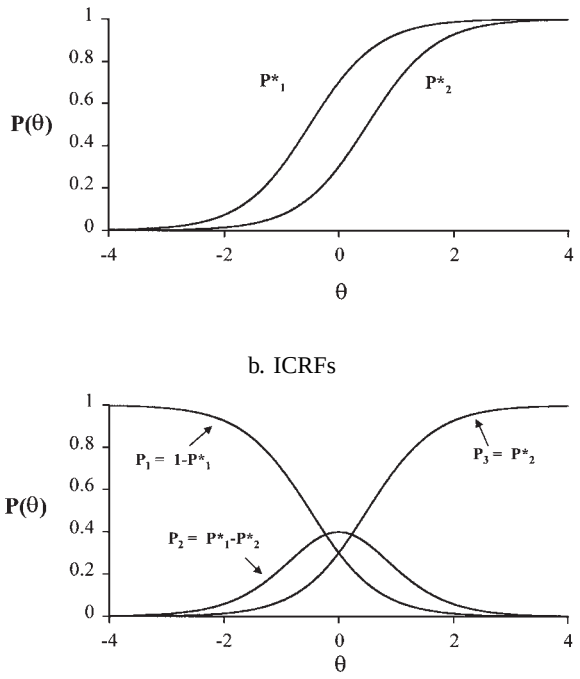
and

$$P^*_{im}(\theta) = 0 , \tag{4}$$

where m equals the number of categories. The probability of responding in the first category (i.e., $k = 1$) for item i is, then,

$$P_{i1}(\theta) = P^*_{i0}(\theta) - P^*_{i1}(\theta) = 1 - P^*_{i1}(\theta) . \tag{5}$$

Figure 1
BRFs and ICRFs for a 3-Category Item ($a = 1$, $b_1 = -.5$, $b_2 = .5$)



The probability of responding in the last category (i.e., $k = m$) for item i is

$$P_{im}(\theta) = P^*_{i(m-1)}(\theta) - P^*_{im}(\theta) = P^*_{i(m-1)}(\theta) - 0 = P^*_{i(m-1)}(\theta) \ . \tag{6}$$

The number of ICRFs per item is equal to the number of categories. A graphic illustration of the ICRFs for a three-category item is given in Figure 1b.

Expected Item and Test Scores

Once the probability for responding in each category (i.e., the ICRF) is estimated, a measure of the expected item score can be calculated. For polytomously scored data, an expected score (ES_{si}) for item i can be computed for examinee s as

$$ES_{si} = \sum_{k=1}^{m} P_{ik}(\theta_s) X_{ik} \ , \tag{7}$$

where

X_{ik} is the score or weight for category k,
m is the number of categories, and
P_{ik} is the probability of responding to category k (see Equation 2).

This is referred to as the *expected item score function* or the IRF. Summing the expected item scores across a test will result in the *expected test score function* for each examinee as

$$T_s = \sum_{i=1}^{n} ES_{si} \ , \tag{8}$$

where n is the number of items in the test.

The only difference between the dichotomous and polytomous DFIT framework is the calculation of the *ES*s. Once the expected item and test scores are known, the DFIT framework for the polytomous framework is identical to the DFIT framework for the dichotomous case.

Definition of DIF and DTF

Chang & Mazzeo (1994) demonstrated that if two items have the same *ES* or IRF in the GRM, then they must have the same number of scoring categories and the same ICRFs. Conversely, an item is considered to be functioning differentially if

$$ES_{iR} \neq ES_{iF} \ , \tag{9}$$

where $ES_{i\mathrm{R}}$ is the item expected score for an examinee in the reference group (R) (i.e., comparison group) with a given θ, and $ES_{i\mathrm{F}}$ is the item expected score for an examinee in the focal group (F) (i.e., the group of interest) with the same θ for item i (see Equation 7). A test functions differentially if

$$T_R \neq T_F \ , \tag{10}$$

where T_{R} and T_{F} are the expected test scores for the reference and focal group examinees, respectively, with the same θ.

Polytomous DFIT

The DFIT framework requires separate item parameter estimation for the reference group and the focal group. As a result, a test will have two sets of item parameters. The reference group item parameters are then linked onto the same metric as the focal group parameters using a linear transformation. The focal group θ distribution is used to calculate two ES_is (Equation 7), one using the focal group parameters and the other using the linked reference group parameters. That is, for a single examinee (with a given θ) who is a member of the focal group (F), an expected score for an item ($ES_{si\mathrm{F}}$) can be calculated using the focal group item parameters. For the same examinee, another expected score ($ES_{si\mathrm{R}}$) is calculated using the linked reference group item parameters. If the item is functioning differentially, the two expected scores will not be equal (Equation 9).

The same reasoning can be applied at the test level. The expected test score (T_s) (Equation 8) is calculated by summing the ES_{si} across all the items in the test. Two expected test scores are calculated for each focal group examinee, one score for the examinee as a member of the focal group ($T_{s\mathrm{F}}$) and one score as if a member of the reference group ($T_{s\mathrm{R}}$). The greater the difference between the two expected scores, the greater the DTF. According to Raju et al. (1995), a measure of DTF at the examinee level may be defined as

$$D_s^2 = (T_{sF} - T_{sR})^2 \ . \tag{11}$$

DTF across the focal group examinees may be defined as

$$\mathrm{DTF} = \mathrm{E}_F D_s^2 = \mathrm{E}_F (T_{sF} - T_{sR})^2 \ , \tag{12}$$

or, equivalently,

$$\text{DTF} = \int_{\theta} D_s^2 f_F(\theta) d\theta \ , \tag{13}$$

where $f_{\text{F}}(\theta)$ is the density function of θ for the focal group. Also,

$$\text{DTF} = \sigma_D^2 + (\mu_{TF} - \mu_{TR})^2 = \sigma_D^2 + \mu_D^2 \ , \tag{14}$$

where

μ_{TF} is the mean expected score for the focal group examinees,
μ_{TR} is the mean expected score for the same examinees (as if they were members of the reference group), and
σ_D^2 is the variance of D.

