The Item Parameter Replication Method for Detecting Differential Functioning in the Polytomous DFIT Framework

Nambury S. Raju, Kristen A. Fortmann-Johnson, Wonsuk Kim, Scott B. Morris, Michael L. Nering and T.C. Oshima

Applied Psychological Measurement 2009 33: 133
DOI: 10.1177/0146621608319514

The Item Parameter Replication Method for Detecting Differential Functioning in the Polytomous DFIT Framework

Nambury S. Raju and Kristen A. Fortmann-Johnson, Illinois Institute of Technology
Wonsuk Kim, Measured Progress
Scott B. Morris, Illinois Institute of Technology
Michael L. Nering, Measured Progress
T. C. Oshima, Georgia State University

The recent study of Oshima, Raju, and Nanda proposes the item parameter replication (IPR) method for assessing statistical significance of the noncompensatory differential item functioning (NCDIF) index within the differential functioning of items and tests (DFIT) framework. Previous Monte Carlo simulations have found that the appropriate cutoff values for determining statistical significance of NCDIF depend on sample size and the item response theory (IRT) model used for the analysis. The IPR method simplifies the process of identifying cutoff values that are tailored to a particular research setting. This approach has been shown to be effective for detecting differential item functioning (DIF) in dichotomous items. The current article extends the IPR method to the polytomous case. *Index terms: item response theory, differential item functioning, statistical significance, differential functioning of items and tests, polytomous data, item bias*

Important decisions in educational, business, and medical settings often rely on standardized tests and questionnaires. Before comparisons among individuals or groups can be made, one must consider if the tests and questionnaires used provide equivalent measurement. Measurement equivalence is obtained when the relations between observed scores and latent constructs are identical across relevant subgroups (Drasgow & Kanfer, 1985). Without measurement equivalence, observed scores from different groups are in different scales and are therefore not comparable. In other words, the meaning of score differences is unclear because they reflect not only meaningful group differences but also systematic measurement error.

Measurement equivalence can be defined either in terms of individual items or the measure as a whole. Systematic measurement errors on a single item are referred to as differential item functioning (DIF). Measures of DIF are useful for scale development by selecting items that are equivalent across groups. Differential test functioning (DTF) refers to systematic measurement errors of an entire test or scale. Because decisions are generally based on scale-level scores, DTF more closely reflects the practical impact of nonequivalence.

Aptitude and achievement testing is one area in which measurement equivalence across groups is of extreme importance (Drasgow & Kanfer, 1985). If some groups (e.g., minority groups)

Applied Psychological Measurement, Vol. 33 No. 2, March 2009, 133–147
DOI: 10.1177/0146621608319514

systematically receive lower mean test scores than other groups (e.g., Whites), one must determine whether these differences are due to true differences on the latent construct or measurement error. If score differences at the item or test level are a reflection of systematic measurement error, an item or test is said to be biased. Test bias could potentially cause adverse impact in the selection rates, even when both groups are equally talented.

Similarly, measurement equivalence should be examined for attitudinal questionnaires (Drasgow & Kanfer, 1985). Questionnaires are used to assess organizational attitudes and identify the need for training and organizational development initiatives. Score differences among individuals from different groups, (e.g., males vs. females, Whites vs. minority groups, managers vs. nonmanagers, etc.) are typically assumed to reflect true group differences on the attitude being measured. Costly organizational interventions are then planned to address these group differences. If the DIF or DTF is present in the questionnaires used, however, observed group differences may be erroneous, and the conclusions drawn based on these results will be inappropriate. In light of the potential consequences, it is therefore of vital importance that practitioners have a readily available means by which to assess differential functioning of items and tests (DFIT).

There are several item response theory (IRT)-based procedures for assessing DIF, such as Lord's chi-square (Cohen, Kim, & Baker, 1993; Lord, 1980), the likelihood ratio test (Thissen, Steinberg, & Wainer, 1988), and area measures (Cohen et al., 1993; Kim & Cohen, 1991; Raju, 1988, 1990). Although these methods have shown to be effective in detecting the DIF, none of these provides a means by which to assess differential functioning at the test level. It is simply assumed the removal of items with significant DIF will result in a test that is unbiased (Raju, van der Linden, & Fleer, 1995). Raju et al. (1995) noted the desirability of having a psychometric measure of DTF and proposed the DFIT framework.

Although the DFIT framework has shown to be an effective mechanism for detecting DIF/DTF in IRT-based tests and questionnaires (e.g., Flowers, Oshima, & Raju, 1999; Oshima, Raju, & Flowers, 1997; Raju et al., 1995), researchers have indicated a need for better procedures for assessing the statistical significance of DIF and DTF indices.

Recently, Oshima, Raju, and Nanda (2006) proposed the item parameter replication (IPR) method for assessing the statistical significance of the noncompensatory differential item functioning (NCDIF) index of DIF for dichotomous items within the DFIT framework. The objective of the current article is to describe how Dr. Raju extended the IPR method to the polytomous case (Raju, Oshima, Fortmann, Nering, & Kim, 2006) and to examine its efficacy in detecting the DIF for polytomous items. First, however, a brief description of the IRT calibration model used and the DFIT framework are provided.

