Applied Psychological Measurement

http://apm.sagepub.com

IRT-Based Internal Measures of Differential Functioning of Items and Tests

Nambury S. Roju, Wim J. van der Linden and Paul F. Fleer

Applied Psychological Measurement 1995; 19; 353
DOI: 10.1177/014662169501900405

The online version of this article can be found at:
http://apm.sagepub.com/cgi/content/abstract/19/4/353

Published by:

http://www.sagepublications.com

Additional services and information for *Applied Psychological Measurement* can be found at:

Email Alerts: http://apm.sagepub.com/cgi/alerts

Subscriptions: http://apm.sagepub.com/subscriptions

Reprints: http://www.sagepub.com/journalsReprints.nav

Permissions: http://www.sagepub.com/journalsPermissions.nav

Citations (this article cites 11 articles hosted on the
SAGE Journals Online and HighWire Press platforms):
http://apm.sagepub.com/cgi/content/refs/19/4/353

Downloaded from http://apm.sagepub.com at GEORGIA STATE UNIVERSITY on January 18, 2008
© 1995 SAGE Publications. All rights reserved. Not for commercial use or unauthorized distribution.

IRT-Based Internal Measures of Differential Functioning of Items and Tests

Nambury S. Raju, Illinois Institute of Technology
Wim J. van der Linden, University of Twente
Paul F. Fleer, Illinois Institute of Technology

Internal measures of differential functioning of items and tests (DFIT) based on item response theory (IRT) are proposed. Within the DFIT context, the new differential test functioning (DTF) index leads to two new measures of differential item functioning (DIF) with the following properties: (1) The compensatory DIF (CDIF) indexes for all items in a test sum to the DTF index for that test and, unlike current DIF procedures, the CDIF index for an item does not assume that the other items in the test are unbiased; (2) the noncompensatory DIF (NCDIF) index, which assumes that the other items in the test are unbiased, is comparable to some of the IRT-based DIF indexes; and (3) CDIF and NCDIF, as well as DTF, are equally valid for polytomous and multidimensional IRT models. Monte carlo study results, comparing these indexes with Lord's χ^2 test, the signed area measure, and the unsigned area measure, demonstrate that the DFIT framework is accurate in assessing DTF, CDIF, and NCDIF. *Index Terms: area measures of DIF, compensatory DIF, differential functioning of items and tests (DFIT), differential item functioning, differential test functioning, Lord's χ^2, noncompensatory DIF, nonuniform DIF, uniform DIF.*

Differential item functioning (DIF) continues to receive significant attention among measurement specialists and practitioners. Several DIF techniques are currently available for determining whether an item is functioning differently in two groups (e.g., black vs. white, female vs. male). Techniques such as the area between two item response functions (IRFs; Kim & Cohen, 1991; Raju, 1988, 1990; Rudner, Geston, & Knight, 1980), Lord's (1980) χ^2 test, and Thissen, Steinberg, & Wainer's (1988) likelihood ratio tests are based on item response theory (IRT), whereas the Mantel-Haenszel (MH) technique (Holland & Thayer, 1988) and the delta method (Angoff & Ford, 1973) do not use IRT. These and other currently available DIF techniques (e.g., Dorans, 1986; Mellenbergh, 1982; Scheuneman, 1979; Shepard, Camilli, & Averill, 1981; Swaminathan & Rogers, 1990) function at the item level; that is, they identify items that have significant DIF or function differently in two groups. A comprehensive review of DIF methods can be found in Millsap & Everson (1993).

In practice, test developers typically either exclude an item with significant DIF (also referred to as "bias") from the final test or modify it so that it no longer exhibits significant DIF and include it in the final test. Although the removal of items with significant DIF is expected to result in a test that is fair to (or unbiased for) various racial, ethnic, and gender subgroups, until recently (Shealy & Stout, 1993) a psychometric measure of differential functioning for an entire test was not available. With an appropriately defined measure of differential test functioning (DTF), it would be possible to determine the effect of removing or adding items with significant DIF on the DTF of the final test.

It is also desirable to have a definition of DIF such that individual DIF values sum to the total test DTF for a given set of items. Current measures of DIF do not possess such an additive property. Therefore, one purpose of this study was to define such IRT-based measures of DTF and DIF within the context of a defini-

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 19, No. 4, December 1995, pp. 353–368
© Copyright 1995 Applied Psychological Measurement Inc.
0146-6216/95/040353-16$2.05

Downloaded from http://apm.sagepub.com at GEORGIA STATE UNIVERSITY on January 18, 2008
© 1995 SAGE Publications. All rights reserved. Not for commercial use or unauthorized distribution.

tion of differential functioning of items and tests (DFIT) originally proposed by Raju, van der Linden, & Fleer (1992).

Measures of Differential Functioning of Items and Tests

Differential Test Functioning

Let $P_i(\theta_s)$ represent the probability of success for examinee s with trait level θ on item i. P_i can be represented either by a one-, two-, or three-parameter logistic model or the normal ogive model (Lord, 1980, pp. 12–13), with item parameters a (discrimination), b (difficulty), and c (pseudoguessing). Let the test consist of n items and have one set of item parameters for each of two groups—the reference (R) group (typically the majority group) and the focal (F) group (typically the minority group). Also assume that the two sets of item parameters are on a common scale. $P_{iR}(\theta_s)$ represents the probability of success on item i at a given θ level for examinee s if examinee s is a member of the reference group; $P_{iF}(\theta_s)$ represents the probability of success on the same item for examinee s if examinee s is a member of the focal group. If an item functions differently in the two groups, P_{iR} and P_{iF} will be different for some examinees.

Within IRT, an examinee's expected proportion correct (EPC, sometimes referred to as the "true" score) can be expressed as

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

Here, each examinee has two EPCs—one as a member of the focal group (T_{sF}) and the other as a member of the reference group (T_{sR}). If $T_{sR} = T_{sF}$, then the examinee's EPC is independent of group membership. The greater the difference between T_{sR} and T_{sF}, the greater the differential functioning of a test. A measure of DTF at the examinee level may be defined as $(T_{sF} - T_{sR})^2$. Therefore, an overall measure of DTF across examinees is

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

where the expectation ($\in$) can be taken over the reference group or the focal group. If it is assumed that the expectation is taken over the focal group, Equation 2 can be rewritten as

$$\text{DTF} = \underset{F}{\in}(T_{sF} - T_{sR})^2. \tag{3}$$

Letting $D_s = T_{sF} - T_{sR}$, Equation 3 can be rewritten as

$$\text{DTF} = \underset{F}{\in} D_s^2 = \int_{\theta} D_s^2 f_F(\theta)\,d\theta = \sigma_D^2 + (\mu_{TF} - \mu_{TR})^2 = \sigma_D^2 + \mu_D^2, \tag{4}$$

where $f_F(\theta)$ is the density function of θ in the focal group, and μ_{TF} and μ_{TR} represent the mean EPC of examinees in the focal and reference groups, respectively.

