*Journal of Educational Measurement*
*Spring 2006, Vol. 43, No. 1, pp. 1–17*

# A New Method for Assessing the Statistical Significance in the Differential Functioning of Items and Tests (DFIT) Framework

**T. C. Oshima**
*Georgia State University*
**Nambury S. Raju**
*Illinois Institute of Technology*
**Alice O. Nanda**
*Georgia State University*

*A new item parameter replication method is proposed for assessing the statistical significance of the noncompensatory differential item functioning (NCDIF) index associated with the differential functioning of items and tests framework. In this new method, a cutoff score for each item is determined by obtaining a $(1 - \alpha)$ percentile rank score from a frequency distribution of NCDIF values under the no-DIF condition by generating a large number of item parameters based on the item parameter estimates and their variance-covariance structures from a computer program such as BIILOG-MG3. This cutoff for each item can be used as the basis for determining whether a given NCDIF index is significantly different from zero. This new method has definite advantages over the current method and yields cutoff values that are tailored to a particular data set and a particular item. A Monte Carlo assessment of this new method is presented and discussed.*

Currently, there are several procedures for assessing differential item functioning (DIF). Some of these procedures are based on item response theory (IRT), while others are non-IRT based. Examples of IRT-based procedures include Lord's $\chi^2$ (Cohen, Kim, & Baker, 1993; Lord, 1980), the likelihood ratio test (Thissen, Steinberg, & Wainer, 1988), area measures (Cohen et al., 1993; Kim & Cohen, 1991; Raju, 1988, 1990), Muraki's methods for polytomous items (Muraki, 1999), and the methods based on the differential functioning of items and tests (DFIT) framework (Flowers, Oshima, & Raju, 1999; Oshima, Raju, & Flowers, 1997; Raju, van der Linden, & Fleer, 1995).

The DFIT framework has several advantages over other methods. First, it can be applied to multidimensional as well as unidimensional data with either dichotomous or polytomous scoring. Second, it can be applied at the test level to identify differential test functioning (DTF) as well as at the item level to identify DIF. Third, it offers two kinds of DIF indices, noncompensatory differential item functioning (NCDIF) and compensatory differential item functioning (CDIF). NCDIF assumes that all other items in the test except the studied item contain no DIF. NCDIF appears

This article is dedicated to our dearest colleague, Dr. Nam Raju, who loved his work in psychometrics and never slowed down until the day he passed away on October 27, 2005. His dedication to the field as well as his most respectable character will always be remembered.

to be similar to most of the other IRT-based DIF indices because other indices also hold the same assumption. On the other hand, CDIF is rather unique since CDIF values add up to the total DTF, enabling the practitioners to examine the net effect of deleting one or more items from the test.

Although the DFIT framework has shown to be an effective mechanism for detecting DIF and DTF in IRT-based tests/questionnaires in several studies (e.g., Flowers et al., 1999; Oshima et al., 1997; Raju et al., 1995), these studies also have pointed out a need for better procedures for assessing the statistical significance of the DIF and DTF indices. Although significance tests based on the $\chi^2$ distribution have been introduced by Raju et al. (1995), researchers pointed out that the $\chi^2$ tests for DTF and NCDIF were overly sensitive for large sample sizes. They then recommended a cutoff value of .006 for dichotomously scored items based on Fleer's (1993) Monte Carlo study where the cutoff value of .006 resulted in falsely identifying approximately 1% of the items as having significant DIF under the no-DIF condition. In a similar fashion, Bolt (2002) and Flowers et al. (1999) generated their own cutoffs for examining the Type I error rates and power rates of several DIF procedures (including the DFIT framework) in the polytomous case. These cutoffs for assessing dichotomous and polytomous DIF appeared to have worked well in the Monte Carlo investigations by Bolt, Flowers et al., and Raju et al., but they are probably not generalizable to other dichotomous and polytomous (with the same or different numbers of response categories) items and sample sizes. In fact, in the dichotomous case, Chamblee (1998) has shown that factors such as the sample size and the type of IRT model can influence the cutoff value. In her Monte Carlo study, she obtained the empirical distributions of NCDIF indices under the no-DIF condition varying the sample size and the type of IRT model (one-parameter (1PL), two-parameter (2PL), and three-parameter (3PL) IRT models). The cutoff values ranged from .003 to .018 with a higher value for a smaller sample size and a higher value for an IRT model with more parameters. This is probably true for the polytomous case as well.

The $\chi^2$ test proposed by Raju et al. (1995) does not appear to be all that useful in identifying significant DIF with acceptable type I error rates and power rates. It is possible in Monte Carlo investigations to come up with cutoff values with desirable statistical characteristics, as Bolt (2002), Flowers et al. (1999), and Raju et al. (1995) did, but it is not a technique that is readily accessible to typical practitioners. Most practitioners would neither have the technical knowledge nor the time to develop cutoffs for use with their particular data sets. Therefore, one of the purposes of this investigation is to describe a procedure for deriving study-based cutoffs for use by practitioners in assessing DIF (within the DFIT framework) in dichotomously scored items. The other purpose is to illustrate this new procedure and to evaluate its efficacy in a Monte Carlo investigation. The CDIF, NCDIF, and DTF indices in the DFIT framework are briefly described next, followed by a detailed presentation of the new significance test for the NCDIF index.