DIF can be derived from Equation 12. If

$$d_{si} = ES_{siF} - ES_{siR} \ , \tag{15}$$

then

$$\text{DTF} = \text{E}\left[\left(\sum_{i=1}^{n} d_{si}\right)^2\right] \ , \tag{16}$$

where n is the number of items in a test. This can be rewritten as

$$\text{DTF} = \sum_{i=1}^{n} [Cov(d_i, D) + \mu_{d_i}\mu_D] \ , \tag{17}$$

where Cov(d_i, D) is the covariance of the difference in expected *ES*s (d_i) and the difference in expected scores (D), and μ_{di} and μ_D are the means of d_{is} and D_s, respectively. In this case, DIF can be written as

$$\text{DIF}_i = Cov(d_i, D) + \mu_{d_i}\mu_D \ . \tag{18}$$

Raju et al. (1995) referred to this DIF as compensatory DIF (CDIF). If DIF in Equation 18 was expressed as CDIF, then Equation 17 can be rewritten as

$$\text{DTF} = \sum_{i=1}^{n} CDIF_i \ . \tag{19}$$

The additive nature of DTF allows for possible cancellation at the test level. This occurs when one item displays DIF in favor of one group and another item displays DIF in favor of the other group. This combination of DIF items will have a canceling effect on the overall DTF. The sum of the CDIF indices reflects the net directionality. For practical applications, a test developer could examine the DTF, then determine which item(s) should be eliminated based on its CDIF value and its overall contribution to DTF.

Raju et al. (1995) proposed a second index, noncompensatory DIF (NCDIF), that assumes that all items other than the one under study are free from differential functioning. In the dichotomous case, NCDIF is closely related to other existing DIF indices such as Lord's χ^2 and the unsigned area

(Raju et al., 1995). If all other items are DIF-free, then $d_j = 0$ for all $j \neq i$, where i is the item being studied, and Equation 18 can be rewritten as

$$NCDIF_i = \sigma_{d_i}^2 + \mu_{d_i}^2 \ . \tag{20}$$

Raju et al. (1995) noted that items having significant NCDIF do not necessarily have significant CDIF, in the sense of contributing significantly to DTF. For example, if one item favors the reference group and another favors the focal group, significant NCDIF occurs for both items even though the two CDIF indices may not be significant because of their canceling effect at the test level. This could lead to a greater number of significant NCDIF items than CDIF items.

In addition to cancellation at the test level, polytomously scored items allow for potential cancellation within an examinee at the item level. Recall that each item has multiple categories in the polytomous case, which leads to multiple probabilities. It is possible for one category to cancel the effects in another category when computing d_i for a given examinee. For example, if P_{1iF} is greater than P_{1iR} and P_{2iF} is less than the P_{2iR}, a cancellation will occur, keeping d_i close to 0, thereby indicating no differential functioning at the item level within an examinee.

DFIT Significance Tests

Assume that the D between expected scores is normally distributed with a mean of μ_D and a standard deviation of σ_D. A Z score for examinee s is

$$Z_s = \frac{D_s - \mu_D}{\sigma_D} \ , \tag{21}$$

where Z_s^2 has a χ^2 distribution with one degree of freedom (DF). The sum of Z_s^2 across N examinees has a χ^2 distribution with N DFs:

$$\chi_N^2 = \Sigma Z_s^2 = \frac{\Sigma (D_s - \mu_D)^2}{\sigma_D^2} \ . \tag{22}$$

The interest is in minimizing the expectation of DTF (i.e., E(DTF) $= \mu_D^2 = 0$), which implies that μ_D must be 0. Then, by substitution,

$$\chi_N^2 = \frac{\Sigma D_s^2}{\sigma_D^2} = \frac{N(DTF)}{\sigma_D^2} \ . \tag{23}$$

If an unbiased estimator is substituted for σ_D^2, then

$$\chi_{N-1}^2 = \frac{N(DTF)}{\hat{\sigma}_D^2} \ . \tag{24}$$

A significant χ^2 value indicates that one or more items are functioning differentially. Raju et al. (1995) suggested removing items that contribute significantly to DTF until the χ^2 value is no longer significant. According to Raju et al., deleted items are designated as having significant CDIF. Therefore, Raju et al. did not propose a separate significance test for CDIF.

Raju et al. (1995) defined a similar χ^2 test for NCDIF. This test was shown to be overly sensitive for large sample sizes (Fleer, 1993). Fleer suggested empirically establishing a critical value (cutoff) for NCDIF. This critical value was determined from a monte carlo study of non-DIF items.

Method

Data Simulation

A GRM with five response categories was used to generate the simulated datasets. Item parameters used in previous studies (Cohen & Kim, 1993; Fleer, 1993) were modified to accommodate the GRM. The modified item parameters are listed in Tables 1 and 2. DIF was modeled by adding a constant to the a and/or b parameters of the focal group.

Item probabilities for five categories per item for a simulated examinee were generated using Equation 1. Recall that five categories result in four probabilities per item. To assign a score for each simulated examinee, the following procedure was used. First, each simulated examinee was randomly assigned a θ from a standard normal distribution. Using the item parameters in Tables 1 and 2, along with the randomly assigned θ, resulted in four probabilities per item for each examinee. Then, for each simulated examinee a single random number (Y) was sampled from a uniform distribution over the interval [0,1]. If the randomly sampled number was less than the calculated probability at the boundary category k but greater than the calculated probability at $k+1$, then the score assigned was the value of category k. This can be expressed as

$$P^*_{ski} > Y_{si} > P^*_{s(k+1)i} \ , \tag{25}$$

where Y_{si} is the single random number for examinee s on item i.

Factors Manipulated

Two different θ distributions were simulated for the focal group. In the first condition, the focal and reference groups had equal θ distributions randomly selected from an N(0, 1) distribution. This condition is referred to as the *Equivalent* condition. In the second condition, the focal group was sampled from an N(−1, 1) distribution, resulting in lower θs than those in the reference group. This condition is referred to as the *Nonequivalent* condition.