Differential Item Functioning

A compensatory DIF index. Based on Equation 1, Equation 2 can be rewritten as

$$\text{DTF} = \in\left[\left(\sum_{i=1}^{n} d_{is}\right)^2\right], \tag{5}$$

where $d_{is} = P_{iF}(\theta_s) - P_{iR}(\theta_s)$. Equation 5 can be rewritten as

$$\text{DTF} = \sum_{i=1}^{n}\left[\text{Cov}(d_i, D) + \mu_{d_i}\mu_D\right], \tag{6}$$

Downloaded from http://apm.sagepub.com at GEORGIA STATE UNIVERSITY on January 18, 2008
© 1995 SAGE Publications. All rights reserved. Not for commercial use or unauthorized distribution.

where $\mathrm{Cov}(d_i, D)$ is the covariance between the difference in item probabilities for item i (d_i) and the difference between the two EPCs (D), and μ_{d_i} and μ_D are the mean of d_{is} and D_s, respectively. Differential functioning at the item level is now defined as

$$\mathrm{CDIF}_i = \in(d_i D) = \mathrm{Cov}(d_i, D) + \mu_{d_i}\mu_D. \tag{7}$$

This definition of DIF will hereafter be referred to as compensatory DIF (CDIF) to distinguish it from noncompensatory DIF (NCDIF), which is defined below. Combining Equations 6 and 7 yields

$$\mathrm{DTF} = \sum_{i=1}^{n} \mathrm{CDIF}_i, \tag{8}$$

which shows that the definition of CDIF_i (given in Equation 7) is additive in the sense that differential functioning at the test level is simply the sum of differential functioning at the item level, and which indicates how much each item's CDIF contributes to DTF. Furthermore, rewriting Equation 5 yields

$$\mathrm{DTF} = \in\left[\sum_{i=1}^{n}\left(P_{iF} - P_{iR}\right)\right]^2 = \in\left[\left(P_{1F} - P_{1R}\right) + \left(P_{2F} - P_{2R}\right) + \ldots + \left(P_{nF} - P_{nR}\right)\right]^2. \tag{9}$$

This equation shows the compensating nature of the proposed index. For example, if $P_{6F} - P_{6R} = -.3$ and $P_{7F} - P_{7R} = +.3$ for a given examinee, then the bias in Item 6 cancels the bias in Item 7 and the two items together contribute a sum of 0 to the examinee's DTF. That is, the proposed DTF index takes into account compensating bias across items at the examinee level. In addition, Equation 8 shows the nature of compensating bias across items at the group level. From a practitioner's point of view, this is a useful feature because it enables the practitioner to assess not only which items have compensating bias and which items to delete due to bias, but also to estimate the net effect of item deletion on DTF. When items with significant CDIF are deleted from the final test, the revised DTF can be computed for the retained items using Equation 4.

There is an important difference between the definition of CDIF given in Equation 7 and other definitions of DIF. As noted above, previous definitions of DIF are item level indexes and therefore do not take into account correlated DIF or item bias that may be inherent in a set of items. DFIT, however, begins with a definition of DTF and then decomposes DTF into differential functioning at the item level (CDIF). Hence, it is not surprising that the definition of CDIF given in Equation 7 includes information about bias from other items in the test. In practice, it is possible that two items with significant CDIF may be quite similar because the stems for the two items are very similar in phrasing or in represented content. In such cases, bias in the two items may have a nonzero correlation that, in turn, will influence differential functioning at the test level.

A noncompensatory DIF index. If it is assumed that all items in the test other than item i are completely unbiased, then it must be true that $d_j = 0$ for all $j \neq i$. Equation 7 can be rewritten as

$$\mathrm{NCDIF}_i = \sigma^2_{d_i} + \mu^2_{d_i}, \tag{10}$$

which does not include information about bias from other items. Three aspects of NCDIF are considered.

First, because d_i was defined above as the difference in item probabilities for item i, NCDIF = 0 if and only if the item parameters for item i are equal for both the focal and reference groups. Lord's (1980) χ^2 test (LC) offers a test of the null hypothesis that the two sets of item parameters are identical. Therefore, LC may be viewed as a test of the hypothesis that $\mathrm{NCDIF}_i = 0$, and LC and Equation 10 may be considered comparable in the sense of providing similar information about DIF.

Second, by letting $f_F(\theta)$ denote the density function of θ in the focal group, Equation 10 can be rewritten as

$$\mathrm{NCDIF}_i = \int_{-\infty}^{\infty}\left[P_{iF}(\theta) - P_{iR}(\theta)\right]^2 f_F(\theta)\mathrm{d}\theta. \tag{11}$$

Downloaded from http://apm.sagepub.com at GEORGIA STATE UNIVERSITY on January 18, 2008
© 1995 SAGE Publications. All rights reserved. Not for commercial use or unauthorized distribution.

This is identical to a definition of DIF recently offered by Wainer (1993).

Third, Equation 11 can be rewritten as

$$\text{NCDIF}_i = \int_{-\infty}^{\infty} \left|P_{iF}(\theta) - P_{iR}(\theta)\right|^2 f_F(\theta)\,d\theta, \tag{12}$$

which, according to the Cauchy-Schwartz inequality, can be expressed as

$$\text{NCDIF}_i \geq \left[\int_{-\infty}^{\infty} \left|P_{iF}(\theta) - P_{iR}(\theta)\right| f_F(\theta)\,d\theta\right]^2. \tag{13}$$

Raju (1988) noted that if $f_F(\theta)$ is rectangular, then the right-hand side of Equation 13 is the square of the absolute or unsigned area between two IRFs. Therefore, the unsigned area definition of DIF may also be viewed as a special case of Equation 10.

The proposed NCDIF index, therefore, appears to be closely related to many of the current methods for assessing DIF within the IRT context. This special case, however, assumes that all items in the test, other than the item under investigation, do not function differentially. This assumption is not likely to be satisfied in most test development situations. NCDIF is noncompensatory because its value for an item can only be non-negative; therefore, it cannot cancel or compensate across items.

Practical Applications of the DFIT Framework

DTF, CDIF, and NCDIF can be useful in practice. The question of which index is more important depends on the purpose. When total test scores are used for determining the effectiveness of an instructional program or for placement and/or selection, DTF is likely to be more valuable than CDIF or NCDIF. CDIF is useful if a test developer is forced to include (for content or other reasons) items with significant DIF in a test in which some items favor the focal group and some favor the reference group. The effect on DTF of counterbalancing items with significant DIF indexes provides critical information because it is generally difficult to exclude all items with significant DIF from the final version of a test. Other DIF procedures do not provide such information. However, if there is concern about the potential offensiveness of certain test items to certain groups of individuals or about why certain types of items are more biased than others, NCDIF is likely to be more valuable than DTF and CDIF. Therefore, it is helpful to have access to all three types of information for a given test. The DFIT framework provides such information whereas other procedures offer only information similar to that provided by NCDIF.