## Indices in the DFIT Framework

According to Raju et al. (1995), the NCDIF index for an item $i$ is defined as

$$\text{NCDIF}_i = E_{\text{F}}[P_{i\text{F}}(\theta) - P_{i\text{R}}(\theta)]^2 = E_{\text{F}}\left(d_i^2\right) = \sigma_{d_i}^2 + \mu_{d_i}^2, \tag{1}$$

where $P_{i\mathrm{F}}(\theta)$ is the probability of a correct response at a given $\theta$ using the item parameter estimates from the focal group and $P_{i\mathrm{R}}(\theta)$ is the probability of a correct response at the same given $\theta$ using the item parameter estimates from the reference group. The expectation ($E$) is taken over the focal group, and $\sigma$ and $\mu$ refer to the standard deviation and mean, respectively. The $d_i$ in the above equation is the difference in probability scores on item $i$ for the same examinee, first treated as a member of the focal group and then treated as a member of the reference group. At the test or scale level (with $n$ items), a similar difference in true scores ($D$) for an examinee may be expressed as

$$D = \sum_{i=1}^{n} d_i. \tag{2}$$

Using this difference at the test/scale level, Raju et al. define DTF as

$$\mathrm{DTF} = E_{\mathrm{F}}(D^2) = \sigma_D^2 + \mu_D^2. \tag{3}$$

Furthermore, the DTF index can also be written as

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

The CDIF index for a given item can be expressed as

$$\mathrm{CDIF}_i = E_{\mathrm{F}}(Dd_i) = \mathrm{Cov}(D, d_i) + \mu_D \mu_{d_i}, \tag{5}$$

where Cov stands for covariance. Additional information about the CDIF, NCDIF, and DTF indices can be found in Flowers et al. (1999), Oshima et al. (1997), and Raju et al. (1995).

Unlike other IRT-based DIF indices such as the area measure (Raju, 1988), NCDIF is a "weighted" measure of the squared difference between the two-item response functions. The typical process of DFIT involves separate calibrations of item parameters for the focal and reference groups, equating of those two sets of item parameters (i.e., the item parameter estimates from the reference group are put on the scale of those from the focal group), and finally the calculation of the DFIT indices.

### The Item Parameter Replication Method for Determining Cutoff Values

When researchers (such as Bolt, 2002; Flowers et al., 1999; Raju et al., 1995) derive a cutoff score for the NCDIF index, a typical procedure has been to simulate data sets using a large number of items similar to their data set (focal and reference groups with no DIF), and to go through the whole DFIT framework including item parameter calibration, linking, and finally the DFIT analysis. An alternative and more labor-intensive approach is to repeat the whole process many times with the same test length as the original data set to obtain a large number of NCDIF values under the no-DIF condition. Chamblee (1998), for example, replicated this process 100 times.

Then, the NCDIF index associated with the 99th (or any other) percentile rank from the distribution of all NCDIF values is defined as the cutoff score. This approach typically produces one cutoff value for all items.

The new method introduced here, the item parameter replication (IPR) method, differs from the methods described above in several aspects. First, a large number of replications of item parameters are generated from the initial set of item parameter estimates obtained from a computer program such as BILOG-MG3 (Zimowski, Muraki, Mislevy, & Bock, 2002), thus eliminating the need for extra calibrations of item parameters, which is one of the most time-consuming aspects of the methods mentioned above. Second, the distribution of NCDIF indices is obtained for each item, making it possible to generate a cutoff value for each item. Finally, it offers a computer program so that the only task that practitioners have to do is to provide estimates of item parameters and their variances and covariances, along with the ability estimates, from a computer program like BILOG-MG3 for the focal group (or the reference group), which they currently need to do for the DFIT analysis anyway. The algorithmic details of this new method are described below.

### *Algorithm of the IPR Method*

The IPR method begins with estimates of item parameters and their variances and covariances obtained from an IRT calibration program. These estimates for each item will form the basis for generating additional item parameters for that item with the restriction that the expectation of the newly generated item parameters equals the initial estimates of item parameters with the same variance and covariance structure. Since the IPR method is the same for all items in a test, it will only be described here for a single item ($i$). The nine major steps involved in the IPR method are as follows.

1. Let the item parameter estimates from the focal group be denoted by a column vector, $M_i$, for item $i$. In the case of the 3PL model, $M_i$ will consist of three elements ($b_i$, $a_i$, and $c_i$ item parameters) as shown below:

$$M_i = \begin{bmatrix} b_i \\ a_i \\ c_i \end{bmatrix}. \tag{6}$$

   In the case of the 1PL or the Rasch model, $M_i$ will be a scalar with an estimate of the $b$ parameter. Associated with each item is a matrix, $V_i$, consisting of the sampling variances and covariances of the item parameter estimates:

$$V_i = \begin{bmatrix} \sigma^2_{b_i} & \sigma_{b_i a_i} & \sigma_{b_i c_i} \\ \sigma_{a_i b_i} & \sigma^2_{a_i} & \sigma_{a_i c_i} \\ \sigma_{c_i b_i} & \sigma_{c_i a_i} & \sigma^2_{c_i} \end{bmatrix}. \tag{7}$$

   The information in $V_i$ is also typically provided by the commercially available IRT calibration programs. Let $R_i$ represent the correlation matrix for the

item parameters of item $i$. These item parameter intercorrelations can be derived from $V_i$:

$$R_i = \begin{bmatrix} 1 & \rho_{b_i a_i} & \rho_{b_i c_i} \\ \rho_{a_i b_i} & 1 & \rho_{a_i c_i} \\ \rho_{c_i b_i} & \rho_{c_i a_i} & 1 \end{bmatrix}. \tag{8}$$

Assuming that $R_i$ is positive definite, it can be expressed as the product of a triangular matrix ($T_i$) and its transpose ($T_i'$) (Graybill, 1969), that is,

$$R_i = T_i' T_i. \tag{9}$$

In the present context, $T_i$ can be expressed as

$$T_i = \begin{bmatrix} 1 & \rho_{b_i a_i} & \rho_{b_i c_i} \\ 0 & \sqrt{1 - \rho_{b_i a_i}^2} & \dfrac{\rho_{a_i c_i} - \rho_{b_i a_i}\rho_{b_i c_i}}{\sqrt{1 - \rho_{b_i a_i}^2}} \\ 0 & 0 & \sqrt{1 - \left[\rho_{b_i c_i}^2 + \dfrac{\left(\rho_{a_i c_i} - \rho_{b_i a_i}\rho_{b_i c_i}\right)^2}{\left(1 - \rho_{b_i a_i}^2\right)}\right]} \end{bmatrix}. \tag{10}$$

For the 2PL model, the above matrix reduces to

$$T_i = \begin{bmatrix} 1 & \rho_{b_i a_i} \\ 0 & \sqrt{1 - \rho_{b_i a_i}^2} \end{bmatrix}. \tag{11}$$

For the Rasch model, $T_i$ becomes a scalar with a unit as its value.