Two test lengths, 20 and 40 items, were simulated. Sample size and scoring options were constant. For each group, 1,000 examinees were simulated. This sample size ensured adequate precision for parameter estimation prior to DIF/DTF analyses (Muraki & Bock, 1993). All items consisted of five scoring options (i.e., 0, 1, 2, 3, and 4). Simulation under each factor combination, referred to as a condition, was replicated five times.

Four proportions of test-wide DIF (0%, 5%, 10%, and 20%) and two conditions of direction of DIF (Unidirectional and Balanced-Bidirectional) were simulated. In the 20-item test, 0, 1, 2, 3, and 4 items were embedded with DIF. In the Unidirectional conditions, all items favored the reference group. In the Balanced-Bidirectional conditions, items favoring the reference group were balanced with items favoring the focal group. In the 5% condition, which had one DIF item, the Bidirectional condition could not be simulated. In addition, items were generated to simulate uniform DIF ($a_{i\mathrm{R}} = a_{i\mathrm{F}}$ and $b_{i\mathrm{R}} \neq b_{i\mathrm{F}}$) and nonuniform DIF ($a_{i\mathrm{R}} \neq a_{i\mathrm{F}}$, either with $b_{i\mathrm{R}} \neq b_{i\mathrm{F}}$ or $b_{i\mathrm{R}} = b_{i\mathrm{F}}$). Only the 20% DIF condition contained nonuniform DIF items. Two nonuniform DIF and two uniform DIF items were embedded in this condition.

Similar conditions were simulated in the 40-item test. DIF was embedded in 0, 2, 4, and 8 items. Directional and Balanced-Bidirectional DIF were simulated using the same method as in the 20-item test. Nonuniform DIF was embedded only in the 20% DIF condition. Figure 2 provides a visual display of the simulation design. Twenty-six conditions were simulated in this study.

It is difficult to judge the impact of embedding DIF by creating differences in the reference and focal group item parameters. A measure of the amount of DIF that has been embedded is reported using the CDIF and NCDIF values based on the true parameters and a standard normal θ distribution

Table 1
Reference Group Item Parameters

20-Item Test						40-Item Test					
Item	a	b_1	b_2	b_3	b_4	Item	a	b_1	b_2	b_3	b_4
1	.55	−1.80	−.60	.60	1.80	1	.55	−1.80	−.60	.60	1.80
2	.73	−2.32	−1.12	.08	1.28	2	.55	−1.80	−.60	.60	1.80
3	.73	−1.80	−.60	.60	1.80	3	.73	−2.32	−1.12	.08	1.28
4[a]	.73	−1.80	−.60	.60	1.80	4	.73	−2.32	−1.12	.08	1.28
4[b]	.73	−1.30	−.10	1.10	2.30	5[a]	.73	−1.80	−.60	.60	1.80
4[c]	1.23	−1.80	−.60	.60	1.80	5[d]	.73	−2.30	−1.10	.10	1.30
5[a]	.73	−1.28	−.08	1.12	2.32	5[e]	.73	−1.80	−.60	.60	1.80
5[b]	.73	−1.80	−.60	.60	1.80	6[a]	.73	−1.80	−.60	.60	1.80
5[c]	.73	−1.80	−.60	.60	1.80	6[d]	.73	−1.30	−.10	1.10	2.30
6[a]	1.00	−2.78	−1.58	−.38	.82	6[e]	1.23	−1.80	−.60	.60	1.80
6[b]	.73	−1.28	−.08	1.12	2.32	7	.73	−1.80	−.60	.60	1.80
6[c]	.73	−1.28	−.08	1.12	2.32	8	.73	−1.80	−.60	.60	1.80
7[a]	1.00	−2.32	−1.12	.08	1.28	9	.73	−1.28	−.08	1.12	2.32
7[b]	1.00	−2.78	−1.58	−.38	.82	10	.73	−1.28	−.08	1.12	2.32
7[c]	1.00	−2.78	−1.58	−.38	.82	11	1.00	−2.78	−1.58	−.38	.82
8	1.00	−2.32	−1.12	.08	1.28	12	1.00	−2.78	−1.58	−.38	.82
9[a]	1.00	−1.80	−.60	.60	1.80	13	1.00	−2.32	−1.12	.08	1.28
9[b]	1.00	−2.32	−1.12	.08	1.28	14	1.00	−2.32	−1.12	.08	1.28
9[c]	1.00	−2.32	−1.12	.08	1.28	15[a]	1.00	−2.32	−1.12	.08	1.28
10[a]	1.00	−1.80	−.60	.60	1.80	15[d]	1.00	−2.57	−1.37	−.17	1.03
10[b]	1.00	−2.07	−.87	.33	1.53	15[e]	1.00	−2.32	−1.12	.08	1.28
10[c]	1.00	−1.80	−.60	.60	1.80	16[a]	1.00	−2.32	−1.12	.08	1.28
11	1.00	−1.80	−.60	.60	1.80	16[d]	1.00	−2.07	−.87	.33	1.53
12[a]	1.00	−1.80	−.60	.60	1.80	16[e]	.50	−2.32	−1.12	.08	1.28
12[b]	1.00	−1.80	−.60	.60	1.80	17	1.00	−1.80	−.60	.60	1.80
12[c]	1.00	−1.28	−.08	1.12	2.32	18	1.00	−1.80	−.60	.60	1.80
13[a]	1.00	−1.28	−.08	1.12	2.32	19	1.00	−1.80	−.60	.60	1.80
13[b]	1.00	−1.80	−.60	.60	1.80	20	1.00	−1.80	−.60	.60	1.80
13[c]	1.00	−.78	.42	1.62	2.82	21	1.00	−1.80	−.60	.60	1.80
14	1.00	−1.28	−.08	1.12	2.32	22	1.00	−1.80	−.60	.60	1.80
15[a]	1.00	−.82	.38	1.58	2.78	23	1.00	−1.80	−.60	.60	1.80
15[b]	1.00	−1.28	−.08	1.12	2.32	24	1.00	−1.80	−.60	.60	1.80
15[c]	1.00	−.82	.38	1.58	2.78	25[a]	1.00	−1.28	−.08	1.12	2.32
16[a]	1.36	−2.32	−1.12	.08	1.28	25[d]	1.00	−1.28	−.08	1.12	2.32
16[b]	1.00	−.82	.38	1.58	2.78	25[e]	1.00	−1.78	−.58	.62	1.82
16[c]	1.00	−.32	.88	2.08	3.28	26[a]	1.00	−1.28	−.08	1.12	2.32
17[a]	1.36	−1.80	−.60	.60	1.80	26[d]	1.00	−1.28	−.08	1.12	2.32
17[b]	1.36	−2.32	−1.12	.08	1.28	26[e]	1.00	−.78	.42	1.62	2.82
17[c]	1.36	−2.32	−1.12	.08	1.28	27	1.00	−1.28	−.08	1.12	2.32
18	1.36	−1.80	−.60	.60	1.80	28	1.00	−1.28	−.08	1.12	2.32
19	1.36	−1.28	−.08	1.12	2.32	29	1.00	−.82	.38	1.58	2.78
20	1.80	−1.80	−.60	.60	1.80	30[a]	1.00	−.82	.38	1.58	2.78