Finally, note that the DFIT framework is equally valid for polytomous and multidimensional datasets. Applications of the DFIT framework to polytomous and multidimensional tests are described in Flowers, Oshima, & Raju (1995), Oshima, Raju, & Flowers (1993), and Oshima, Raju, Flowers, & Monaco (1995).

Significance Tests for DTF and NCDIF

Although DTF, CDIF, and NCDIF are defined in terms of true parameters, only estimates of θ, a, b, and c (denoted $\hat{\theta}$, $\hat{a}$, $\hat{b}$, and $\hat{c}$, respectively) are available. Therefore, the proposed DTF, CDIF, and NCDIF indexes are computed using estimated person and item parameters in practice. Estimates of DTF, CDIF, and NCDIF (denoted $\widehat{\text{DTF}}$, $\widehat{\text{CDIF}}$, and $\widehat{\text{NCDIF}}$, respectively) are computed using D_s and d_{is} for examinee s. That is,

$$D_s = \sum_{i=1}^{n} d_{is}, \tag{14}$$

$$d_{is} = \hat{P}_{iF} - \hat{P}_{iR}, \tag{15}$$

$$\widehat{\text{DTF}} = \hat{\sigma}_D^2 + \hat{\mu}_D^2, \tag{16}$$

Downloaded from http://apm.sagepub.com at GEORGIA STATE UNIVERSITY on January 18, 2008
© 1995 SAGE Publications. All rights reserved. Not for commercial use or unauthorized distribution.

$$\widehat{\text{CDIF}}_i = \widehat{\text{Cov}}(d_i, D) + \hat{\mu}_{d_i}\hat{\mu}_D, \tag{17}$$

and

$$\widehat{\text{NCDIF}}_i = \hat{\sigma}^2_{d_i} + \hat{\mu}^2_{d_i}, \tag{18}$$

where $\hat{P}_{iF}$ and $\hat{P}_{iR}$ are item probabilities computed using estimated person and item parameters, and $\hat{\sigma}$, $\hat{\mu}$, and $\widehat{\text{Cov}}$ represent the unbiased estimates of σ, μ, and Cov, respectively.

According to the above definitions, $\widehat{\text{DTF}}$, $\widehat{\text{CDIF}}$, and $\widehat{\text{NCDIF}}$ have two distinct sources of error: (1) estimation error resulting from the use of person and item parameter estimates, and (2) sampling error resulting from using a sample from a population of examinees. If true person and item parameters were known, $\widehat{\text{DTF}}$, $\widehat{\text{CDIF}}$, and $\widehat{\text{NCDIF}}$ would include only sampling error. In that case, $D_s = 0$ with probability 1 for all s in the focal group when the null condition (i.e., DTF = 0) is true. Using estimated person and item parameters in the computation of DTF, CDIF, and NCDIF, it is very unlikely for $D_s = 0$ for all s, even under the null condition. Hopefully, future research will be successful in proposing significance tests that fully account for the errors associated with the estimation of person and item parameters. Note, however, that although the present approach is not ideal, such approaches are currently used in the unidimensional case in computing standard errors for person and item parameters by LC for DIF and by the significance tests for the exact signed area (ESA) and exact unsigned area (EUA) measures (Raju, 1990).

χ^2 test for $\widehat{\text{DTF}}$. Assuming that D is normally distributed with a mean of μ_D and a finite standard deviation of σ_D, for examinee s

$$z_s = \frac{D_s - \mu_D}{\sigma_D}. \tag{19}$$

Because it is well known that z_s^2 has a χ^2 distribution with 1 degree of freedom (df), the sum of z_s^2 across N_F examinees in the focal group has a χ^2 distribution with N_F df, where N_F is the focal group sample size and N_R is the reference group sample size. Algebraically, this can be expressed as

$$\chi^2_{N_F} = \sum_{s=1}^{N_F} z_s^2 = \frac{\sum_{s=1}^{N_F}(D_s - \mu_D)^2}{\sigma_D^2}. \tag{20}$$

In the present context, the interest is in minimizing the expectation of $\widehat{\text{DTF}}$ or approaching

$$\in(\widehat{\text{DTF}}) = \mu_{D^2} = 0, \tag{21}$$

which implies that μ_D also must be 0. Note that $\mu_D = 0$ is a necessary but not sufficient condition for the validity of Equation 21. Substituting $\mu_D = 0$ into Equation 20 yields

$$\chi^2_{N_F} = \frac{\sum_{s=1}^{N_F} D_s^2}{\sigma_D^2}, \tag{22}$$

which, according to the definition of $\widehat{\text{DTF}}$ for N_F examinees (Equation 4), can be expressed as

$$\chi^2_{N_F} = \frac{N_F(\widehat{\text{DTF}})}{\sigma_D^2}. \tag{23}$$

Substituting the sample-based estimate of the variance of D, Equation 23 can be rewritten as

Downloaded from http://apm.sagepub.com at GEORGIA STATE UNIVERSITY on January 18, 2008
© 1995 SAGE Publications. All rights reserved. Not for commercial use or unauthorized distribution.

$$\chi^2_{N_F} = \frac{N_F(\widehat{\mathrm{DTF}})}{\hat{\sigma}^2_D}. \tag{24}$$

Because of the sample-based estimate of the variance of D in Equation 24, the *df* for this χ^2 is probably less than N_F. This χ^2 test may prove useful in determining if an observed (or sample-based) DTF is significantly different from 0.

t test for $\widehat{DTF}$. Another statistical test that may prove useful in the present context is the t test, which can be expressed as

$$t = \frac{(N_F)^{1/2}(\hat{\mu}_D - \mu_D)}{\hat{\sigma}_D}. \tag{25}$$

Under the null hypothesis that $\mu_D = 0$, Equation 25 can be rewritten as

$$t = \frac{(N_F)^{1/2}(\hat{\mu}_D)}{\hat{\sigma}_D}. \tag{26}$$

According to the previously stated assumptions about the distribution of D and the asymptotic normality of $\hat{\mu}_D$ with variance equal to σ^2_D/N_F, Equation 26 is expected to have an asymptotic t distribution with $N_F - 1$ *df*. Because the *N*s are generally large in IRT analyses, t and χ^2 tests are likely to lead to very similar conclusions.