2. Let $k$ represent the IRT model under consideration. For the Rasch model, $k = 1$, for 2PL, $k = 2$, and for 3PL, $k = 3$. Now, let $X_{1i}$ represent a column vector of $k$ elements, with each element drawn at random from one of $k$ independent, standardized (mean of 0 and standard deviation of 1), and normally distributed populations. Let $X_{2i}$ represent a second vector of $k$ elements similarly drawn.
3. Using the $T_i$ matrix in Equation (9), transform the two $X$ vectors into two $Z$ (column) vectors as follows:

$$Z_{1i} = T_i' X_{1i}, \tag{12}$$

$$Z_{2i} = T_i' X_{2i}. \tag{13}$$

Each $Z$ vector now represents a random element from a $k$-dimensional standardized multivariate normal distribution with a correlation structure for the $k$ dimensions conforming to the correlation structure in the $R_i$ matrix.

4. By definition, each element in the $Z$ vectors is standardized in that its expectation and variance are 0 and 1, respectively. Each $Z$ vector is now transformed to a $Y$ vector so that the elements in the new vector will have the appropriate mean and variance as shown in the $M_i$ and $V_i$ matrices above. To achieve this transformation, let $D_i$ represent a diagonal matrix consisting of the diagonal elements (variances) in $V_i$. Now, let

$$Y_{1i} = D_i^{1/2} Z_{1i} + M_i, \tag{14}$$

$$Y_{2i} = D_i^{1/2} Z_{2i} + M_i. \tag{15}$$

5. Vectors $Y_{1i}$ and $Y_{2i}$ represent two estimates of item parameters from two populations with identical item parameters; these vectors may be thought as representing item parameter estimates for the focal and reference groups when the true DIF is zero. That is, any difference in these two sets of estimates is simply due to sampling error. Therefore, an NCDIF index for item $i$ can be obtained with the help of the two $Y$ vectors and the estimates of thetas for the focal group, using the computations spelled out in Raju et al. (1995).
6. Steps 1–5 can be replicated as many times as one wishes (e.g., 100, 1,000, . . . , or 10,000 times).
7. NCDIF values from all replications obtained in Step 6 will be rank ordered and the 90th, 95th, 99th, and 99.9th percentile rank scores are recorded to establish the cutoff values for alpha levels at .10, .05, .01, and .001, respectively.
8. Once the alpha level is chosen, the cutoff associated with it will be used as the cutoff for assessing statistical significance of the initial NCDIF value obtained for item $i$.
9. Steps 1–8 are repeated for all items in the test, thus potentially resulting in different cutoffs for different items.

Figure 1 displays the above-described algorithm graphically. A SAS-IML program named "DIFCUT" (Nanda, Oshima, & Gagne, in press) is available to execute the above algorithm. This program can be used with the 1PL, 2PL, and 3PL unidimensional models with dichotomous scoring. In the next section, the performance of this new method will be demonstrated.

## Method

### *Item Parameter and Data Generation*

The specific item parameters used for simulating the 20- and 40-item tests are shown in Table 1. These prespecified (40) $a$ and $b$ item parameters are identical to those used in a study by Raju et al. (1995). Raju et al. did not have the 20-item condition, but it was added for this study to demonstrate the effect of the number of items ($n$) on the IPR-based cutoff values. The item parameters for the 20-item test were a subset of the item parameters for the 40-item test. As shown in Table 1, the item parameters associated with the odd-numbered items formed the basis for the 20-item test. For convenience, items in the 20-item test were assigned their own identifying