The Graded Response Model (GRM)

The DFIT framework can be used with any polytomous IRT model. Samejima's (1969) GRM was used in the current investigation and will therefore be described. The GRM assumes item response categories can be rank ordered or, in other words, represent ordinal data (Zickar, 2002). The probability of person s responding above response category k to item i may be expressed as

$$P_{ik}^{*}(\theta_s) = \frac{e^{Da_i\left(\theta_s - b_{ik}\right)}}{1 + e^{Da_i(\theta_s - b_{ik})}}, \tag{1}$$

where b_{ik} is the location parameter that designates the boundary between response categories k and $k+1$ for item i, a_i is the item discrimination parameter, and θ_s is the person (ability)

parameter. The number of *b* parameters for an item is equal to the number of response categories minus one. A five-response category item, therefore, is represented by one *a* parameter and four *b* parameters. Each *b* parameter may be thought of as the amount of attitude or ability needed to choose one response category over another or to choose one set of available categories over another set of available categories within an item. Equation (1) is commonly referred to as the boundary response function (BRF); the number of BRFs for each item is one less than the number of response categories. The BRFs are a cumulative probability of a response above category *k*. To calculate the probability of responding in a particular response category as a function of θ, the adjacent BRF is subtracted from the cumulative probability. This function is referred to as the category response function (CRF):

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

Because the first and last response categories lack an adjacent boundary, Samejima (1969) defined $P^*_{i0}(\theta)$ and $P^*_{im}(\theta)$ as equal to 1 and 0, respectively, where *m* represents the number of response categories. There will be as many CRFs in an item as there are response categories.

The DFIT Framework

The DFIT framework offers several advantages to other DIF detection methods (Flowers et al., 1999). First, it offers a means by which to assess differential functioning at both the item and test levels. Second, this method can be applied to both dichotomous and polytomous data and to unidimensional and multidimensional data. Third, the DFIT framework offers two indices for assessing DIF: NCDIF and compensatory differential item functioning (CDIF). The NCDIF index assumes all items in the test except for the item under consideration contain no DIF (Raju et al., 1995). Although this assumption is commonly made by other measures of DIF as well, it may not be plausible. CDIF, on the other hand, does not make this assumption. The CDIF index is additive in the sense that DTF equals the sum of the CDIF indices across the items in a test. This is advantageous in the sense that it provides a means by which to assess the overall effect of removing an item from a test. A description of these indices follows.

The DFIT methodology (Raju et al., 1995) begins by defining differential functioning at the test level and then decomposing it into differential functioning at the item level. For a given test, assume item parameters have been estimated separately for the focal group and reference group, and that these two sets of item parameters have been placed on a common metric. For examinee *s*, it is now possible to compute two test true scores, one treating the examinee as a member of the focal group and the other treating the examinee as a member of the reference group. If these true scores are not equal, the examinee's true score is not independent of group membership and the test is said to exhibit DTF. The greater the difference between these two true scores, the greater the magnitude of DTF. Raju et al. (1995) defined the DTF index as follows:

$$\mathrm{DTF} = E_F\left(D_s^2\right) = \sigma_D^2 + \mu_D^2, \tag{3}$$

where D_s represents the difference in test true scores for examinee *s* and the expectation (*E*) is taken over the focal group.

For a polytomous item, the expected score of individual *s* on item *i* (ES_{si}) can be defined as a weighted average of the category values, where the weights reflect the probability of the individual selecting each category. The difference between item true scores computed for a given

examinee, first treated as a member of the focal group and then treated as a member of the reference group, is defined as

$$d_{si} = ES_{si_F} - ES_{si_R} = \left(\sum_{k=1}^{m} P_{ik_F}(\theta_s) X_{ik} - \sum_{k=1}^{m} P_{ik_R}(\theta_s) X_{ik} \right). \quad (4)$$

Derived from this, CDIF is defined as

$$\text{CDIF}_i = E(d_{si} D_s) = \text{Cov}(d_{si}, D_s) + \mu_{d_i} \mu_D, \quad (5)$$

where Cov stands for covariance. As previously stated, the CDIF index is additive such that

$$\text{DTF} = \sum_{i=1}^{n} \text{CDIF}_i. \quad (6)$$

The NCDIF index for a given item is expressed as

$$\text{NCDIF}_i = E\left(d_{si}^2\right) = \sigma_{d_i}^2 + \mu_{d_i}^2. \quad (7)$$

Raju et al. (1995) recommended chi-square significance tests for the NCDIF and DTF indices. Because the CDIF indices sum to DTF, a separate significance test was not proposed for this index. Raju et al. instead recommended deletion, one at a time, of items with large, positive CDIF values. Following deletion of a single item, the chi-square test of DTF is recomputed based on the remaining items. This process is repeated until the DTF becomes nonsignificant. Deleted items are labeled as having significant CDIF. Additional information about the DTF, CDIF, and NCDIF indices may be found in Flowers et al. (1999) and Raju et al.

Monte Carlo examination of the proposed indices suggested the chi-square tests for DTF and NCDIF to be overly sensitive for large sample sizes (Fleer, 1993). At the .01 level of significance, substantially greater than 1% of the items in the no-DIF condition were falsely identified as having DIF. This suggested the need to establish cutoff values for DTF and NCDIF indices. Cutoff values provide a means to identify findings of differential functioning that are not only statistically significant but also practically nontrivial.

Cutoff values were established by creating a frequency distribution of observed NCDIF values across 50 replications of the no-DIF condition (Fleer, 1993). A cutoff value of .006 was associated with the 99th percentile and so resulted in falsely identifying approximately 1% of items as exhibiting DIF. Based on this result from Fleer's Monte Carlo study, Raju et al. (1995) recommended items with NCDIF > .006 and a statistically significant chi-square be designated as exhibiting DIF; DTF > .006 and a significant chi-square suggests differential functioning at the test level. Other researchers have similarly generated study-based cutoff values for other dichotomous (Chamblee, 1998) and polytomous (Bolt, 2002; Flowers et al., 1999) data sets. Although these study-based cutoff values have shown to be effective in identifying the DIF/DTF, they are probably not generalizable to other samples and items. In working with dichotomous items, Chamblee's study in fact suggested that optimal cutoff values are related to sample size and the specific IRT model used in the investigation. The cutoff values obtained from the distribution of observed NCDIF values ranged from .003 to .018, with smaller sample sizes and a larger number of item parameters in the IRT model yielding higher cutoff values.