When an observed $\widehat{\mathrm{DTF}}$ is statistically significant, the search for items that may be causing the significant t or χ^2 can begin. After identifying and removing such items from the test, $\widehat{\mathrm{DTF}}$ and its χ^2 should be recomputed with the remaining items. Because the value for $\widehat{\mathrm{Cov}}(d_i, D)$ depends on, among other things, the number of items that are still in the test, it is recommended that a single item at a time be selected and that the procedure be continued until the χ^2 associated with the revised index becomes nonsignificant. Because $\widehat{\mathrm{CDIF}}$s sum to the total test $\widehat{\mathrm{DTF}}$, when a given $\widehat{\mathrm{DTF}}$ is found statistically significant, items with large, positive $\widehat{\mathrm{CDIF}}$ should be deleted, one item at a time, until $\widehat{\mathrm{DTF}}$ based on the remaining items is statistically nonsignificant. All deleted items should be labeled "biased" or characterized as having significant $\widehat{\mathrm{CDIF}}$. No separate significant test, therefore, is proposed for $\widehat{\mathrm{CDIF}}$. Because this sequential deletion of items is likely to capitalize on chance, a cross-validation of the end result is desirable whenever the sample is of adequate size for such an analysis.

χ^2 and t tests for $\widehat{NCDIF}$. Based on the significance test defined above for $\widehat{\mathrm{DTF}}$, a χ^2 significance test, given that d_i is normally distributed with a finite variance, may be similarly defined for (sample-based) $\widehat{\mathrm{NCDIF}}$ for item i as

$$\chi^2_{N_F} = \frac{N_F(\widehat{\mathrm{NCDIF}})}{\hat{\sigma}^2_{d_i}} \tag{27}$$

with N_F *df*. A t test for $\widehat{\mathrm{NCDIF}}$ is

$$t = \frac{(N_F)^{1/2}\left(\hat{\mu}_{\hat{d}_i}\right)}{\hat{\sigma}_{\hat{d}_i}} \tag{28}$$

with $N_F - 1$ *df*.

An exploratory monte carlo examination of the χ^2 test for $\widehat{\mathrm{DTF}}$ and $\widehat{\mathrm{NCDIF}}$ showed that these indexes were overly sensitive for large *N*s (Fleer, 1993). In a no-bias condition (i.e., identical true item parameters

Downloaded from http://apm.sagepub.com at GEORGIA STATE UNIVERSITY on January 18, 2008
© 1995 SAGE Publications. All rights reserved. Not for commercial use or unauthorized distribution.

in the focal and reference groups), the percent of items identified as biased at the .01 level of significance was substantially greater than 1%. Therefore, after several replications under the no-bias condition, Fleer found that a cut-off value of .006 for both indexes resulted in falsely identifying approximately 1% of the items as biased. Therefore, the criterion of $\leq .006$ or nonsignificant χ^2 was used with $\widehat{\text{DTF}}$ in the successive deletion of items. For $\widehat{\text{NCDIF}}$, items with $\widehat{\text{NCDIF}} > .006$ and statistically significant χ^2s were designated as differentially functioning items.

Method

Data Generation

A two-parameter logistic model (2PLM) was used to generate simulated datasets using the computer program RANGEN (Fleer, Kiley, & Raju, 1991). Item response data for two groups of equal θ—the reference group and the focal group—were generated. RANGEN was used to randomly select the values for the underlying (true) examinee θ from a normal (0,1) distribution and, for each value selected, calculate the item responses. An item response for a simulated examinee was determined by comparing the calculated probability of a correct response to an item, based on a randomly selected θ parameter and preselected item parameters, with a number sampled at random from the uniform distribution on the [0,1] interval. If the sampled number was less than the calculated probability, the simulated item response was scored as correct; otherwise, it was scored as incorrect.

Test Length and Sample Size

The simulated test consisted of 40 items. Two sample sizes were used—$N = 500$ (the small sample condition) and $N = 1{,}000$ (the large sample condition)—to allow a comparison of the effects of sample size. $N = 500$ corresponded to an apparent minimum size for relatively accurate recovery of item parameters estimated with marginal Bayesian procedures under the 2PLM (Baker, 1990; Cohen & Kim, 1993; Kim & Cohen, 1992; Lim & Drasgow, 1990).

Generation of DIF

The data were generated to simulate four proportions of test-wide DIF (0%, 5%, 10%, and 20%). Datasets contained 0, 2, 4, or 8 differentially functioning items, depending on the proportion of DIF condition (0%, 5%, 10%, or 20%, respectively) simulated. Two conditions of test-wide DIF (unidirectional vs. bidirectional) were simulated. Because both CDIF and NCDIF were investigated, separate datasets were required to reflect a condition of bidirectional differential functioning at the test level in which items favoring one group were balanced with those favoring the other, and unidirectional test level differential functioning in which differential functioning at the item level pervasively favored the reference group over the focal group. For example, two items with unidirectional differential functioning would be considered biased according to both the CDIF and NCDIF definitions of bias. However, two items with bidirectional but balanced differential functioning would be considered biased only within the NCDIF definition of bias; that is, in view of the bidirectional, balanced definition of bias, the two items (one item favoring the reference group to the same degree as the other item favoring the focal group) would cancel each other, thus making no contribution to the total DTF. The unidirectional and bidirectional bias conditions were treated separately in this investigation in order to provide a more complete and adequate assessment of CDIF.

In addition, items were generated to simulate uniform DIF ($a_{iR} = a_{iF}$ and $b_{iR} \neq b_{iF}$) and nonuniform DIF ($a_{iR} \neq a_{iF}$ and either $b_{iR} = b_{iF}$ or $b_{iR} \neq b_{iF}$). Only the 20% proportion of DIF condition included nonuniform DIF items because it contained the largest possible number of items (i.e., 8 items) with DIF and thus allowed for a more reasonable intracondition comparison of the effects of DIF uniformity on the detection of differentially functioning items. Under the bidirectional DIF condition, the nonuniformly biased items were designed so that the

Downloaded from http://apm.sagepub.com at GEORGIA STATE UNIVERSITY on January 18, 2008
© 1995 SAGE Publications. All rights reserved. Not for commercial use or unauthorized distribution.

generating *b*s were equal in both the reference and focal groups. The intention was to produce, at once, off-setting areas inscribed by IRFs for each item (i.e., a signed area of 0.0) and (mostly) off-setting areas across specific pairs of items. (i.e., the CDIF indexes for the two items in the pair were equal in magnitude but opposite in sign).

The generating item parameter values for the unidirectional DIF condition (see Table 1) replicated those used by Cohen & Kim (1993). Four focal group datasets were generated: Focal 0 is not listed in Table 1 because it contained the same item parameters as the reference group; Focal 1 contained two uniform DIF items (Items 5 and 10); Focal 2 contained four uniform DIF items (Items 5, 10, 15, 20); Focal 3 contained four uniform DIF items (Items 5, 10, 25, and 30) and four nonuniform DIF items (Items 15, 20, 35, and 40). In this condition, uniform and nonuniform DIF items favored the reference group.