continued on next page

of 1,000 examinees. This provides an indication of how "large" or "small" DIF is in the embedded items. Table 3 shows the differences in the item parameters between the reference and focal groups, as well as the true CDIF and NCDIF values.

Table 1, continued
Reference Group Item Parameters

20-Item Test						40-Item Test					
Item	a	b_1	b_2	b_3	b_4	Item	a	b_1	b_2	b_3	b_4
						30[d]	1.00	−.82	.38	1.58	2.78
						30[e]	1.00	−.32	.88	2.08	3.28
						31	1.36	−2.32	−1.12	.08	1.28
						32	1.36	−2.32	−1.12	.08	1.28
						33	1.36	−1.80	−.60	.60	1.80
						34	1.36	−1.80	−.60	.60	1.80
						35	1.36	−1.80	−.60	.60	1.80
						36	1.36	−1.80	−.60	.60	1.80
						37	1.36	−1.28	−.08	1.12	2.32
						38	1.36	−1.28	−.08	1.12	2.32
						39	1.80	−1.80	−.60	.60	1.80
						40	1.80	−1.80	−.60	.60	1.80

[a]Item parameters used in Conditions 1, 2, and 3.
[b]Item parameters used in Condition 4.
[c]Item parameters used in Condition 5.
[d]Item parameters used in Conditions 4 and 5.
[e]Item parameters used in Condition 6.

Parameter Estimation and Linking

Item and θ parameters were estimated using PARSCALE 2 (Muraki & Bock, 1993). The maximum marginal likelihood procedure and EM algorithm were used to estimate the item parameters. Default values were used for all estimation. Estimated a posteriori Bayesian procedures with normal priors were used to estimate θ.

The estimation of linking coefficients was based on Baker's modified test characteristic curve method as implemented in EQUATE 2.0 (Baker, 1993). All parameter estimates for the reference group in this study were equated to the underlying metric of the focal group.

Several researchers have shown that an iterative linking procedure improves identification of DIF items (e.g., Candell & Drasgow, 1988; Drasgow, 1987; Lautenschlager & Park, 1988; Lord, 1980; Miller & Oshima, 1992). To minimize error introduced by the equating procedure, a two-stage linking procedure was used in this study. After the initial linking with all test items, a DIF analysis was performed. If items were identified as displaying DIF, as indicated by an NCDIF index that exceeded the critical value, the linking procedure was performed again without these DIF items. Finally, all items were transformed using the linking coefficients obtained in the second iteration. A FORTRAN program written by Raju (1995) was used to calculate the DFIT indices.

Establishing Critical Values

Because the χ^2 for NCDIF was found to be overly sensitive for large sample sizes, an empirical critical value was established for all DIF indices to protect against Type I error. Two thousand DIF-free items were simulated and DIF analyses were conducted. An alternative cutoff was established by finding the value at the 99th percentile. This resulted in an alternative cutoff value of .016. This value was used for both DIF and DTF items.

Table 2
Focal Group Item Parameters, by Condition (Items Not Listed Used the Same Item Parameters as the Reference Group)

20-Item Test						40-Item Test					
Item	a	b_1	b_2	b_3	b_4	Item	a	b_1	b_2	b_3	b_4
Condition 1											
3	.73	−.80	.40	1.60	2.80	5	.73	−.80	.40	1.60	2.80
10	.73	−.28	.92	2.12	3.32						
Condition 2											
3	.73	−1.30	−.10	1.10	2.30	5	.73	−.80	.40	1.60	2.80
8	1.00	−1.32	−.12	1.08	2.28	10	.73	−.78	.42	1.62	2.82
15	1.00	−1.32	−.12	1.08	2.28						
20	1.00	−1.30	−.10	1.10	2.30						
Condition 3											
3	.73	−.80	.40	1.60	2.80	5	.73	−.80	.40	1.60	2.80
8	.50	−1.82	−.62	.58	1.78	10	.73	−.78	.42	1.62	2.82
13	1.00	−.78	.42	1.62	2.82	15	.50	−1.82	−.62	.58	1.78
18	.86	−1.80	−.60	.60	1.80	20	.50	−1.80	−.60	.60	1.80
25	1.00	−.28	.92	2.12	3.32						
30	1.00	−.32	.88	2.08	3.28						
35	.86	−1.30	−.10	1.10	2.30						
40	1.30	−1.80	−.60	.60	1.80						
Condition 4											
3	.73	−1.30	−.10	1.10	2.30	5	.73	−1.30	−.10	1.10	2.30
4	.73	−1.80	−.60	.60	1.80	6	.73	−2.30	−1.10	.10	1.30
Condition 5											
3	1.23	−1.80	−.60	.60	1.80	5	.73	−1.30	−.10	1.10	2.30
4	.73	−1.80	−.60	.60	1.80	6	.73	−2.30	−1.10	.10	1.30
12	1.00	−.78	.42	1.62	2.82	15	1.00	−2.07	−.87	.33	1.53
13	1.00	−1.28	−.08	1.12	2.32	16	1.00	−2.57	−1.37	−.17	1.03
Condition 6											
5	1.23	−1.80	−.60	.60	1.80						
6	.73	−1.80	−.60	.60	1.80						
15	.50	−2.32	−1.12	.08	1.28						
16	1.00	−2.32	−1.12	.08	1.28						
25	1.00	−.78	.42	1.62	2.82						
26	1.00	−1.78	−.58	.62	1.82						
29	1.00	−.32	.88	2.08	3.28						
30	1.00	−.82	.38	1.58	2.78						