This dilemma concerning appropriate cutoff values for NCDIF and DTF indices poses an obstacle to using the DFIT framework in practice. Using Monte Carlo methods, researchers have achieved success in generating cutoff values (Bolt, 2002; Flowers et al., 1999; Raju et al., 1995) that are appropriate for the specific conditions simulated in a given study. In practice, researchers

may need to identify cutoff values for situations not covered in past research, and typical practitioners may not have the expertise or time to conduct their own simulations. As such, across the literature, there has been a call for better procedures for assessing the statistical significance of DIF and DTF indices (Bolt, 2002; Flowers et al., 1999; Raju et al., 1995). In response to this call, Oshima et al. (2006) proposed the IPR method.

The IPR Method for Determination of Cutoff Values

The IPR method (Oshima et al., 2006) provides a means for deriving study-based cutoff values for use in assessing differential functioning within the DFIT framework. The IPR method begins with estimates of item parameters for the focal group and the sampling variances and covariances of these estimates. The item parameters and variances can be obtained from the output of an IRT calibration program, such as PARSCALE. Unfortunately, the covariances among item parameters are not available as part of the standard output of programs such as PARSCALE and MULTILOG. These values, however, can be derived using procedures that will be described below.

Based on these initial estimates, a large number of replications of item parameters are then generated with the restriction that the expectation of the newly generated item parameters equals the initial estimates of focal group item parameters with the same sampling variance/covariance structure. That is, any differences in the sets of estimates must be due to sampling error. Pairs of samples can then be used to compute DIF statistics. This produces an empirical sampling distribution of NCDIF under the null hypothesis that focal and reference groups have identical parameters.

It should be noted that this approach does not adjust for differences in sample size between the reference and focal groups, which could produce unequal covariance matrices even when item parameters are identical. Thus, using the focal-group covariance matrix to represent both groups may lead to some inaccuracy when sample sizes differ. However, prior simulation research by Oshima et al. (2006) found accurate results using the IPR method even with substantial sample size differences. This issue will be explored further in the simulations reported below.

The IPR method consists of nine major steps that will be described here for a single polytomous item i. The IPR method is identical for all items in a test.

1. Let the item parameter estimates be denoted by a column vector $\mathbf{M}_i$ In the case of Samejima's (1969) GRM, a polytomous item with five response categories will be represented by one a parameter and four b parameters. Therefore, $\mathbf{M}_i$ will consist of five elements:

$$\mathbf{M}_i = \begin{bmatrix} a_i \\ b_{i1} \\ b_{i2} \\ b_{i3} \\ b_{i4} \end{bmatrix}. \tag{8}$$

Each item is also associated with a matrix consisting of the sampling variances and covariances of the item parameter estimates. Let this be represented as

$$\mathbf{V}_i = \begin{bmatrix} \sigma^2_a & \sigma_{ab_{i1}} & \sigma_{ab_{i2}} & \sigma_{ab_{i3}} & \sigma_{ab_{i4}} \\ \sigma_{b_{i1}a} & \sigma^2_{b_{i1}} & \sigma_{b_{i1}b_{i2}} & \sigma_{b_{i1}b_{i3}} & \sigma_{b_{i1}b_{i4}} \\ \sigma_{b_{i2}a} & \sigma_{b_{i2}b_{i1}} & \sigma^2_{b_{i2}} & \sigma_{b_{i2}b_{i3}} & \sigma_{b_{i2}b_{i4}} \\ \sigma_{b_{i3}a} & \sigma_{b_{i3}b_{i1}} & \sigma_{b_{i3}b_{i2}} & \sigma^2_{b_{i3}} & \sigma_{b_{i3}b_{i4}} \\ \sigma_{b_{i4}a} & \sigma_{b_{i4}b_{i1}} & \sigma_{b_{i4}b_{i2}} & \sigma_{b_{i4}b_{i3}} & \sigma^2_{b_{i4}} \end{bmatrix}. \tag{9}$$

The correlation matrix $(\mathbf{R}_i)$ for the item parameters can then be derived from V_i:

$$\mathbf{R}_i = \begin{bmatrix} 1 & \rho_{ab_{i1}} & \rho_{ab_{i2}} & \rho_{ab_{i3}} & \rho_{ab_{i4}} \\ \rho_{b_{i1}a} & 1 & \rho_{b_{i1}b_{i2}} & \rho_{b_{i1}b_{i3}} & \rho_{b_{i1}b_{i4}} \\ \rho_{b_{i2}a} & \rho_{b_{i2}b_{i1}} & 1 & \rho_{b_{i2}b_{i3}} & \rho_{b_{i2}b_{i4}} \\ \rho_{b_{i3}a} & \rho_{b_{i3}b_{i1}} & \rho_{b_{i3}b_{i2}} & 1 & \rho_{b_{i3}b_{i4}} \\ \rho_{b_{i4}a} & \rho_{b_{i4}b_{i1}} & \rho_{b_{i4}b_{i2}} & \rho_{b_{i4}b_{i3}} & 1 \end{bmatrix}. \tag{10}$$