Values used for the bidirectional DIF condition (see Table 2) were based on a modified version of the unidirectional set. Again, four focal group datasets were generated: Focal 0 is not listed in Table 2 because it contained the same item parameters as the reference group; Focal 1 contained two uniform DIF items (Items 5 and 6) with Item 5 favoring the reference group and Item 6 favoring the focal group; Focal 2 contained four uniform DIF items (Items 5, 6, 15, 16) with Items 5 and 15 favoring the reference group and Items 6 and 16 favoring the focal group; Focal 3 contained four uniform DIF items (Items 25, 26, 29, and 30) with Items 25 and 29 favoring the reference group and Items 26 and 30 favoring the focal group, and four nonuniform DIF items (Items 5, 6, 15, and 16) with Items 6 and 15 favoring the reference group and Items 5 and 16 favoring the focal group.

A total of 16 ($2 \times 4 \times 2$) datasets were generated and analyzed: one for each of the sample size ($N = 500$ and $N = 1{,}000$) × proportion of DIF (0%, 5%, 10%, or 20%) × test-wide DIF (unidirectional and bidirectional) conditions.

Parameter Estimation

Item and θ parameters were estimated using the computer program PC-BILOG 3.04 (Mislevy & Bock, 1990). The program's default Bayesian procedure, MMAP, and the default priors and their hyperparameters were used to estimate 2PLM item parameters. Estimates of θ used the program's default Bayesian EAP procedure with a unit normal prior. BILOG's goodness-of-fit indexes were examined for model-data fit.

The accuracy of item and θ parameter estimates was assessed using a recovery analysis. This was conducted on each dataset by calculating the root mean squared differences (RMSDs) and product-moment correlations between generating (true) and estimated parameters.

Parameter Linking

The estimation of equating coefficients used Stocking & Lord's (1983) test characteristic curve method as implemented by the computer program EQUATE (Baker, Al-Karni, & Al-Dosary, 1991). In this study, all parameter estimates for the reference group were equated to the underlying metric of the focal group. The EQUATE program was applied iteratively to determine the final linking coefficients using the procedure reported by Candell & Drasgow (1988). The final linking constants were separately and iteratively obtained for NCDIF, ESA, EUA, and LC. In order to use the iterative process at the item level, the DIF procedure under consideration must have a significance test. Because CDIF does not have a significance test (and items with significant CDIF are identified with the help of a significance test for the DTF index), the final linking constants obtained with the NCDIF procedure were used to transform the item and θ parameters in the DFIT framework.

Measurement of Differential Item and Test Functioning

Each of the datasets was analyzed for DIF with LC, Raju's (1990) z statistics for ESA and EUA, the χ^2 statistic for $\widehat{\text{DTF}}$ (Equation 24), and the cut-off value of .006 for $\widehat{\text{NCDIF}}$. For all these measures, items were examined for significant differential functioning at $\alpha = .01$. The relative effectiveness of DIF detection across methods was determined by examining the number of false positives (FPs; i.e., incorrectly identifying an item as

Downloaded from http://apm.sagepub.com at GEORGIA STATE UNIVERSITY on January 18, 2008
© 1995 SAGE Publications. All rights reserved. Not for commercial use or unauthorized distribution.

Table 1
Item Parameters for Generating Unidirectional DIF Conditions (Blanks Indicate That the Same Parameters Were Used for the Focal Group as for the Reference Group)

	Reference		Focal 1		Focal 2		Focal 3	
Item	*a*	*b*	*a*	*b*	*a*	*b*	*a*	*b*
1	.55	0.00						
2	.55	0.00						
3	.73	−1.04						
4	.73	−1.04						
5	.73	0.00	.73	1.00	.73	1.00	.73	1.00
6	.73	0.00						
7	.73	0.00						
8	.73	0.00						
9	.73	1.04						
10	.73	1.04	.73	2.04	.73	1.54	.73	1.54
11	1.00	−1.96						
12	1.00	−1.96						
13	1.00	−1.04						
14	1.00	−1.04						
15	1.00	−1.04			1.00	−.04	.50	−.54
16	1.00	−1.04						
17	1.00	0.00						
18	1.00	0.00						
19	1.00	0.00						
20	1.00	0.00			1.00	.50	.50	0.00
21	1.00	0.00						
22	1.00	0.00						
23	1.00	0.00						
24	1.00	0.00						
25	1.00	1.04					1.00	2.04
26	1.00	1.04						
27	1.00	1.04						
28	1.00	1.04						
29	1.00	1.96						
30	1.00	1.96					1.00	2.46
31	1.36	−1.04						
32	1.36	−1.04						
33	1.36	0.00						
34	1.36	0.00						
35	1.36	0.00					.86	.50
36	1.36	0.00						
37	1.36	1.04						
38	1.36	1.04						
39	1.80	0.00						
40	1.80	0.00					1.30	0.00

functioning differentially) and false negatives (FNs; i.e., failing to identify items with true differential functioning) produced under each method.

Results

Recovery of Item and θ Parameters

Correlations and RMSDs between generating parameters and parameter estimates were examined. In general, the results of the recovery analysis suggested acceptable recapturing of the underlying θ and item

Downloaded from http://apm.sagepub.com at GEORGIA STATE UNIVERSITY on January 18, 2008
© 1995 SAGE Publications. All rights reserved. Not for commercial use or unauthorized distribution.

Table 2
Item Parameters for Generating Bidirectional DIF Conditions (Blanks Indicate That the Same Parameters Were Used for the Focal Group as for the Reference Group)

	Reference		Focal 1		Focal 2		Focal 3	
Item	*a*	*b*	*a*	*b*	*a*	*b*	*a*	*b*
1	.55	0.00						
2	.55	0.00						
3	.73	−1.04						
4	.73	−1.04						
5	.73	0.00	.73	1.00	.73	1.00	1.23	0.00
6	.73	0.00	.73	−1.00	.73	−1.00		
6[a]	1.23	0.00					.73	0.00
7	.73	0.00						
8	.73	0.00						
9	.73	1.04						
10	.73	1.04						
11	1.00	−1.96						
12	1.00	−1.96						
13	1.00	−1.04						
14	1.00	−1.04						
15	1.00	−1.04			1.00	−.54	.50	−1.04
16	1.00	−1.04			1.00	−1.54		
16[a]	.50	−1.04					1.00	−1.04
17	1.00	0.00						
18	1.00	0.00						
19	1.00	0.00						
20	1.00	0.00						
21	1.00	0.00						
22	1.00	0.00						
23	1.00	0.00						
24	1.00	0.00						
25	1.00	1.04					1.00	2.04
26	1.00	1.04					1.00	.04
27	1.00	1.04						
28	1.00	1.04						
29	1.00	1.96					1.00	2.46
30	1.00	1.96					1.00	1.46
31	1.36	−1.04						
32	1.36	−1.04						
33	1.36	0.00						
34	1.36	0.00						
35	1.36	0.00						
36	1.36	0.00						
37	1.36	1.04						
38	1.36	1.04						
39	1.80	0.00						
40	1.80	0.00						

[a]For Focal 3, the item parameters for Items 6 and 16 were not the same as those under Focal 1 and 2.

parameters. These results were consistent with the findings of previous studies (e.g., Cohen & Kim, 1993; Kim & Cohen, 1992) using the same estimation and equating procedures. None of the datasets yielded recovery indexes extreme enough to warrant exclusion from further analyses. (Data from this phase of the study are not presented here, but they can be obtained from the authors.)