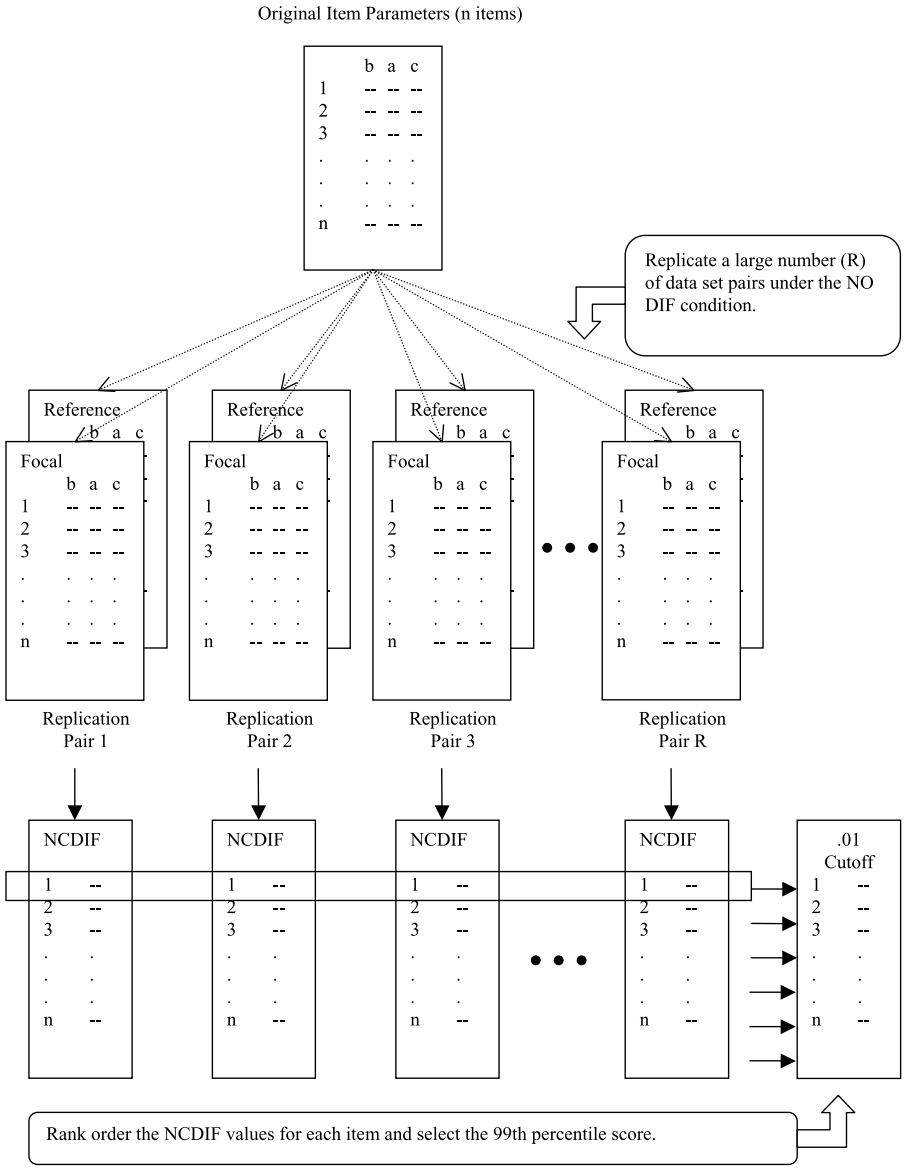


FIGURE 1. *A graphic representation of the IPR algorithm.*

numbers, as shown in column 2 of Table 1. In Raju et al.'s study, a 2PL model was used to generate the data sets. In this study, 1PL and 3PL models were also added to demonstrate the effect of the IRT model type on the IPR-based cutoff values. For the 1PL model, only the *b* parameters from Table 1 were used. For the 3PL model, a constant of .20 was used as the *c* parameter. The same data generation method was used as described in Raju et al. In addition to two sample-size combinations used

TABLE 1
*Item Parameters for the 40-Item Test and the 20-Item Test*

| Item | | Reference | | Focal (10%) | | Focal (20%) | |
|---|---|---|---|---|---|---|---|
| 40 | 20 | *a* | *b* | *a* | *b* | *a* | *b* |
| 1 | 1 | .55 | .00 | | | | |
| 2 | | .55 | .00 | | | | |
| 3 | 2 | .73 | −1.04 | | | | |
| 4 | | .73 | −1.04 | | | | |
| 5 | 3 | .73 | .00 | .73 | 1.00 | .73 | 1.00 |
| 6 | | .73 | .00 | | | | |
| 7 | 4 | .73 | .00 | | | | |
| 8 | | .73 | .00 | | | | |
| 9 | 5 | .73 | 1.04 | | | | |
| 10 | | .73 | 1.04 | .73 | 1.54 | .73 | 1.54 |
| 11 | 6 | 1.00 | −1.96 | | | | |
| 12 | | 1.00 | −1.96 | | | | |
| 13 | 7 | 1.00 | −1.04 | | | | |
| 14 | | 1.00 | −1.04 | | | | |
| 15 | 8 | 1.00 | −1.04 | 1.00 | −.04 | .50 | −.54 |
| 16 | | 1.00 | −1.04 | | | | |
| 17 | 9 | 1.00 | .00 | | | | |
| 18 | | 1.00 | .00 | | | | |
| 19 | 10 | 1.00 | .00 | | | | |
| 20 | | 1.00 | .00 | 1.00 | .50 | .50 | .00 |
| 21 | 11 | 1.00 | .00 | | | | |
| 22 | | 1.00 | .00 | | | | |
| 23 | 12 | 1.00 | .00 | | | | |
| 24 | | 1.00 | .00 | | | | |
| 25 | 13 | 1.00 | 1.04 | | | 1.00 | 2.04 |
| 26 | | 1.00 | 1.04 | | | | |
| 27 | 14 | 1.00 | 1.04 | | | | |
| 28 | | 1.00 | 1.04 | | | | |
| 29 | 15 | 1.00 | 1.96 | | | | |
| 30 | | 1.00 | 1.96 | | | 1.00 | 2.46 |
| 31 | 16 | 1.36 | −1.04 | | | | |
| 32 | | 1.36 | −1.04 | | | | |
| 33 | 17 | 1.36 | .00 | | | | |
| 34 | | 1.36 | .00 | | | | |
| 35 | 18 | 1.36 | .00 | | | .86 | .50 |
| 36 | | 1.36 | .00 | | | | |
| 37 | 19 | 1.36 | 1.04 | | | | |
| 38 | | 1.36 | 1.04 | | | | |
| 39 | 20 | 1.80 | .00 | | | | |
| 40 | | 1.80 | .00 | | | 1.30 | .00 |

in Raju et al. ($N = 500{:}500$, $N = 1{,}000{:}1{,}000$), where the first number indicates the number of examinees in the focal group and the second number indicates the number of examinees in the reference group, one more combination ($N = 500{:}1{,}000$) was added for this study.

### *DIF Levels*

Three different DIF levels were investigated: No (0%) DIF, 10% DIF, and 20% DIF. In the 0% DIF condition, all item parameters were assumed to be the same for both the reference and focal groups. That is, the *a* and *b* item parameters given in columns 3 and 4, respectively, of Table 1 were used with the reference group as well as with the focal group. In the 10% DIF level, two items (items 3 and 8) in the 20-item test and four items (items 5, 10, 15, and 20) in the 40-item test had different item parameters in the focal group, indicating DIF, as shown in columns 5 and 6 of Table 1. It should be noted that items 3 and 8 in the 20-item test were identical to items 5 and 15, respectively, in the 40-item test. All four items in this level reflected uniform DIF, that is, only the *b* parameters were different between the reference and focal groups. In all four cases, the *b* parameters for the focal group were higher than those for the reference group, that is, these four items were more difficult for the focal group. In both tests, the remaining items were assumed to have the same item parameters for the reference and focal groups.