Assuming $\mathbf{R}_i$ is positive definite, it can be expressed as the product of a triangular matrix $(\mathbf{T}_i)$ and its transpose $(\mathbf{T}_i'$; Graybill, 1969):

$$\mathbf{R}_i = \mathbf{T}_i'\mathbf{T}_i. \tag{11}$$

2. Let m represent the number of response categories. Let $\mathbf{X}_{1i}$ represent a column vector of m elements, with each element drawn at random from one of m independent, standardized, and normally distributed populations. Let $\mathbf{X}_{2i}$ represent a second vector of m elements similarly drawn.
3. Using the $\mathbf{T}_i$ matrix in equation (11), transform the two $\mathbf{X}$ vectors into two $\mathbf{Z}$ vectors as follows:

$$\mathbf{Z}_{1i} = \mathbf{T}_i'\mathbf{X}_{1i}, \tag{12}$$

$$\mathbf{Z}_{2i} = \mathbf{T}_i'\mathbf{X}_{2i}. \tag{13}$$

Each $\mathbf{Z}$ vector now represents a random element from an m dimensional standardized multivariate normal distribution with a correlation structure for the m dimensions conforming to the correlation structure in the $\mathbf{R}_i$ matrix.

4. By definition, each element in the $\mathbf{Z}$ vectors is standardized in that its expectation and variance are 0 and 1, respectively. Each $\mathbf{Z}$ vector is now transformed to a $\mathbf{Y}$ vector so that the elements in the new vector will have the appropriate mean and variance as shown in the $\mathbf{M}_i$ and $\mathbf{V}_i$ matrices above. To achieve this transformation, let $\mathbf{D}_i$ represent a diagonal matrix consisting of the variances contained in $\mathbf{V}_i$. Now, let

$$\mathbf{Y}_{1i} = \sqrt{\mathbf{D}_i}\mathbf{Z}_{1i} + \mathbf{M}_i, \tag{14}$$

$$\mathbf{Y}_{2i} = \sqrt{\mathbf{D}_i}\mathbf{Z}_{2i} + \mathbf{M}_i. \tag{15}$$

5. Vectors $\mathbf{Y}_{1i}$ and $\mathbf{Y}_{2i}$ represent two estimates of item parameters from two populations with identical item parameters. In other words, these vectors may represent item parameter estimates for the focal and reference groups when there is no DIF. Any differences in these estimates must be due to sampling error. Therefore, an NCDIF index for item i can be obtained with the help of the two $\mathbf{Y}$ vectors and the estimates of thetas for the focal group using the equations previously described.
6. Steps 1–5 can be replicated as many times as desired.
7. The NCDIF values obtained from all replications can be rank ordered, and the 90th, 95th, 99th, 99.5th, and 99.9th percentile rank scores are recorded to establish the cutoff values for alpha levels at .10, .05, .01, .005, and .001, respectively.
8. Once the alpha level is chosen, the cutoff associated with it is used as the cutoff for assessing statistical significance of the initial NCDIF value obtained for item i.
9. This process is repeated for all items in the test, thus potentially resulting in different cutoff values for different items.

Oshima et al. (2006) noted several distinctions between the new IPR method and the method used to generate cutoff values in previous research (e.g., Bolt, 2002; Fleer, 1993; Flowers et al., 1999; Raju et al., 1995). First, a large number of replications of item parameters was generated

from the initial set of estimates obtained from the IRT calibration program. This eliminated the need for extra calibrations of item parameters, which is one of the most time-consuming aspects of the previous method. Further, this offered a theoretical advantage in the sense that it is tailored to a particular data set. As such, other unknown factors that may influence the error associated with parameter estimation were taken into account. Second, the distribution of NCDIF values was obtained for each item on a test, and so it was possible to generate a cutoff value for each item. Lastly, the IPR method is easier to use from a practitioner's standpoint because the procedure is implemented within a computer program such as DFIT7[1] (Raju, Oshima, & Wolach, 2005). The only task practitioners have is to provide estimates of item parameters, their sampling variances and covariances, and ability estimates for the focal (or reference) group.

A Monte Carlo study was conducted to examine the efficacy of the IPR method in generating cutoff values for dichotomous items (Oshima et al., 2006). Overall results suggested the IPR method performed well and provided a practical means of assessing differential functioning within the DFIT framework. The next section examines the IPR method with polytomous data.

Monte Carlo Simulation

A Monte Carlo simulation was conducted to evaluate the performance of the IPR-based NCDIF significance test. The simulated test consisted of 20 items, each having 5 response categories. The simulation was repeated under three sample size conditions: two with an equal sample size in the reference and focal groups ($n = 1{,}000$ or $n = 500$ in each group) and one with an unequal sample size (focal group $n = 1{,}000$, reference group $n = 500$).

The DIF analyses are often of greatest interest in situations where scores differ across groups, differences which may be attributed either to DIF or true differences in the underlying latent variable. Group differences in the latent variable are often referred to as impact. One set of simulations with no impact was conducted; that is, the distribution of θ was identical for the focal and reference groups. A second set of simulations was conducted with a true group difference of 0.5 *SD*. The two factors (sample size and impact) were fully crossed, resulting in six conditions. For each condition, 100 samples were generated and analyzed as described below.

Data Generation

Item responses were generated using a Fortran program created for this research. For each of n observations in each group, a person (θ) parameter was randomly generated from a normal distribution with $\sigma = 1$ and mean determined by the condition. In the no-impact condition, the mean θ was 0 for both reference and focal groups. In the impact condition, the mean θ was 0 for the reference group and –0.5 for the focal group.