Downloaded from http://apm.sagepub.com at GEORGIA STATE UNIVERSITY on January 18, 2008
© 1995 SAGE Publications. All rights reserved. Not for commercial use or unauthorized distribution.

Detection of DIF

Table 3 shows the number of FPs and FNs and the number of equating iterations to a final solution for each DIF method for the unidirectional and bidirectional DIF conditions. In addition, the methods were compared for identification of uniformly biased and nonuniformly biased items (see Table 4). Only the frequency counts are reported in Tables 3 and 4 (information about the specific items identified as biased or unbiased can be obtained from the authors).

Table 3
Identification Errors (FPs and FNs) and Number of Iterations (NI)* for Each Method by Proportion of Test-Wide DIF and Sample Size for Unidirectional and Bidirectional DIF Conditions

% DIF, Errors, and NI	N = 500					N = 1,000				
	NCDIF	CDIF	ESA	EUA	LC	NCDIF	CDIF	ESA	EUA	LC
Unidirectional DIF										
0% DIF										
FP	0	0	0	0	0	0	0	0	0	0
FN	0	0	0	0	0	0	0	0	0	0
NI	0	–	0	0	0	0	–	0	0	0
5% DIF										
FP	2	1	0	0	1	0	0	0	0	0
FN	0	0	0	0	0	0	0	0	0	0
NI	2	–	2	2	1	1	–	1	1	1
10% DIF										
FP	0	0	0	0	0	0	0	0	0	0
FN	1	1	1	1	1	0	0	0	0	0
NI	2	–	1	1	1	1	–	1	1	1
20% DIF										
FP	0	0	0	0	0	0	0	1	0	0
FN	2	5	4	4	3	2	3	5	3	1
NI	1	–	2	3	1	2	–	1	1	1
Bidirectional DIF										
0% DIF										
FP	1	1	0	0	0	0	0	0	1	0
FN	0	0	0	0	0	0	0	0	0	0
NI	1	–	0	0	0	0	–	0	1	0
5% DIF										
FP	0	1	0	0	0	0	0	0	0	0
FN	0	0	0	0	0	0	0	0	0	0
NI	1	–	1	1	1	1	–	1	1	1
10% DIF										
FP	1	1	0	0	0	0	1	0	0	0
FN	0	0	0	0	0	0	0	0	0	0
NI	2	–	1	1	1	1	–	1	1	1
20% DIF										
FP	0	1	0	0	0	0	1	1	1	1
FN	2	0	6	2	2	1	0	4	0	0
NI	1	–	1	1	1	1	–	1	1	1

*Because the same final linking constants from NCDIF were also used in the CDIF analysis, NI is not reported for CDIF.

Downloaded from http://apm.sagepub.com at GEORGIA STATE UNIVERSITY on January 18, 2008
© 1995 SAGE Publications. All rights reserved. Not for commercial use or unauthorized distribution.

Effects of Sample Size

Unidirectional DIF. The results for the unidirectional DIF datasets indicated that a considerably smaller number of FPs than FNs were identified across sample sizes (see Table 3). The largest number of FPs was found for the $N = 500$, 5% DIF condition. These resulted from NCDIF (2 FPs), CDIF (1 FP), and LC (1 FP). The single FP identification for the $N = 1{,}000$ condition resulted from ESA and occurred under the 20% DIF condition.

The number of FNs under the unidirectional DIF condition tended to decrease as N increased and to increase as the proportion of test-wide DIF increased. The latter trend was most evident for $N = 500$; the largest total number of FN identifications was for CDIF (6 total; 1 at the 10% DIF condition and 5 at the 20% DIF condition). The fewest total numbers of FNs occurred with NCDIF (3) and LC (4).

The number of equating iterations required to reach a final solution tended to decrease across methods with an increase in N. For both $N = 500$ and $N = 1{,}000$, the number of iterations remained relatively consistent across the proportion of test-wide DIF conditions. The largest single number of iterations (3) across all conditions was required by EUA under the $N = 500$, 20% DIF condition. Furthermore, under $N = 500$ and across the test-wide DIF conditions, the fewest total number (3) of iterations was required by LC. For $N = 1{,}000$ and across the proportion of test-wide DIF conditions, the largest total number (4) of iterations was required by NCDIF.

Bidirectional DIF. As described above, simulation of DIF under this condition was done so as to produce bidirectional, but balanced, differential functioning at the test level. At the item level, item parameters were selected to create adjacent pairs of differentially functioning items with (mostly) off-setting bias. Therefore, such items would not be considered biased within the CDIF definition of bias. Hence, all items in the bidirectional condition are considered unbiased within the context of CDIF. Results across the bidirectional DIF condition should be considered within this framework.

In addition, items were constructed to reflect only uniform DIF for the 5% and 10% proportion of DIF conditions, and uniform and nonuniform DIF under the 20% proportion of DIF condition. The nonuniformly biased items for the bidirectional DIF condition were designed so that the generating bs were equal in both the reference and focal groups (i.e., $a_R \neq a_F$, $b_R = b_F$). The intention was to produce both offsetting areas inscribed by the IRFs for each biased pair element and (mostly) offsetting areas across a biased pair. The desired net effect, again, was to produce compensated differential functioning at the test level with differentially functioning items.

Under the bidirectional DIF condition, the number of FPs was markedly smaller than the number of FNs across sample sizes (see Table 3). Both $N = 500$ and $N = 1{,}000$ had a total of 6 FPs across all methods. For both $N = 500$ and $N = 1{,}000$, FPs were observed under the null DIF condition (no FPs were observed under the null DIF condition for unidirectional DIF). For $N = 500$ and null DIF, NCDIF and CDIF identified the same, single item (Item 26). For $N = 1{,}000$ and null DIF, EUA identified a single item (Item 24). The largest number of FPs occurred under the $N = 1{,}000$, 20% DIF condition. ESA, EUA, and LC identified the same, single item (Item 39), although CDIF identified a different item (Item 1). Note, however, that NCDIF produced no FPs for $N = 1{,}000$.