Detection of DIF

Two indicators were calculated to determine the accuracy of DIF detection, true positive (TP) and false positive (FP). A TP was an embedded DIF item with a DIF index value that exceeded the cutoff value. An FP was a non-DIF item with a DIF index value that exceeded the criterion established for DIF. TP rates were determined by tallying the total number of detected embedded DIF items across the five replications and dividing by the total number of embedded DIF items across the five replications. FP rates were determined from the total number of erroneously identified non-DIF items across the five replications, divided by the total number of non-DIF items across the five replications.

For comparison, a significance test was done using the true item parameters. These analyses bypassed the PARSCALE estimation and linking procedures and are referred to as *True* conditions. True conditions consisted of one analysis per condition, as opposed to the Estimated conditions that consisted of five replications per condition.

Figure 2
Simulation Design

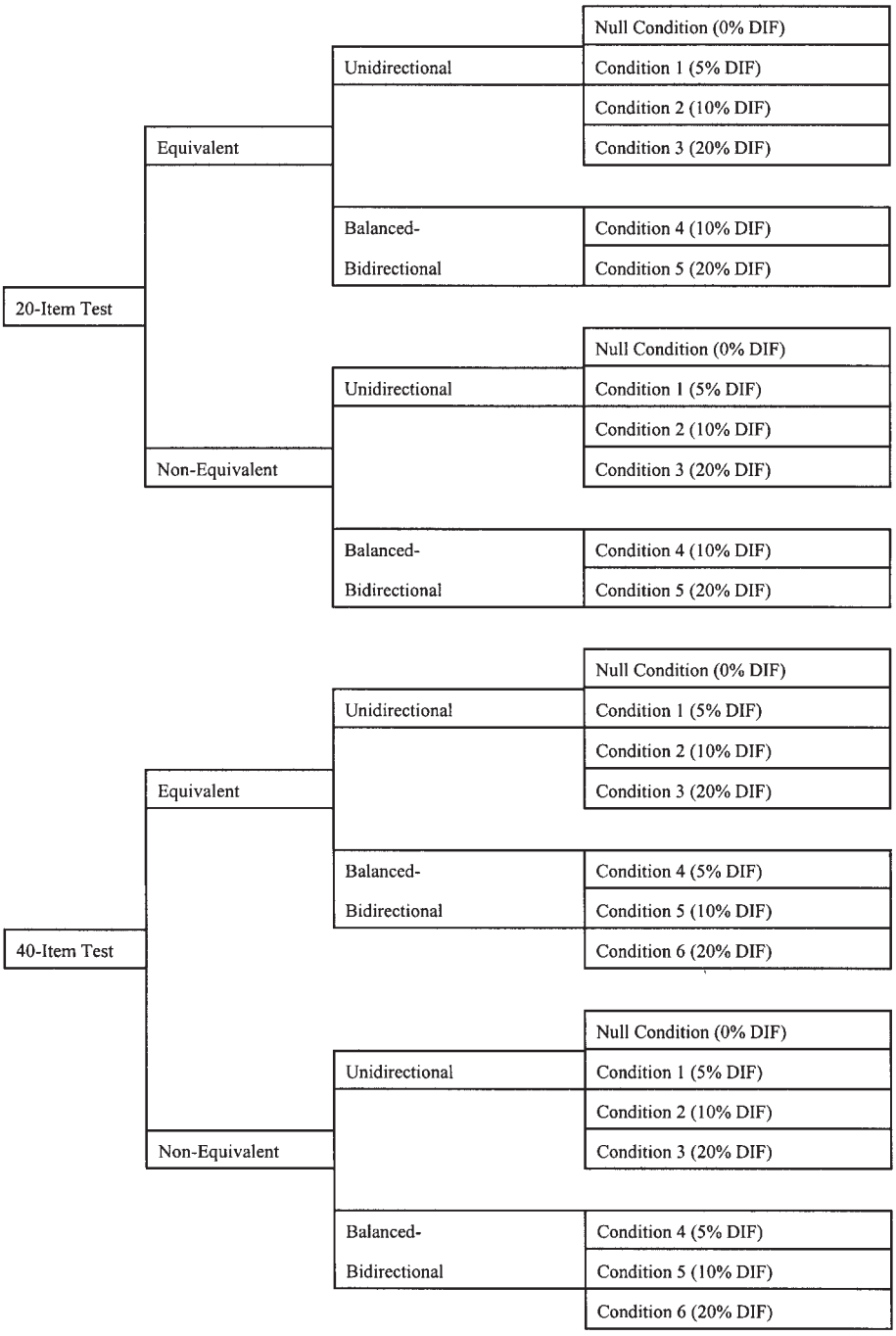


Table 3
Difference Between Focal and Reference Group Item Parameters (Focal Group Minus Reference Group) for True CDIF and NCDIF Values for a 20-Item Test and a 40-Item Test