Item responses were then generated according to Samejima's (1969) GRM. The item parameters were taken from Flowers et al. (1999)[2] and are shown in Table 1. In the 20-item test, 4 items with DIF were embedded. The DIF was modeled by adding a constant to the a and/or b parameters of the focal group. All DIF items favored the reference group, which had higher a and/or lower b parameters. Items 3 and 13 demonstrated uniform DIF, with differences on the b parameters of 1.0 and 0.5, respectively. Items 8 and 18 had nonuniform DIF. Item 8 had a 0.5 difference on both a and b parameters, whereas on Item 18 there was a 0.5 difference on the a parameter only. These items represent a wide range of magnitudes of DIF, as indicated by the true NCDIF values in Table 1. True NCDIF was computed according to equation (7) using the known population parameters for each item.

Table 1
Item Parameters for the Simulated 20-Item Test

	Reference Group					Focal Group[a]					True
Item	a	b_1	b_2	b_3	b_4	a	b_1	b_2	b_3	b_4	NCDIF
1	0.55	–1.80	–0.60	0.60	1.80						
2	0.73	–2.32	–1.12	0.08	1.28						
3	0.73	–1.80	–0.60	0.60	1.80	0.73	–0.80	0.40	1.60	2.80	0.47
4	0.73	–1.80	–0.60	0.60	1.80						
5	0.73	–1.28	–0.08	1.12	2.32						
6	1.00	–2.78	–1.58	–0.38	0.82						
7	1.00	–2.32	–1.12	0.08	1.28						
8	1.00	–2.32	–1.12	0.08	1.28	0.50	–1.82	–0.62	0.58	1.78	0.16
9	1.00	–1.80	–0.60	0.60	1.80						
10	1.00	–1.80	–0.60	0.60	1.80						
11	1.00	–1.80	–0.60	0.60	1.80						
12	1.00	–1.80	–0.60	0.60	1.80						
13	1.00	–1.28	–0.08	1.12	2.32	1.00	–0.78	0.42	1.62	2.82	0.13
14	1.00	–1.28	–0.08	1.12	2.32						
15	1.00	–0.82	0.38	1.58	2.78						
16	1.36	–2.32	–1.12	0.08	1.28						
17	1.36	–1.80	–0.60	0.60	1.80						
18	1.00	–1.80	–0.60	0.60	1.80	0.50	–1.80	–0.60	0.60	1.80	0.03
19	1.36	–1.28	–0.08	1.12	2.32						
20	1.80	–1.80	–0.60	0.60	1.80						

Note. NCDIF = noncompensatory differential item functioning; DIF = differential item functioning.
a. Item parameters are listed for the true DIF items only. Values not listed were the same as the item parameters used for the reference group.

Item Parameter Estimation

Item and θ parameters were estimated using PARSCALE (Muraki & Bock, 1997). As previously mentioned, the standard PARSCALE output provides only the parameter estimates and their standard errors and does not provide the parameter covariances required for the IPR method. However, it is possible to obtain parameter covariances from the SDRV matrix in the PARSCALE diagnostic output.

Further complicating matters, the PARSCALE output provides separate item and category parameters, whereas combined item-specific category boundaries are needed for the IPR method. That is, PARSCALE estimated a single location parameter for each item, b_i, as well as a set of four category parameters, c_k, reflecting the location of category boundaries. This analysis had PARSCALE estimate fixed category parameters that were the same across items, although it is also possible to estimate the category parameters separately for each item. In either case, one must transform the PARSCALE parameter estimates into a separate set of category boundaries for each item to apply the IPR method.

The item step parameter b_{ik} can be computed from the item location parameter b_i and a category parameter c_k,

$$b_{ik} = b_i - c_k. \tag{16}$$

The covariance matrix of sampling errors for the combined item parameters (b_{ik}) were estimated as functions of the covariance of item parameters and the covariance among category parameters, which were obtained from PARSCALE. For these calculations, it was assumed that the sampling errors of item parameters (a_i and b_i) were independent of the fixed category parameters (c_k). Derivations of the formulas are provided in the Appendix. The sampling variance of the item step parameter b_{ik} was computed as

$$\mathrm{VAR}(b_{ik}) = \mathrm{VAR}(b_i) + \mathrm{VAR}(c_k). \tag{17}$$

The covariance of slope with each step parameter was estimated as

$$\mathrm{COV}(a_i, b_{ik}) = \mathrm{COV}(a_i, b_i). \tag{18}$$

The covariance of the step parameters for categories k and m was estimated using

$$\mathrm{COV}(b_{ik}, b_{im}) = \mathrm{VAR}(b_i) + \mathrm{COV}(c_k, c_m). \tag{19}$$

A two-stage linking procedure was used to establish a common metric for reference and focal groups. Baker's (1993) EQUATE 2.1 program was used to link the two samples based on the test characteristic function method. Using the obtained linking coefficients, the DFIT7 (Raju et al., 2005) program computed NCDIF values for all items. The DFIT7 program provided cutoff values for NCDIF utilizing the IPR method with 1,000 replications. Based on the results, a reduced set of anchor items was created using only DIF-free items. To keep the anchor set as large as possible, only items significant at the .001 level were excluded. Using only the items deemed DIF-free from this previous step, second-stage linking coefficients were obtained from EQUATE 2.1.

Using the second-stage linking coefficients, the DFIT analysis was conducted again for all items, and items exhibiting DIF at $\alpha = .01$ were identified. Significance cutoffs for the IPR method with 1,000 replications were obtained from DFIT7. For comparison, significance was also determined using the fixed cutoff of .016 recommended by Flowers et al. (1999).