At the 20% DIF level, items 3, 8, 13, and 18 in the 20-item test and items 5, 10, 15, 20, 25, 30, 35, and 40 in the 40-item test had different item parameters for the focal group, as shown in columns 7 and 8 of Table 1. There were both uniform DIF and nonuniform DIF items at this level; items 8 and 18 exhibited nonuniform DIF in the 20-item test, whereas items 15, 20, 35, and 40 reflected nonuniform DIF in the 40-item test. As before, the focal group item parameters for the other (non-DIF) items were identical to those of the reference group item parameters. It should be noted that, due to the process of creating the 20-item test from the 40-item test by choosing the odd-numbered items, the 40-item test contained items with smaller amounts of DIF than the 20-item test. For example, among the four DIF items in the 40-item test (10% DIF), item 5 had a larger amount of DIF (*b* difference of 1) than item 10 (*b* difference of .5), and item 15 (*b* difference of 1) had a larger amount of DIF than item 20 (*b* difference of .5). Items 5 and 15 (renamed as items 3 and 8 for the 20-item test) were used for the two DIF items for the 20-item test (10% DIF). Similarly, four of the eight DIF items in the 40-item test were used for the 20-item test for the 20% DIF condition. Those four items (items 5, 15, 25, and 35) had larger amounts of DIF than the four items not used (items 10, 20, 30, and 40).

### *Impact*

Two different ability distributions were considered: identical ability distributions for the focal and reference groups and different ability distributions for the focal and reference groups. In the first case, the focal- and reference-group $\theta$ distributions were assumed to be normal with a mean of zero and a standard deviation of one. In the second case, the focal group had a lower theta mean (by .5) than the reference group, which is commonly known as impact.

### *Determining the Number of Replications in DIFCUT*

Prior to applying the new IPR algorithm for determining cutoff scores, the number of replications needed to achieve stable cutoff scores had to be determined. Using one condition ($n = 40$, $N = 1{,}000$, DIF = 10%), cutoff scores were obtained over

various replication sizes (100–1,000 by an increment of 100) to determine the minimum number of replications. In addition, cutoff scores were obtained with 10,000 replications as approximations to the population values. These cutoff data were used to come up with a number for a minimum number of replications needed for reliable estimation of item-level cutoffs for use in this investigation. Although a larger replication size is desirable (such as 10,000), given today's typical computer speed (about 1 minute and 36 seconds per 100 replications using a machine with Pentium 4, 1G RAM), a practical compromise had to be made.

### *Procedures*

As described above, the current investigation included two test lengths ($n = 20$ items and $n = 40$ items), three sample-size combinations ($N = 500{:}500$, $N = 1{,}000{:}1{,}000$, and $N = 500{:}1{,}000$), three different levels of DIF (0%, 10%, and 20%), three IRT models (1PL, 2PL, and 3PL), and two ability distribution combinations (no-impact and impact). This completely crossed design resulted in a total of 108 ($2 \times 3 \times 3 \times 3 \times 2$) conditions. However, the 20% DIF condition was not considered for the 1PL model, since nonuniform DIF items were included in the 20% condition. Thus, the final total was 96 ($108 - 12$) conditions.

For each condition, each pair of generated 1–0 data sets (one for the focal group and another for the reference group) was calibrated with BILOG-MG3. For the 3PL model, the $c$ parameter was constrained at .20. BILOG-MG3 is an extension of BILOG and is designed for a multiple-group analysis as well as a single-group analysis. A conventional single-group IRT analysis was used for each group. While it may be possible to conduct a concurrent multiple-group IRT analysis with BILOG-MG3, the method of separate IRT calibrations was employed here because some investigators may use an older version of BILOG or an IRT calibration program without the concurrent calibration option. In the case of a concurrent calibration, the equating step described next would not be required. Second, item parameter estimates from the two groups were put on the same scale using the IPLINK program (Lee & Oshima, 1996). Third, using the item parameter estimates from the focal group, cutoff values were determined using the DIFCUT program. Fourth, DIFCUT calculated the NCDIF values for the pair of original data sets (reference and focal groups), and identified items with NCDIF values larger than the cutoff values. Fifth, using only the items deemed DIF-free by the fourth step, second-stage linking coefficients were obtained from IPLINK. Finally, using the linking coefficients from the second-stage linking, NCDIF values were calculated for all items and DIF was determined using the cutoff values obtained in the third step above. This process is identical to the traditional DFIT analysis except the third step. In this investigation, a significance level of .01 was used.

## Results

Reported in Figures 2 and 3 are results from the investigation as to how many replications are necessary as a minimum to determine the cutoff values. A graph similar to Figure 2 was plotted for each of the 40 items. Due to space limitation, only

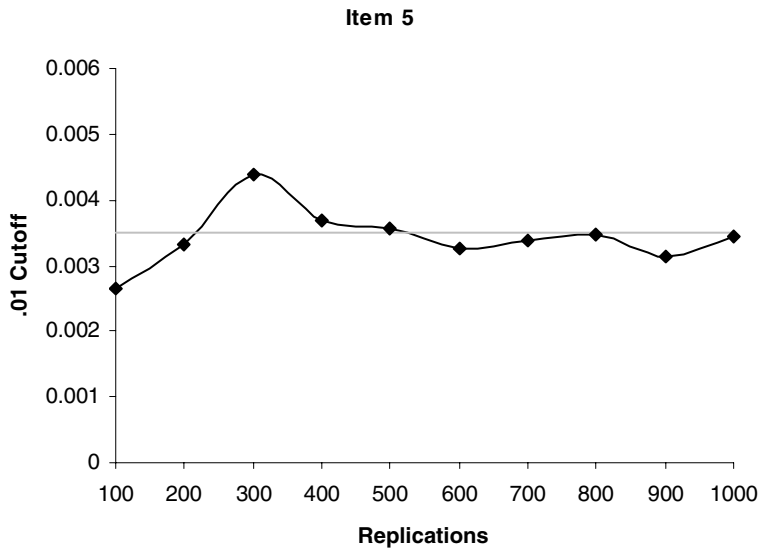


FIGURE 2. *The .01 cutoff value produced by each number of replications for item 5.*