20-Item Test					40-Item Test				
	Difference		True	True		Difference		True	True
Item	*a*	*b*	CDIF	NCDIF	Item	*a*	*b*	CDIF	NCDIF
Unidirectional Conditions									
Condition 1									
3	—	+1.0	.48	.48	5	—	+1.0	.92	.48
					10	—	+1.0	.87	.43
Condition 2									
3	—	+.5	.39	.12	5	—	+1.0	1.49	.48
8	—	+1.0	.84	.58	10	—	+.5	.73	.12
					15	—	+1.0	1.64	.57
					20	—	+.5	.82	.14
Condition 3									
3	—	+1.0	1.00	.48	5	—	+1.0	1.97	.48
8	−.5	+.5	.59	.17	10	—	+.5	.97	.12
13	—	+.5	.53	.13	15	−.5	+.5	1.17	.17
18	−.5	—	.02	0.00	20	−.5	—	.11	.03
					25	—	+.5	2.04	.50
					30	—	+.5	.99	.12
					35	−.5	+.5	1.08	.14
					40	−.5	—	.02	0.00
Balanced-Bidirectional Conditions									
Condition 4									
3	—	+.5	0.00	.12	5	—	+1.0	0.00	.49
4	—	−.5	0.00	.12	6	—	−1.0	0.00	.49
Condition 5									
3	+.5	—	0.00	.01	5	—	+1.0	0.00	.49
4	−.5	—	0.00	.01	6	—	−1.0	0.00	.49
12	—	+.5	0.00	.13	15	—	+.5	0.00	.14
13	—	−.5	0.00	.13	16	—	−.5	0.00	.14
Condition 6									
					5	+.5	—	0.00	.01
					6	−.5	—	0.00	.01
					15	−.5	—	0.00	.03
					16	+.5	—	0.00	.03
					25	—	+1.0	0.00	.56
					26	—	−1.0	0.00	.56
					29	—	+.5	0.00	.12
					30	—	−.5	0.00	.12

Results

CDIF

Because CDIFs sum to DTF, when a given DTF was found statistically significant (at $p < .01$), items with large and positive CDIF indices were removed one at a time until the DTF index based on the remaining items was statistically nonsignificant. Items that were removed were classified as having significant CDIF. The Balanced-Bidirectional tests should not have any items identified as DIF because of CDIF cancellation. Therefore, TPs were relevant only in the 20- and 40-item Unidirectional conditions (Conditions 1, 2, and 3; see Figure 2). Tables 4 and 5 show the aggregated results at both the condition level and the item level, respectively, for CDIF analyses.

CDIF True conditions. For the 20-item conditions, all items with significant CDIF were identified, except in Condition 3. In Condition 3, .75 of the true CDIF items were detected (Table 4). Item level results (Table 5) indicated that Item 18, an item with a very small amount of DIF, was not detected. No FPs were detected in any of the conditions.

Table 4
True Positive (TP) and False Positive (FP) Rates for CDIF by Condition

Test and Condition	No. of DIF Items	Equivalent				Nonequivalent			
		True CDIF		Est. CDIF		True CDIF		Est. CDIF	
		TP	FP	TP	FP	TP	FP	TP	FP
20-Item Test									
Null Condition	0	—	0.00	—	0.00	—	0.00	—	0.00
Unidirectional									
Condition 1	1	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00
Condition 2	2	1.00	0.00	.90	.03	1.00	0.00	.90	.03
Condition 3	4	.75	0.00	.65	.18	.75	0.00	.65	.03
Balanced-Bidirectional									
Condition 4	2	—	0.00	—	0.00	—	0.00	—	.01
Condition 5	4	—	0.00	—	.02	—	0.00	—	.01
40-Item Test									
Null Condition	0	—	0.00	—	0.00	—	0.00	—	0.00
Unidirectional									
Condition 1	2	1.00	0.00	1.00	.01	1.00	0.00	1.00	.02
Condition 2	4	1.00	0.00	.80	.01	1.00	0.00	.50	.02
Condition 3	8	.75	0.00	.68	.01	.75	0.00	.68	.01
Balanced-Bidirectional									
Condition 4	2	—	0.00	—	.01	—	0.00	—	.01
Condition 5	4	—	0.00	—	0.00	—	0.00	—	.01
Condition 6	8	—	0.00	—	.03	—	0.00	—	.03

Similar results were obtained in the 40-item conditions. Again, all significant CDIF items were identified, except in Condition 3. Items with a small amount of DIF were not detected (Items 20 and 40). No FPs were observed.

CDIF Estimated conditions. In the Estimated 20-item/Equivalent conditions, there was a decrease for the TPs in Conditions 2 and 3 as compared to the True conditions. In Condition 2, the TP rate decreased from 1.00 to .90. In Condition 3, the TP rate dropped from .75 to .65 (Table 4). Additionally, the FP rates increased in Conditions 2 and 3. In Condition 2, the FP rate increased slightly from 0.0 to .03. In Condition 3, the FP rate had a much larger increase from 0.0 to .18. This was due to two repetitions within this condition that identified four and six non-DIF items. The remaining three repetitions identified zero or one FP items.

For the 20-item/Nonequivalent conditions, the results were identical to the 20-item/Equivalent conditions, except for the FP rate in Condition 3. A lower FP rate (.03) was detected in the Nonequivalent condition, compared to the Equivalent condition (.18).

A similar trend was observed in the 40-item conditions. In the 40-item/Equivalent conditions, the TP rates decreased in both Conditions 2 and 3. The TP rate decreased from 1.00 to .80 and from .75 to .68 for Conditions 2 and 3, respectively. The item-level analyses (Table 5) revealed that items with a small amount of DIF were not detected. The FP rates increased slightly in almost all conditions (ranging from 0.0 to .03).

The 40-item/Nonequivalent conditions had similar results to the 40-item/Equivalent conditions, except for two instances. In Condition 2, the TP rate decreased from .80 to .50. Because of the