Results

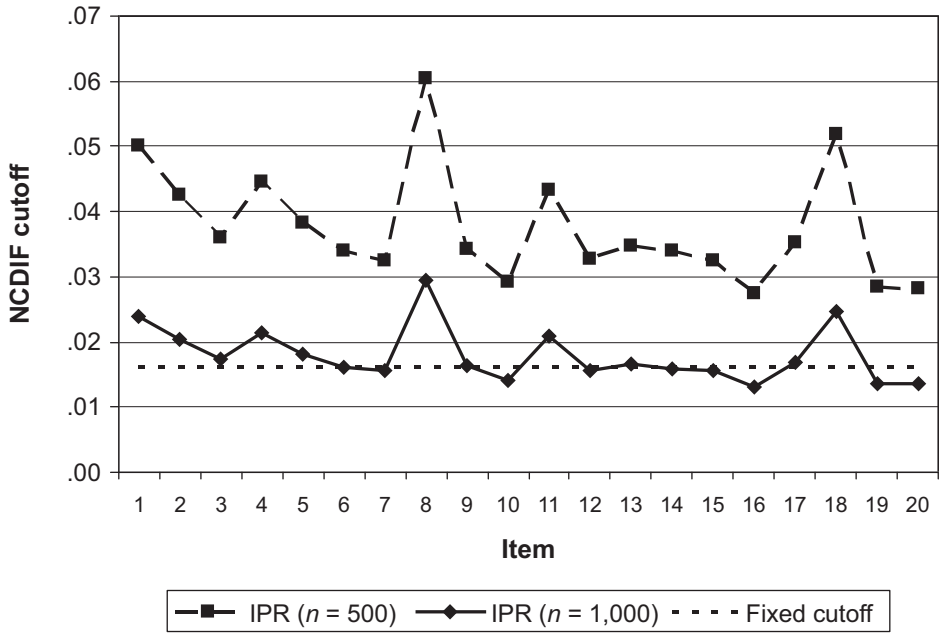
As expected, the IPR-based cutoff values varied as a function of the sample size and item characteristics. Figure 1 presents the average NCDIF cutoff at $\alpha = .01$ for each item in the no-impact conditions. Because the IPR-based cut scores are based on the focal-group covariance matrix, they are influenced only by the focal group sample size. Therefore, the cutoffs in the mixed sample size conditions were identical to those shown in Figure 1 for $n = 1{,}000$. In addition, results were nearly identical in the conditions with impact.

When the sample size in the focal group was 1,000, the average IPR-based cutoff ($\alpha = .01$) ranged from .013 to .029, with an overall mean of 0.018. This is quite similar to the .016 value obtained by Flowers et al. (1999) for the same condition. When the focal group sample size was 500, the cutoffs were substantially higher, ranging from .028 to .060 ($M = 0.038$). In addition, the cutoff values tended to be lower for items with higher a parameters ($r = -.50$, $p < .001$).

Empirical rejection rates from the simulations are presented in Table 2. For the DIF items, these reflect true positive rates or the statistical power of the test. For items with moderate to large DIF (Items 3, 8, and 13), rejection rates were close to 1.0 across conditions for both the IPR-based and fixed cutoff. These results suggest that the method is quite sensitive to this level of DIF.

When the degree of DIF was small (Item 18), true positive rates were considerably lower for the IPR method and were dependent on the sample size and the degree of impact. With a sample size of 1,000 in each group, the true positive rate was fairly high (.72) when impact was present,

Figure 1.
A Comparison of the Average IPR Method Cutoff Values for NCDIF and the Fixed Cutoff Value (.016) Proposed by Flowers et al. (1999)



Note. n = focal group sample size.
NCDIF = noncompensatory differential item functioning; IPR = item parameter replication.

but only .5 in the no-impact condition. When the sample size was 500 in each group, the true positive rate for the IPR method was less than .10 regardless of impact. True positive rates for Item 18 were considerably higher when using a fixed cutoff (0.81–1.00), particularly in the smaller sample size conditions. Thus, the fixed cutoff was considerably more sensitive than the IPR method when the magnitude of DIF was small.

False positive rates were unaffected by the presence of impact but differed as a function of the sample size condition. At the largest sample size, false positive rates were similar and somewhat conservative for both the IPR and fixed cutoff approaches. The overall rejection rate for non-DIF items ranged from .002 to .003, which is lower than the nominal $\alpha = .01$. No item had a false positive rate of more than .01 for the IPR method or more than .02 when using a fixed cutoff.

The IPR method remained conservative when the sample size was 500 in each group. False positives ranged from .003 to .004, and no item had a false positive rate of more than .02. The rejection rates for the fixed cutoff were inflated (.03–.04) relative to the nominal α, and the worst performing items exhibited rejection rates as high as .09.

When sample size was larger in the focal group than the reference group, IPR false positive rates were higher than the other conditions and were very close to the nominal α level. However, some items exhibited slightly inflated rejection rates (up to .03). Similarly, the overall false positive rate when using a fixed cutoff was close to the nominal level (.013–.014), but some items exhibited rejection rates as high as .07.

Table 2
Rejection Rates for IPR and Fixed Cutoff NCDIF Tests ($\alpha = .01$)

	$n_F = n_R = 1{,}000$		$n_F = n_R = 500$		$n_F = 1{,}000, n_R = 500$	
Item	IPR	Fixed	IPR	Fixed	IPR	Fixed
	No Impact					
3	1.00	1.00	1.00	1.00	1.00	1.00
8	1.00	1.00	0.98	1.00	1.00	1.00
13	1.00	1.00	0.99	1.00	1.00	1.00
18	0.50	0.99	0.01	0.94	0.17	0.81
Average false positive rate	0.002	0.003	0.004	0.029	0.012	0.014
Max false positive rate	0.01	0.01	0.02	0.09	0.03	0.05
	Impact (focal group mean 0.5 *SD* below reference group mean)					
3	1.00	1.00	1.00	1.00	1.00	1.00
8	1.00	1.00	0.93	1.00	1.00	1.00
13	1.00	1.00	1.00	1.00	1.00	1.00
18	0.72	1.00	0.08	0.97	0.38	0.85
Average false positive rate	0.002	0.003	0.003	0.040	0.010	0.013
Max false positive rate	0.01	0.02	0.02	0.08	0.03	0.07

Note. Max false positive rate refers to the highest rejection rate among the non-DIF items. IPR = item parameter replication.