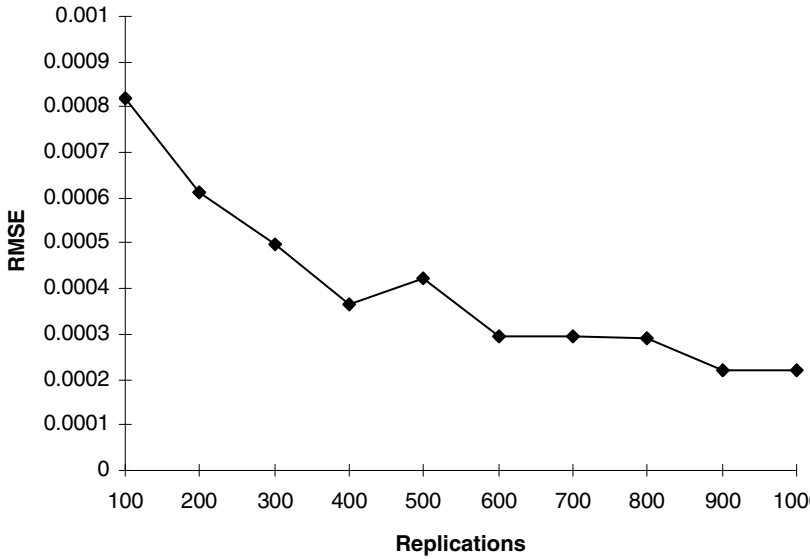


FIGURE 3. *RMSE across all items by each number of replications.*

the graph for item 5 is presented here as an example. The horizontal line (parallel to the *x*-axis) in Figure 2 represents a cutoff of .0035 based on 10,000 replications. The cutoff values from other sets of replications (100, 200, . . . , 1,000) are plotted against the cutoff of .0035 (based on 10,000 replications) in Figure 2. Although there were

some variations over items, generally it appeared that after 500–600 replications, the cutoff values did not change much and also stayed close to the value from 10,000 replications. To summarize numerically across all the items, root mean square error (RMSE) was calculated over items for each level of replications. Here, RMSE was defined as the square root of the average squared difference between the cutoff value from each level of replications and the cutoff value from the 10,000 replications. By a visual inspection, it appeared that the RMSE decreased rapidly up to 600 replications, and the decrease seemed rather small after 600 replications as seen in Figure 3. The cutoff values obtained from 1,000 replications were used in the current investigation.

Table 2 shows the number of false positives (FPs; identifying non-DIF items as having significant DIF) and false negatives (FNs; identifying DIF items as having no DIF) for the 48 conditions with no impact. This table was made similar to Table 1 presented in Raju et al. (1995) for ease of comparison. Also reported in Table 2 in the parentheses are the overall-cutoff values for each condition under the new method. The overall-cutoff value for a given condition is the 99th percentile rank score for *all* NCDIF values from 1,000 replications across all items within a test. Although a separate cutoff score was determined for each item, the overall cutoff value offers a summary statistic across items, providing a feel for the magnitude of the cutoff

TABLE 2
*False Positive (FP) and False Negative (FN) for the No-Impact Conditions*

| | $N$ = 500:500 | | | $N$ = 1,000:1,000 | | | $N$ = 500:1,000 | | |
|---|---|---|---|---|---|---|---|---|---|
| | 1PL | 2PL | 3PL | 1PL | 2PL | 3PL | 1PL | 2PL | 3PL |
| | | | | (a) Test length = 20 items | | | | | |
| 0% | (.0054)[a] | (.0074) | (.0100) | (.0027) | (.0036) | (.0050) | (.0054) | (.0074) | (.0100) |
| FP | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| FN | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 10% | (.0054) | (.0073) | (.0098) | (.0027) | (.0036) | (.0050) | (.0054) | (.0073) | (.0098) |
| FP | 1 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 |
| FN | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 20% | (NA) | (.0075) | (.0118) | (NA) | (.0037) | (.0054) | (NA) | (.0075) | (.0118) |
| FP | NA | 0 | 0 | NA | 0 | 0 | NA | 0 | 0 |
| FN | NA | 0 | 1 | NA | 0 | 0 | NA | 0 | 1 |
| | | | | (b) Test length = 40 items | | | | | |
| 0% | (.0052)[a] | (.0068) | (.0088) | (.0026) | (.0034) | (.0042) | (.0052) | (.0068) | (.0088) |
| FP | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 |
| FN | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 10% | (.0051) | (.0068) | (.0089) | (.0026) | (.0033) | (.0043) | (.0051) | (.0068) | (.0089) |
| FP | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| FN | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 |
| 20% | (NA) | (.0071) | (.0095) | (NA) | (.0035) | (.0044) | (NA) | (.0071) | (.0095) |
| FP | NA | 0 | 2 | NA | 0 | 0 | NA | 0 | 0 |
| FN | NA | 2 | 4 | NA | 2 | 2 | NA | 2 | 5 |

[a]The overall-cutoff value is the 99th percentile rank score for *all* NCDIF values from 1,000 replications across all items in a given condition.

values across conditions.

A few observations can be made from Table 2. First, the new method seemed to work quite well in terms of the numbers of the FPs and FNs except for the 20% DIF condition with the 3PL model for the 40-item test. The relatively poor performance of this particular condition (with respect to FNs) can be due to a combination of several factors: the smaller magnitude of DIF embedded in the 20% condition for the 40-items test, the potentially poor parameter estimation due to small sample size ($N = 500$) for the 3PL model, and the way the $c$ parameters were simulated and estimated. The last two factors may have contributed to the inflation of the cutoff values for the 3PL conditions, thus leading to higher FN rates. Additional research is certainly needed to better understand the FN rates for the 3PL model.

Second, the overall cutoff values varied from .0026 to .0118. The effects of the sample size and the type of IRT models were evident. The smaller sample size resulted in higher cutoff values. The more parameters in the IRT model resulted in higher cutoff values. These results coincide with Chamblee's (1998) results, and indicate that her replication method and the IPR method produce similar results. In addition, the shorter test length had slightly higher cutoff values.