There were no FNs across sample size and proportion of DIF conditions, with the exception of the 20% DIF condition. A larger total number of FNs was observed under this condition for $N = 500$ (12 FNs) than for $N = 1{,}000$ (5 FNs). 10 of the total 15 FNs resulted from ESA, whereas NCDIF produced 3 FNs and EUA and LC produced 2 FNs each. CDIF produced no FNs. The findings for ESA were not unanticipated (discussed below).

The number of required equating iterations was generally constant across sample sizes. Only NCDIF required more than a single iteration to reach a solution ($N = 500$, 10% DIF condition).

Effects of Uniformity

Unidirectional DIF. Table 4 provides a summary of results obtained across DIF detection methods and

Downloaded from http://apm.sagepub.com at GEORGIA STATE UNIVERSITY on January 18, 2008
© 1995 SAGE Publications. All rights reserved. Not for commercial use or unauthorized distribution.

DIF conditions with respect to the identification of items with uniform and nonuniform bias. For the unidirectional DIF condition, all DIF methods, including CDIF, had the same items as biased. For the unidirectional DIF condition, the largest total number (12) of identifications of truly biased items across *N*s occurred equally for NCDIF and LC. For $N = 500$ and unidirectional DIF, NCDIF identified 6 of the 8 biased items, and LC identified 5 items. NCDIF identified 3 each of the uniformly and nonuniformly biased items. LC identified 2 uniformly biased items as well as the same 3 nonuniformly biased items identified by NCDIF. For $N = 500$, NCDIF and LC were the only methods to identify a nonuniformly biased item characterized by equal *b*s (Item 20). CDIF identified the fewest total number (3) of biased items, two of which were uniformly biased. CDIF, ESA, and EUA all failed to identify a single nonuniformly biased item characterized by equal *b*s.

The results for the $N = 1{,}000$, unidirectional DIF condition were somewhat different. Again, NCDIF and LC identified the largest total numbers (6 and 7, respectively) of biased items. However, these methods were joined by ESA in identifying the largest number of nonuniformly biased items (each identified the same three items—Items 15, 20, and 35); Item 20 was the only item of the three with equal *b*s. CDIF also identified Items 15 and 20. Note that CDIF identified one less total number of items than NCDIF (5 vs. 6), and identified the same uniformly biased items as NCDIF. ESA identified the fewest total number (4) of biased items; it identified

Table 4
Number of Items Identified with Significant Bias by DIF Condition, Detection Method, and Simulated Uniformity of Bias (The Numbers in Parentheses Refer to the Uniform, Nonuniform, and Total Number of Truly Biased Items)

DIF Condition, *N*, and Method	Total Identified*	Number of Truly Biased Items Identified		
		Total (8)	Uniformly Biased (4)	Nonuniformly Biased (4)
Unidirectional DIF				
$N = 500$				
NCDIF	6	6	3	3
CDIF	3	3	2	1
ESA	4	4	2	2
EUA	4	4	2	2
LC	5	5	2	3
$N = 1{,}000$				
NCDIF	6	6	3	3
CDIF	5	5	3	2
ESA	4	3	2	1
EUA	5	5	2	3
LC	7	7	4	3
Bidirectional DIF				
$N = 500$				
NCDIF	6	6	3	3
CDIF	1	0	0	0
ESA	2	2	2	0
EUA	6	6	2	4
LC	6	6	3	3
$N = 1{,}000$				
NCDIF	7	7	4	3
CDIF	1	0	0	0
ESA	5	4	4	0
EUA	9	8	4	4
LC	9	8	4	4

*Total number of items identified as having significant DIF regardless of whether they were truly biased.

Downloaded from http://apm.sagepub.com at GEORGIA STATE UNIVERSITY on January 18, 2008
© 1995 SAGE Publications. All rights reserved. Not for commercial use or unauthorized distribution.

only two uniformly biased items (Items 5 and 25) and a single nonuniformly biased item. In addition, ESA failed to identify the nonuniformly biased items characterized by equal *b*s. This appears to be a reasonable finding because ESA is sensitive only to items with differences in the *b*s (Raju, 1988). Finally, Item 40 was the only nonuniformly biased item to elude detection across all conditions; this item may have been missed because of its extreme true *a* value (see Table 1).

Bidirectional DIF. Under the bidirectional DIF condition, the largest total number (14) of DIF identifications of truly biased items across *N*s was made by LC and EUA. NCDIF identified two less items than these methods. For $N = 500$, LC, EUA, and NCDIF each identified 6 truly biased items. Of these, EUA identified 2 of the uniformly biased items and all of the nonuniformly biased items. Recall that under the CDIF condition, all nonuniformly biased items were characterized by equal *b*s. For $N = 500$, NCDIF and LC identified the same three uniformly biased and nonuniformly biased items. ESA failed to identify any items with nonuniform DIF and identified two of the four uniformly biased items. CDIF identified only one item as biased for each sample size condition. This was expected because there were no items simulated with significant CDIF under the bidirectional condition (recall that CDIF, by definition, has no biased items under the bidirectional DIF condition).

The identification of truly biased items improved for $N = 1{,}000$. EUA and LC identified all the biased items; NCDIF identified all but a single nonuniformly biased item (Item 6). ESA identified only half of the truly biased items but all of the uniformly biased items.

Discussion

In general, the number of detection errors was relatively low across simulated conditions and methods. For the unidirectional DIF condition, the number of FPs was markedly smaller than the number of FNs. The largest number of FPs was found for the $N = 500$ condition. The number of FNs, the primary indicator of threat to the validity of a measure, tended to decrease as *N* increased and to increase as the proportion of test-wide DIF increased. The most problematic condition across methods appeared to be the $N = 500$, 20% DIF condition. Here, NCDIF performed better than the other methods, and CDIF performed least well. For the $N = 1{,}000$, 20% DIF condition, the poorest performance was exhibited by ESA. This appeared to be related to the presence of nonuniform DIF characterized by equal *b*s.

Detection errors under the bidirectional condition were minimal for almost all of the methods across simulated conditions. The exception was ESA. ESA failed to identify items with nonuniform bias characterized by equal *b*s. This finding, across bidirectional conditions of DIF, also provides strong support for the theoretical expectations (see Raju, 1988) but contradicts the results of a previous study reported by Cohen & Kim (1993). They found no differences between LC, ESA, and EUA with respect to FN identification of uniform versus nonuniform DIF. In addition, they reported that the use of EUA appeared more likely to result in identification errors than either of the other two methods. The present results do not generally support this finding across the conditions examined.