Discussion

The current investigation extended the IPR method for determining the NCDIF cutoff values in the DFIT framework to polytomous items. Monte Carlo simulations supported the efficacy of the IPR procedure in detecting DIF at sample sizes of both 500 and 1,000 per group, while maintaining good control over false positive rates. The procedure was very sensitive to both uniform and nonuniform DIF, as long as the magnitude of DIF was not extremely small. The IPR method was less effective at detecting DIF in Item 18 than the other items. This is likely due to the small magnitude of DIF on this item.

The IPR method tended to produce cutoff scores that were higher than the fixed cutoff recommended by Flowers et al. (1999). Consequently, when the magnitude of DIF was small, the fixed cutoff method was considerably more sensitive than the IPR method. However, this greater sensitivity came at the cost of inflated false positive rates under some conditions. When the sample size in both groups was 1,000, false positive rates for the fixed cutoff were similar to those of the IPR method. This is not surprising, as the fixed cutoff was generated by Flowers et al. (1999) using simulations of essentially the same condition. When the sample size in one or both groups was smaller, overall false positive rates were two to three times higher than the nominal alpha level of .01, and some items exhibited false positive rates as high as .09. These results reinforce the need to identify significance cutoffs that are tailored to the characteristics of the data set, as with the IPR method.

Although the results are generally supportive of the IPR method, they also suggest some potential areas where the procedure might be improved. First, the false positive rates for the IPR method

tended to be conservative. When sample sizes in the two groups were equal, the false positive rates were only 20%–40% of the nominal alpha level. This suggests that the cutoffs could be lowered somewhat, resulting in greater statistical power although maintaining the desired false positive rate.

It is also worth noting that because the IPR-based cutoffs are based on the focal-group covariance matrix they are unaffected by the sample size of the reference group. If the sample sizes are unequal across groups, the smaller or larger n for the reference group will affect the sampling variance of the NCDIF statistic, but this will not be reflected in the cutoff value. Despite this, our simulations demonstrate that the IPR-based cutoff values are effective even with substantially different sample sizes ($n_F = 1{,}000$, $n_R = 500$). Similar results were found in prior research on the IPR method with dichotomous items (Oshima et al., 2006). Still, the false positive rates under the mixed sample size condition tended to be higher than when sample sizes were equal, and this may become problematic when the sample size difference is more extreme. Future work should refine the IPR method to incorporate sample size information from both groups. Until this issue has been resolved, it is recommended that the IPR method, as implemented in DFIT7, be used only when sample sizes are roughly the same in reference and focal groups.

The current study examined DFIT analysis in the context of the rating scale model with fixed category parameters. That is, the distance between adjacent categories was assumed to be the same across all items. However, it is important to note that the DFIT framework can also readily accommodate situations in which category parameters are allowed to vary across items. In addition, DFIT statistics are based on the expected score functions, which can be obtained from any parameterization of the IRT model. Thus, the DFIT analysis can be generalized to work with the generalized partial credit model (Muraki, 1992) or the nominal response model (Bock, 1972).

A challenge in applying the IPR method is obtaining item-parameter covariances from existing IRT software. To our knowledge, no current software provides this information as part of the standard output. From the diagnostic output of PARSCALE the necessary information was obtained, but doing so required considerable effort and hand calculation. The authors hope that vendors of IRT software will make this information more readily available in the future. Alternatively, analytic solutions may provide means to compute covariances directly from item parameters (Li & Lissitz, 2004; Morris, Fortmann, & Oshima, 2007).

The need for sound DIF techniques for polytomous items has increased recently, and it is expected to continue. Due to the No Child Left Behind Act of 2001, standardized assessments have been mandated across the United States. As a result, there has been a dramatic increase in the use of polytomous IRT models to accommodate constructed response types of items. These types of items allow the assessment to accumulate information quickly although, at the same time, measuring more complex aspects of the construct of interest. Furthermore, there are many applications of polytomous DIF techniques in psychological assessments, including assessing measurement equivalence. It is likely that the performance of polytomous DFIT will be enhanced with the new significance test. The enhanced DIFT would be a useful tool for practitioners in educational and psychological measurement.

Appendix

The item parameter sampling error covariance matrix required by the differential functioning of items and tests (DFIT) program includes a separate item step parameter b_{ik} for each BRF of each item. However, the diagnostic output of PARSCALE provides two separate covariance matrices: one covariance matrix with a discrimination parameter (a_i) and a single location parameter (b_i) for

each item and a second covariance matrix of the category parameters (c_k). Therefore, it was necessary to integrate information from these two types of matrices. No information on the covariance of category parameters with a or b parameters could be located. Therefore, these covariances were assumed to be 0.