Table 5
True Positive Rates for CDIF at the Item Level

Test Condition, and Item	True CDIF Value	Equivalent True CDIF	Equivalent Est. CDIF	Nonequivalent True CDIF	Nonequivalent Est. CDIF
20-Item Test					
Unidirectional Conditions					
Condition 1					
3	.48	1.0	1.0	1.0	1.0
Condition 2					
3	.39	1.0	.8	1.0	.8
8	.84	1.0	1.0	1.0	1.0
Condition 3					
3	1.00	1.0	.8	1.0	1.0
8	.59	1.0	.8	1.0	.6
13	.53	1.0	.8	1.0	1.0
18	.02	0.0	.2	0.0	0.0
40-Item Test					
Unidirectional Conditions					
Condition 1					
5	.92	1.0	1.0	1.0	1.0
10	.87	1.0	1.0	1.0	1.0
Condition 2					
5	1.49	1.0	.8	1.0	.4
10	.73	1.0	.6	1.0	.4
15	1.64	1.0	1.0	1.0	.8
20	.82	1.0	.8	1.0	.4
Condition 3					
5	1.97	1.0	1.0	1.0	1.0
10	.97	1.0	1.0	1.0	1.0
15	1.17	1.0	.8	1.0	.2
20	.11	0.0	0.0	0.0	0.0
25	2.04	1.0	1.0	1.0	1.0
30	.99	1.0	.8	1.0	1.0
35	1.08	1.0	1.0	1.0	1.0
40	.02	0.0	0.0	0.0	0.0

substantial decrease in detection rate, an additional five repetitions were simulated. The results of the additional repetitions were similar to the finding in the 40-item/Equivalent condition. For the additional repetitions in this condition, the TP rate was .80 and the FP rate was .03.

NCDIF

NCDIF True conditions. Tables 6 and 7 contain the results of the TPs and FPs for NCDIF. In the True 20-item conditions, the TP rates were 1.0, except for Conditions 3 and 5, which had a TP rate of .75 and .50, respectively. Analyses at the item level revealed that the DIF items not detected were Item 18 (Condition 3) and Items 3 and 4 (Condition 5). These items had a small amount of DIF. No FP items were detected.

For the True 40-item conditions, all conditions had perfect TP detection rates except Conditions 3 and 6. In Condition 3 the TP detection rate was .88 (Item 40 not detected); in Condition 6, the TP rate was .75 (Items 5 and 6 not detected). Again, these items had the smallest amount of DIF. No FPs were detected.

Table 6
The True Positive (TP) and False Positive (FP) Rates for NCDIF by Condition

Test and Condition	No. of DIF Items	Equivalent				Nonequivalent			
		True NCDIF		Estimated NCDIF		True NCDIF		Estimated NCDIF	
		TP	FP	TP	FP	TP	FP	TP	FP
20-Item Test									
Null Condition	0	—	0.00	—	0.00	—	0.00	—	0.00
Unidirectional									
Condition 1	1	1.00	0.00	1.00	.01	1.00	0.00	1.00	.01
Condition 2	2	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00
Condition 3	4	.75	0.00	.75	0.00	.75	0.00	.80	0.00
Balanced-Bidirectional									
Condition 4	2	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00
Condition 5	4	.50	0.00	.50	0.00	.50	0.00	.55	0.00
40-Item Test									
Null Condition	0	—	0.00	—	0.00	—	0.00	—	0.00
Unidirectional									
Condition 1	2	1.00	0.00	1.00	0.00	1.00	0.00	1.00	.01
Condition 2	4	1.00	0.00	1.00	0.00	1.00	0.00	1.00	.01
Condition 3	8	.88	0.00	.88	.01	.88	0.00	.88	0.00
Balanced-Bidirectional									
Condition 4	2	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00
Condition 5	4	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00
Condition 6	8	.75	0.00	.70	.01	.75	0.00	.80	0.00

NCDIF Estimated conditions. The results of the Estimated conditions were similar to the True conditions. In the 20-item/Equivalent conditions, the results were identical to the True conditions except in Condition 1 in which the FP rate slightly increased from 0.0 to .01. In the 20-item/Nonequivalent case, Conditions 3 and 5 showed a slight increase in the TP rates, from .75 to .80 and from .50 to .55, respectively.

In the 40-item/Equivalent condition, the Estimated conditions were similar to the True conditions. There was a slight decrease in TP detection rate in Condition 6, from .75 to .70. There was also a slight increase in FP rates in Conditions 3 and 6, from 0.0 to .01.

For the 40-item/Nonequivalent case, the results were identical to the True condition except in Condition 6, in which the TP detection rate increased from .75 to .80. Additionally, the FP rates in Conditions 1 and 2 increased slightly, from 0.0 to .01 for both conditions.

Conclusions

The DFIT framework was effective in identifying DTF and DIF in polytomously scored data for the conditions simulated. Test length (20 and 40 items), focal group distribution (equivalent and nonequivalent), number of DIF items (0%, 5%, 10%, and 20%), and direction of DIF (unidirectional and balanced-bidirectional) had little effect on the true positive and false positive detection rates across all conditions. As expected, items with large amounts of DIF were detected, and items with small amounts of DIF were not detected.

Overall, CDIF was not as stable as NCDIF. This finding is similar to the findings for the unidimensional case (Fleer, 1993) and the multidimensional-dichotomous cases (Oshima et al., 1997). In the present study, CDIF had two conditions that varied from what was expected. For the 20-item/Equivalent condition, CDIF erroneously identified 18% of the non-DIF items as DIF. For the 40-item/Nonequivalent condition, CDIF identified only 50% of the DIF items. When additional

Table 7
True Positive NCDIF Rates at the Item Level

Test Condition, and Item	True NCDIF Value	Equivalent		Nonequivalent	
		True NCDIF	Est. NCDIF	True NCDIF	Est. NCDIF
20-Item Test					
Unidirectional Conditions					
Condition 1					
3	.48	1.0	1.0	1.0	1.0
Condition 2					
3	.12	1.0	1.0	1.0	1.0
8	.58	1.0	1.0	1.0	1.0
Condition 3					
3	.48	1.0	1.0	1.0	1.0
8	.17	1.0	1.0	1.0	1.0
13	.13	1.0	1.0	1.0	1.0
18	0.00	0.0	0.0	0.0	.2
Balanced-Bidirectional Conditions					
Condition 4					
3	.12	1.0	1.0	1.0	1.0
4	.12	1.0	1.0	1.0	1.0
Condition 5					
3	.01	0.0	0.0	0.0	0.0
4	.01	0.0	0.0	0.0	0.0
12	.13	1.0	1.0	1.0	1.0
13	.13	1.0	1.0	1.0	1.0
40-Item Test					
Unidirectional Conditions					
Condition 1					
5	.48	1.0	1.0	1.0	1.0
10	.43	1.0	1.0	1.0	1.0
Condition 2					
5	.48	1.0	1.0	1.0	1.0
10	.12	1.0	1.0	1.0	1.0
15	.57	1.0	1.0	1.0	1.0
20	.14	1.0	1.0	1.0	1.0
Condition 3					
5	.48	1.0	1.0	1.0	1.0
10	.12	1.0	1.0	1.0	1.0
15	.17	1.0	1.0	1.0	1.0
20	.03	1.0	1.0	1.0	1.0
25	.50	1.0	1.0	1.0	1.0