Reported in Table 3 are the results from the impact condition. Overall, the number of FPs and FNs tended to be slightly more for some of the conditions than those

TABLE 3
*False Positive (FP) and False Negative (FN) for the Impact Conditions*

| | $N = 500{:}500$ | | | $N = 1{,}000{:}1{,}000$ | | | $N = 500{:}1{,}000$ | | |
|---|---|---|---|---|---|---|---|---|---|
| | 1PL | 2PL | 3PL | 1PL | 2PL | 3PL | 1PL | 2PL | 3PL |
| | | | | (a) Test length = 20 items | | | | | |
| 0% | (.0053)[a] | (.0074) | (.0129) | (.0027) | (.0037) | (.0060) | (.0053) | (.0074) | (.0129) |
| FP | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| FN | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 10% | (.0053) | (.0073) | (.0110) | (.0026) | (.0037) | (.0061) | (.0053) | (.0073) | (.0110) |
| FP | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 |
| FN | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 20% | (NA) | (.0075) | (.0153) | (NA) | (.0037) | (.0067) | (NA) | (.0075) | (.0153) |
| FP | NA | 0 | 0 | NA | 0 | 0 | NA | 0 | 0 |
| FN | NA | 0 | 2 | NA | 0 | 1 | NA | 0 | 2 |
| | | | | (b) Test length = 40 items | | | | | |
| 0% | (.0050)[a] | (.0068) | (.0107) | (.0026) | (.0033) | (.0046) | (.0050) | (.0068) | (.0107) |
| FP | 0 | 1 | 1 | 0 | 1 | 1 | 0 | 0 | 0 |
| FN | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 10% | (.0050) | (.0069) | (.0104) | (.0025) | (.0034) | (.0047) | (.0050) | (.0069) | (.0104) |
| FP | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| FN | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 1 |
| 20% | (NA) | (.0072) | (.0138) | (NA) | (.0034) | (.0051) | (NA) | (.0072) | (.0138) |
| FP | NA | 0 | 2 | NA | 0 | 0 | NA | 0 | 0 |
| FN | NA | 2 | 5 | NA | 2 | 4 | NA | 3 | 6 |

[a]The overall-cutoff value is the 99th percentile rank score for *all* NCDIF values from 1,000 replications across all items in a given condition.

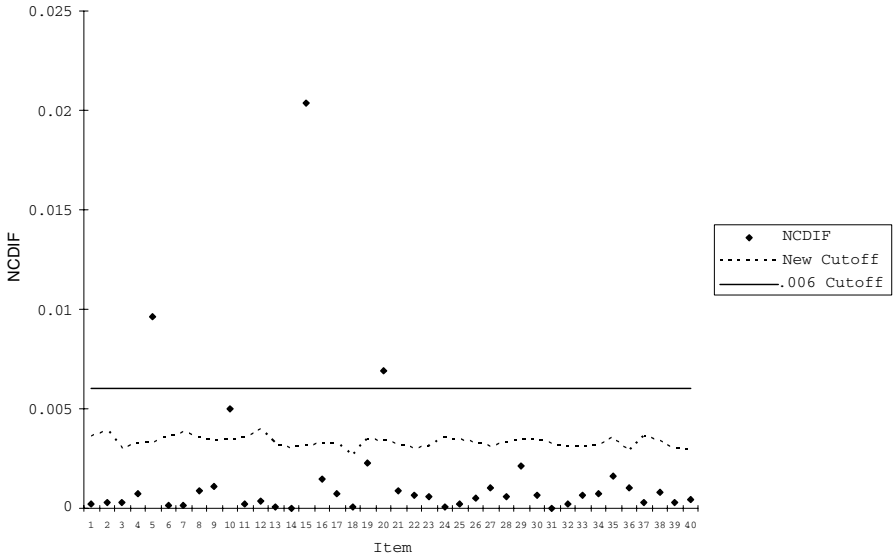


FIGURE 4. *A graphic presentation of the new cutoff method (the IPR method) and the old cutoff method (the .006 method) for NCDIF with n = 40, N = 1,000, and 10% DIF.*

shown in Table 2. The cutoff values for the 1PL and 2PL models were comparable between the no-impact and impact conditions. However, the cutoff values for the 3PL model were larger, which may have contributed to the higher number of FNs.

To further investigate the case where the new and old methods make a difference, an additional condition was created with less magnitude of DIF for one 2PL condition ($n = 40$, $N = 1{,}000$, 10% DIF). The magnitude of DIF was cut in half for all four DIF items. The $b$ parameters for the focal group were .50, 1.29, $-.54$, and .25 for items 5, 10, 15, and 20, respectively. With an overall cutoff value of .0033 for the new method, the old method (with the cutoff value of .006) missed one DIF item (item 10). The new method picked all the DIF items while keeping the FP rate to zero. Figure 4 shows graphically the difference between the new and old cutoff values for this condition (after the first-stage linking). The straight line at .006 is the cutoff line for the old method. The zigzag line below it indicates the varying cutoff values based on the new method. The dots indicate the calculated NCDIF values. In the old method, the cutoff value always stays at .006 over items across IRT models and sample sizes. In the new method, the cutoff values could vary from item to item (as shown by the zigzag line) and furthermore, as shown in Tables 2 and 3, the overall cutoff values could vary from 1PL to 2PL to 3PL and also from one sample size to the next.

## Discussion

A new method to determine the cutoff values for NCDIF in the DFIT framework was introduced in this article. Under the conditions examined in the study, this new cutoff method, based on the IPR concept, controlled the FN and FP rates reason-

ably well. Furthermore, the new method appeared to be robust against factors that contribute to differences in the standard errors across groups (such as differences in ability distributions and sample size's). The new method offers several theoretical advantages over the old method where a fixed value of .006 was used as the cutoff score for dichotomous items. First, it is tailored to a particular data set a user presents. For example, a data set with smaller sample size would receive higher cutoff values, and a data set calibrated with an IRT model with more parameters to estimate (e.g., the 3PL model) would receive higher cutoff values. In general, the larger the item parameter estimate standard errors are, the larger the cutoff values are. In other words, other unknown factors that may influence the standard errors of item parameter estimates are taken into account with the new method. Other possible factors influencing the standard error include the fit of the IRT model to the data and the difficulty level of the test. Second, this new method produces the cutoff score for each item. This again is an advantage over the old method, since the standard errors of item parameter estimates may differ from item to item.