In the bidirectional balanced DIF condition, CDIF did not identify any items with true DIF for $N = 1{,}000$, and only a single item for $N = 500$. These findings provide strong support for the intended meaning and purpose of the CDIF index.

The number of equating iterations required to reach a final solution was minimal across all simulated conditions. Generally, only a single iteration was required following the initial equating of item parameters. Note, however, that all four DIF methods (NCDIF, ESA, EUA, and LC) required at least one iteration beyond initial equating to arrive at a stable solution under all proportions of test-wide DIF beyond the 0% DIF condition. This finding provides additional support for the findings of other studies (see Candell & Drasgow, 1988; Cohen & Kim, 1993).

Although the results of this study provide strong support for the theoretical expectations for the func-

Downloaded from http://apm.sagepub.com at GEORGIA STATE UNIVERSITY on January 18, 2008
© 1995 SAGE Publications. All rights reserved. Not for commercial use or unauthorized distribution.

tioning of the new measures, a number of important issues were raised that have implications for further research. The first concerned the need for a significance test that takes into account the estimation errors associated with $\hat{\theta}$, $\hat{a}$, $\hat{b}$, and $\hat{c}$. Although the results from the proposed significance tests and/or empirically determined cutoff levels appear to be promising, there is still a need for further research on the distribution of D (assumed to be normal) and on the empirically determined cutoff level used with DTF and NCDIF. The critical value at $\alpha = .01$ was established by the percentage of FPs observed in several monte carlo studies in which both the focal and reference groups had the same θ distributions and identical item parameters. It is not presently known if this critical value was optimal for detection of differential functioning across the experimental conditions. More comprehensive research is required to either develop a different statistical significance test that is less sensitive to sample size or to establish precisely the critical values for various α levels. The influence of different decision rules on the rates of FPs and FNs across the new procedures also should be examined. It may be that each procedure requires a somewhat different rule for determining which items to remove and when (sequentially) to remove them.

The impact of the α level selection on identification errors was not addressed in this study. Two recent studies (Cohen & Kim, 1993; Kim & Cohen, 1992) indicated that the number of FNs tended to decrease across the examined IRT-based methods with an increase in α level from .01 to .05. A comparative study of identification errors at these two levels for the new measures would be helpful. Because the DIF conditions simulated in this study were selected to provide only a partial replication of those reported in Cohen & Kim (1993), they do not provide for an examination of the influence of unequal reference and focal group θ means (impact) on the detection of DIF. Cohen & Kim's study (1993), however, suggested relatively small differences in DIF detection between matched and nonmatched reference and focal groups across the conditions examined. Nonetheless, the new DFIT indexes need to be investigated within the context of various degrees of impact.

In addition, further study on variables that impact the power of IRT-based measures is required for these measures. The influence of sample size (smaller than those used in this study), relative amounts of bias in items (e.g., minimum required for detection), number and mix (unidirectional and bidirectional) of biased items, and different test lengths should be systematically investigated. Also, future investigations of the DFIT framework should include the MH technique, Shealy & Stout's (1993) SIBTEST, and the likelihood ratio tests (Thissen et al., 1988).

References

Angoff, W. H., & Ford, S. F. (1973). Item-race interaction on a test of scholastic aptitude. *Journal of Educational Measurement, 10,* 95–105.

Baker, F. B. (1990). Some observations on the metric of PC-BILOG results. *Applied Psychological Measurement, 14,* 139–150.

Baker, F. B., Al-Karni, A., & Al-Dosary, I. M. (1991). EQUATE: A computer program for the test characteristic method of IRT equating. *Applied Psychological Measurement, 15,* 78.

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

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

Dorans, N. J. (1986, April). *Two new approaches to assessing unexpected differential item preference: Standardization and the Mantel-Haenszel methods.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

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

Fleer, P. F., Kiley, K. A., & Raju, N. S. (1991). *RANGEN* [Computer program]. Unpublished computer program, Illinois Institute of Technology, Chicago.

Flowers, C. P., Oshima, T. C., & Raju, N. S. (1995, April). *A monte carlo assessment of DFIT with polytomously scored unidimensional tests.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Downloaded from http://apm.sagepub.com at GEORGIA STATE UNIVERSITY on January 18, 2008
© 1995 SAGE Publications. All rights reserved. Not for commercial use or unauthorized distribution.

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 129–145). Hillsdale NJ: Erlbaum.

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

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

Lim, R. G., & Drasgow, F. (1990). Evaluation of two methods for estimating item response theory parameters when assessing differential item functioning. *Journal of Applied Psychology, 75,* 164–174.

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

Mellenbergh, G. J. (1982). Contingency table models for assessing item bias. *Journal of Educational Measurement, 7,* 105–108.

Millsap, R. E., & Everson, H. T. (1993). Methodology review: Statistical approaches for assessing measurement bias. *Applied Psychological Measurement, 17,* 297–334.

Mislevy, R. J., & Bock, R. D. (1990). *BILOG: Item analysis and test scoring with binary logistic models.* Mooresville IN: Scientific Software.

Oshima, T. C., Raju, N. S., & Flowers, C. P. (1993, April). *Evaluation of a multidimensional IRT-based DIF/DTF index.* Paper presented at the annual meeting of the American Educational Research Association, Atlanta.

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

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

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

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

Rudner, L. M., Geston, P. R., & Knight, D. L. (1980). Biased item detection techniques. *Journal of Educational Measurement, 17,* 213–233.

Scheuneman, J. D. (1979). A method for assessing bias in test items. *Journal of Educational Measurement, 16,* 143–152.

Shealy, R. T., & Stout, W. F. (1993). A model-based standardization approach that separates true bias/DIF from ability group differences and detects test bias/DTF as well as item bias/DIF. *Psychometrika, 58,* 159–194.

Shepard, L. A., Camilli, G., & Averill, M. (1981). Comparison of procedures for detecting test-item bias with both internal and external ability criteria. *Journal of Educational Statistics, 6,* 317–375.

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 7,* 201–210.

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

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

Wainer, H. (1993). Model-based standardized measurement of an item's differential impact. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 123–135). Hillsdale NJ: Erlbaum.

Acknowledgments

Portions of this paper were previously presented at the 1992 and 1995 Annual Meetings of the American Educational Research Association. The authors express their appreciation to two anonymous reviewers for their helpful comments.

Author's Address

Send requests for reprints or further information to Nambury S. Raju, Institute of Psychology, Illinois Institute of Technology, Chicago IL 60616, U.S.A.

Downloaded from http://apm.sagepub.com at GEORGIA STATE UNIVERSITY on January 18, 2008
© 1995 SAGE Publications. All rights reserved. Not for commercial use or unauthorized distribution.