The item step parameter b_{ik} can be computed from the item location parameter b_i and a category parameter c_k, where $b_{ik} = b_i - c_k$. Therefore, the variance of the item step parameter b_{ik} is

$$\mathrm{VAR}(b_{ik}) = \mathrm{VAR}(b_i) + \mathrm{VAR}(c_k) - 2\mathrm{COV}(b_i, c_k). \tag{A1}$$

Assuming COV $(b_i, c_k) = 0$,

$$\mathrm{VAR}(b_{ik}) = \mathrm{VAR}(b_i) + \mathrm{VAR}(c_k). \tag{A2}$$

The covariance of the slope with each step parameter may be derived as follows:

$$\mathrm{COV}(a_i, b_{ik}) = E[a_i(b_i - c_k)] - E(a_i)E(b_i - c_k). \tag{A3}$$
$$\mathrm{COV}(a_i, b_{ik}) = E(a_i b_i) - E(a_i c_k) - E(a_i)E(b_i) + E(a_i)E(c_k), \tag{A4}$$
$$\mathrm{COV}(a_i, b_{ik}) = \mathrm{COV}(a_i b_i) - \mathrm{COV}(a_i, c_k). \tag{A5}$$

Assuming $\mathrm{COV}(a_i, c_k) = 0$,

$$\mathrm{COV}(a_i, b_{ik}) = \mathrm{COV}(a_i, b_i). \tag{A6}$$

The covariance of the step parameters for categories k and m is

$$\mathrm{COV}(b_{ik}, b_{im}) = E[(b_i - c_k)(b_i - c_m)] - E(b_i - c_k)E(b_i - c_m), \tag{A7}$$
$$\mathrm{COV}(b_{ik}, b_{im}) = E\left(b_i^2\right) - E(b_i c_k) - E(b_i c_m) + E(c_k c_m) - \left[E(b_i)^2 - E(b_i)E(c_k) - E(b_i)E(c_m) + E(c_k)E(c_m)\right], \tag{A8}$$
$$\mathrm{COV}(b_{ik}, b_{im}) = \left[E\left(b_i^2\right) - E(b_i)^2\right] - [E(b_i c_k) - E(b_i)E(c_k)] - [E(b_i c_m) - E(b_i)E(c_m)] + [E(c_k c_m) - E(c_k)E(c_m)], \tag{A9}$$
$$\mathrm{COV}(b_{ik}, b_{im}) = \mathrm{VAR}(b_i) - \mathrm{COV}(b_i, c_k) - \mathrm{COV}(b_i, c_m) + \mathrm{COV}(c_k, c_m). \tag{A10}$$

Assuming $\mathrm{COV}(b_i, c_k) = 0$ for all k,

$$\mathrm{COV}(b_{ik}, b_{im}) = \mathrm{VAR}(b_i) + \mathrm{COV}(c_k, c_m). \tag{A11}$$

Equations (A2), (A6), and (A11) were used accordingly to estimate the parameter covariance in equation (9).

Notes

1. Commercial release of the DFIT7 software is expected to occur soon. Contact John Scott (JScott@appliedpsych.com) for information on obtaining the program.
2. Item 18 was modified for the current study to produce a slightly greater magnitude of DIF (NCDIF = .03). The a parameters used in Flowers et al. (1999) were 1.36 and 0.86 for the reference and focal groups, respectively, which produced expected score functions that were nearly identical (NCDIF = .003).

References

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.

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 1 error rates of differential functioning of items and tests. 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.

Drasgow, F., & Kanfer, R. (1985). Equivalence of psychological measurement in heterogeneous populations. *Journal of Applied Psychology*, *70*, 662-680.

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

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.

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

Li, Y. H., & Lissitz, R. W. (2004). Applications of the analytically derived asymptotic standard errors of item response theory item parameter estimates. *Journal of Educational Measurement*, *41*, 85-117.

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

Morris, S. B., Fortmann, K. A., & Oshima, T. C. (2007, April). *An evaluation of the item parameter replication method for DFIT analysis of polytomous items.* Paper presented at the annual conference of the National Council on Measurement in Education, Chicago.

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

Muraki, E., & Bock, R. D. (1997). PARSCALE: IRT based test scoring and item analysis for graded, open-ended exercises and performance tasks [Computer software]. Chicago: Scientific Software International.

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

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

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., Oshima, T. C., Fortmann, K., Nering, M., & Kim, W. (2006, February). *The new significance test for Raju's polytomous DFIT.* Poster session presented at the Georgia Institute of Technology New Directions in Psychological Measurement with Model-Based Approaches conference, Atlanta, GA.

Raju, N. S., Oshima, T. C., & Wolach, A. (2005). Differential functioning of items and tests (DFIT): Dichotomous and polytomous [Computer program]. Chicago: Illinois Institute of Technology.

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.

Samejima, F. (1969). *Estimation of latent ability using a response pattern of graded scores* (Psychometric Monograph No. 17). Iowa City, IA: Psychometric Society.

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.

Zickar, M. J. (2002). Modeling data with polytomous item response theory. In F. Drasgow & N. Schmitt (Eds.), *Measuring and analyzing behavior in organizations: Advances in measurement and data analysis* (pp. 123-155). San Francisco: Jossey-Bass.

Authors' Address

This article presents the work of Dr. Nambury S. Raju, who suddenly passed away in October 2005 while working on this research. His tireless contribution to the field of psychometrics will be always remembered. This article is dedicated to Dr. Raju and his family. All other authors are listed in alphabetical order. Portions of the manuscript were presented at the New Directions in Psychological Measurement With Model-Based Approaches Conference (February 2006, Atlanta, GA) and the Society for Industrial and Organizational Psychology Conference (April 2006, Dallas, TX). Correspondence regarding this article should be directed to Scott B. Morris, Institute of Psychology, Illinois Institute of Technology, 3101 S. Dearborn, Chicago, IL 60616; e-mail: scott.morris@iit.edu.