This new method also has an advantage over other simulation methods designed for empirically deriving cutoff scores. This new method bypasses the time-consuming repeated calibrations of item parameters. A computer program, DIFCUT (Nanda et al., in press), is available and the labor required for the user is minimal. The DIFCUT program provides not only the cutoff values but also compares the cutoff values to the calculated NCDIF values and flags the DIF items.

The current IPR algorithm is based on replicating (or cloning) a single file that contains item parameter estimates and their variance-covariance information. For this study we used a file from the focal group. Anonymous reviewers made an insightful suggestion of making use of the variance-covariance (or just the variance) information from the *reference* group as well as that from the focal group, while using the item parameter estimates from the focal group only. The alternative algorithms were speculated to work better when the sample size differs between the focal and reference groups. We applied these alternative algorithms to the $N = 500{:}1{,}000$ conditions. Under the conditions we examined, however, the alternative methods did not always outperform the current IPR method. The alternative methods had slightly lower FN rates with slightly higher FP rates. Although it is only a speculation, the reason as to why the alternative methods did not improve the original method could be related to how one defines the null condition (i.e., the no-DIF condition). With the original method, two sets of simulated item parameters came from *identical* distributions based on the focal group only. For the alternative methods, the strict null condition was relaxed to accommodate the reference group variance-covariance information, and it is not clear what impact these varying distributions would bring.

In the general IPR framework, item-level cutoffs for NCDIF indices can be obtained either with the sampling (error) variances and covariances and $\theta$ estimates from the focal group or with the sampling (error) variances and covariances and $\theta$ estimates from the reference group; the cutoffs can even be obtained with the sampling (error) variances and covariances and $\theta$ estimates from the combined (focal plus reference) sample. Based on the results from this study as well as that from Chamblee (1998), these different sets of cutoffs could be different because of the

differences in sample sizes for the focal and reference groups. Other things being equal, larger samples will result in smaller sampling errors and hence lower cutoff values. The question of which of these sets of cutoffs offers a better mechanism for accurately identifying DIF is an empirical one. The current investigation only briefly touched on this issue, and there is definitely a need for additional research on this question. Meanwhile, we recommend that practitioners eyeball the standard errors for substantial differences across groups when performing their analysis.

The current investigation did not extend the IPR method to generate cutoff values for the DTF index. This extension is currently under way. Also under consideration is the extension of the IPR method to polytomously scored items. We hope to work on these and other extensions in the near future.

The only probable drawback of this new method is the computer time. However, by keeping the number of replications to 1,000, the computer run time is manageable at the current time. With a rapid improvement of computer technology, this problem should soon dissipate. With a faster computer, the user can increase the number of replications, if desired.

## Note

Portions of this manuscript were presented at the 2005 International Meeting of the Psychometric Society in Tilburg, the Netherlands. The authors would like to express their appreciation to Terry Ackerman for presenting this research in Tilburg since none of the authors could make it to the 2005 International meeting. The authors are very grateful to three anonymous reviewers and the Editor for their many helpful comments on an earlier version of this manuscript.

## References

Bolt, D. M. (2002). A Monte Carlo comparison of parametric and nonparametric polytomous DIF detection methods. *Applied Measurement in Education, 2,* 113–141.

Chamblee, M. C. (1998). *A Monte Carlo investigation of conditions that impact Type I error rates of DFIT*. Unpublished doctoral dissertation, Georgia State University.

Cohen, A. S., Kim, S., & Baker, F. B. (1993). Detection of differential item functioning in the graded response model. *Applied Psychological Measurement, 17,* 335–350.

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,* 54-04, 2266B.

Flowers, C. P., Oshima, T. C., & Raju, N. S. (1999). A description and demonstration of the polytomous-DFIT framework. *Applied Psychological Measurement, 23,* 309–326.

Graybill, F. A. (1969). *Introduction to matrices with applications in statistics*. Belmont, CA: Wadsworth Publishing.

Kim, S.-H., & Cohen, A. S. (1991). A comparison of two area measures for detecting differential item functioning. *Applied Psychological Measurement, 15,* 269–278.

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

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

Muraki, E. (1999). Stepwise analysis of differential item functioning based on multiple-group partial credit model. *Journal of Educational Measurement, 36,* 217–232.

Nanda, A. O., Oshima, T. C., & Gagné, P. (in press). DIFCUT: A SAS-IML program for calculating cutoff scores for DFIT. *Applied Psychological Measurement*.

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

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. (1995). An IRT-based internal measure of test bias with applications for differential item functioning. *Applied Psychological Measurement, 19,* 353–368.

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. I. Braun (Eds.), *Test validity* (pp. 147–169). Hillsdale, NJ: Erlbaum.

Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (2002). BILOG-MG3 [Computer software]. Chicago, IL: Scientific Software International.

## Authors

T. C. OSHIMA is an Associate Professor, Department of Educational Policy Studies, College of Education, Georgia State University, PO Box 3977, Atlanta, GA 30302-3977; oshima@gsu.edu. Her primary research interests include item response theory and differential item functioning.

NAMBURY S. RAJU was a Distinguished Professor, Institute of Psychology, Illinois Institute of Technology, Chicago, IL 60616-3793. His primary research interests included psychometric theory, test development, and industrial/organizational psychology.

ALICE O. NANDA is a Graduate Student, Department of Educational Psychology and Special Education, College of Education, Georgia State University, PO Box 3979, Atlanta, GA 30302-3979; aliowens@aol.com. Her primary research interests include item response theory, differential item functioning, and reading acquisition.