continued on next page

simulations were performed, the results were consistent with theoretical expectations. A possible explanation for the occasional erratic detection rate is that the estimation and linking errors associated with the Estimated conditions accumulated across the entire test. The calculation of DTF involves summing the CDIF values across the entire test, which includes all the errors related to each item. For example, a linking error would magnify the error in the same direction throughout the test. If the linking additive component was overestimated by .2, then .2 would be added to each item. NCDIF, which had stable results across all conditions, is calculated from information related to only one item; consequently, this led to more stable results.

Table 7, continued
True Positive NCDIF Rates at the Item Level

Test Condition, and Item	True NCDIF Value	Equivalent		Nonequivalent	
		True NCDIF	Est. NCDIF	True NCDIF	Est. NCDIF
Balanced-Bidirectional Conditions					
Condition 4					
5	.49	1.0	1.0	1.0	1.0
6	.49	1.0	1.0	1.0	1.0
Condition 5					
5	.49	1.0	1.0	1.0	1.0
6	.49	1.0	1.0	1.0	1.0
15	.14	1.0	1.0	1.0	1.0
16	.14	1.0	1.0	1.0	1.0
Condition 6					
5	.01	0.0	0.0	0.0	.2
6	.01	0.0	0.0	0.0	.2
15	.03	1.0	1.0	1.0	1.0
16	.03	1.0	1.0	1.0	1.0
25	.56	1.0	1.0	1.0	1.0
26	.56	1.0	1.0	1.0	1.0
29	.12	1.0	1.0	1.0	1.0
30	.12	1.0	1.0	1.0	1.0

Limitations

Although this study supports the validity of the polytomous-DFIT framework, the results are specific to the conditions simulated. In this study, the method in which DIF was embedded (i.e., placing differences in each category) might be unrealistic and might provide optimal conditions for detecting DIF. This high detection rate created a ceiling effect that limited the investigation of the influence of factors that were manipulated in this study. θ group distribution and values of the a and b parameters should have an influence in the detection of DIF/DTF. The efficacy of the DFIT framework should be examined in more conditions with other IRT models.

Future Research

The findings of this study encourage future research areas for DFIT. First, critical (cutoff) values for CDIF and NCDIF should be investigated. In this study, the critical value was established by using an empirical method that was optimal for the detection of DIF/DTF specific to this study. A Type I and Type II error simulation study should be performed. For DFIT to be of practical use, critical values at various α levels with different IRT models should be established.

The reason for the occasional instability of CDIF needs to be determined. CDIF offers a unique method for assessing the overall effect of removing or adding an item to a test. Finally, many conditions need to be experimentally manipulated. Sample size, amount of DIF, length of test, distribution of focal group, and many other conditions need to be systematically investigated. Additionally, the DFIT framework should be applied to tests with mixed item formats (i.e., dichotomous and polytomous items). These systematic investigations would help establish guidelines and limitations of the DFIT procedure.

References

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 43,* 561–573.

Baker, F. B. (1993). *EQUATE2: Computer program for equating two metrics in item response theory* [Computer program]. Madison: University of Wisconsin, Laboratory of Experimental Design.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika, 37,* 29–51.

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.

Chang, H., & Mazzeo, J. (1994). The unique correspondence of the item response function and item category response functions in polytomously scored item response models. *Psychometrika, 59,* 391–404.

Cohen, A. S., & Kim, S. H. (1993). A comparison of Lord's χ^2 and Raju's area measures on detection of DIF. *Applied Psychological Measurement, 17,* 39–52.

Drasgow, F. (1987). Study of the measurement of bias of two standardized psychological tests. *Journal of Applied Psychology, 72,* 19–29.

Dunbar, S. B., Koretz, D. M., & Hoover, H. D. (1991). Quality control in the development and use of performance assessments. *Applied Measurement in Education, 4,* 289–303.

Fleer, P. F. (1993). A monte carlo assessment of a new measure of item and test bias (Doctoral dissertation, Illinois Institute of Technology, 1993) *Dissertation Abstracts International, 54-04B,* 2266.

Lautenschlager, G. J., & Park, D. G. (1988). IRT item bias detection procedures: Issues of model misspecification, robustness, and parameter linking. *Applied Psychological Measurement, 12,* 365–376.

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

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47,* 149–174.

Miller, M. D., & Oshima, T. C. (1992). Effect of sample size, number of biased items, and magnitude of bias on a two-stage item bias estimation method. *Applied Psychological Measurement, 16,* 381–388.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16,* 159–176.

Muraki, E., & Bock, R.D. (1993). *PARSCALE2: IRT based test scoring and item analysis for graded open-ended exercises and performance tasks* [Computer program]. Chicago: Scientific Software International.

Oshima, T. C., Raju, N., & 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.

Raju, N. (1995). *DFITPU: A FORTRAN program for calculating DIF/DTF* [Computer program]. Atlanta: Georgia Institute of Technology.

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

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph, No. 17.*

Samejima, F. (1972). A general model for free-response data. *Psychometrika Monograph, No. 18.*

Zwick, R., Donoghue, J. R., & Grima, A. (1993). Assessment of differential item functioning for performance tasks. *Journal of Educational Measurement, 30,* 233–251.

Acknowledgments

The authors thank two anonymous reviewers for their many valuable comments.

Author's Address

Send requests for reprints or further information to Claudia Flowers, Department of Educational Administration, Research, and Technology, University of North Carolina, Charlotte NC 28223, U.S.A. Email: cpflower@email.uncc.edu.