academicJournals

Vol. 11(13), pp. 1251-1261, 10 July, 2016
DOI: 10.5897/ERR2016.2803
Article Number: 9E279A959269
ISSN 1990-3839

http://www.academicjournals.org/ERR

Educational Research and Reviews

Full Length Research Paper

A comparison of four differential Item functioning procedures in the presence of multidimensionality

Özlem Yeşim Özbek Baştuğ

Çankırı Karatekin University, Turkey.

Received 3 April, 2016; Accepted 7 June, 2016

Differential item functioning (DIF), or item bias, is a relatively new concept. It has been one of the most controversial and the most studied subject in measurement theory. DIF occurs when people who have the same ability level but from different groups have a different probability of a correct response. According to Item Response Theory (IRT), DIF occurs when item characteristic curves (ICC) of two groups are not identical or do not have the same item parameters after rescaling. Also, DIF might occur when latent ability space is misspecified. When the groups have different multidimensional ability distributions and test items chosen to discriminate among these abilities, using unidimensional scoring, might flag items as DIF items. The purpose of this study was to compare four DIF procedures the Mantel Haenszel (MH), the Simultaneous Item Bias Test (SIBTEST), the IRT, the Logistic Regression (LR) when the underlying ability distribution is erroneously assumed to be homogenous. To illustrate the effect of assuming a homogenous ability distribution for the groups while they differ in terms of their underlying multidimensional ability levels on the DIF procedures, two different data sets were generated; one set in which DIF occurs, and one set in which no DIF occurs by using 2PL model. The UNIGEN program was used to generate the data. Each of the data sets contained 1000 examinees and 25 items. Item parameters where chosen to be capable of measuring a two dimensional ability distribution of the two groups. The MH, the SIBTEST, the AREA and the LR procedures were applied to the data both with DIF and without DIF. The study showed that all the four methods identified items as biased when the ability space was misspecified.

Key words: Item response theory, simultaneous ıtem bias test (SIBTEST), differential item functioning, differential item functioning (DIF), Mantel Haenszel (MH), logistic regression (LR).

INTRODUCTION

Differential item functioning (DIF), or item bias, is a relatively new concept. It has been one of the most controversial and the most studied subject in measurement theory. DIF occurs when people who have the same ability level but from different groups have a different probability of a correct response. According to Item Response Theory (IRT), DIF occurs when item characteristic curves (ICC) of two groups are not identical

E-mail:ozacik@yahoo.com.

or do not have the same item parameters after rescaling (Hambleton and Swaminathan, 1985). DIF is categorized as uniform and non-uniform, according to interactions between group and ability levels. Non-uniform DIF occurs when there is an interaction between group membership and ability levels, whereas uniform DIF occurs when there is no interaction between the two. In uniform DIF, the ICC do not cross and one group is always superior to the other group, but in non-uniform DIF the ICC cross and both reference and focal groups might be superior to each other in different parts of the ability scale (Hambleton and Swaminathan, 1985; 1989). Also, DIF might occur when latent ability space is misspecified. When the groups have different multidimensional ability distributions and test items chosen to discriminate among these abilities, using unidimensional scoring, might flag items as DIF items (Ackerman, 1992). Stout et al. (2001) described the dimensionality as "…minimum number of dimensions of θ required to produce a locally independent and monotone latent variable model...When the dimensionality of a test is one, the latent variable model is called *unidimensional*, and when the dimensionality of a test is greater than one, the latent variable model is called *multidimensional*." (p. 359).

Practitioners sometimes incorrectly presume that test takers in their groups have the same underlying *unidimensional* distribution when it is a *multidimensional* distribution. When the numbers of examinees and items are large, this assumption cannot be satisfied. Researchers need to check unidimesionality in the data before adopting any of the DIF methods and they should use multidimensional models for parameter estimation in the violation of unidimensionaltiy to understand the behavior of an item (Ackerman, 1992). Otherwise, several valid items may have been flagged as DIF items and can be wrongly eliminated from the test.

This paper compares four DIF procedures- the Mantel Haenszel (MH), the Simultaneous Item Bias Test (SIBTEST), the Item Response Theory (IRT), the Logistic Regression (LR) - when the latent ability space is misspecified and the multidimensional ability space is scored by using unidimensional scaling.

The MH, the SIBTEST and the Logistic Regression are non-parametric methods and the IRT is a parametric method. Herein, the four methods will be explained in some depth and their advantages and disadvantages will be identified.

Mantel Haenszel Method (MH)

When using the MH procedure, examinees are matched according to their observed correct score and then contingency tables are prepared for each test item. Table 1 shows a contingency table that contains the number of examinees in each group who correctly or incorrectly respond to an item (Clauser and Mazor, 1998). The null

Table 1. Contingency table.

Gorup	Score on studied item		
	1	0	Total
Reference	A_j	B_j	N_{rj}
Focal	C_j	D_j	$N_{f\,j}$
Total	$N_{1.j}$	$N_{0.j}$	$N_{.j}$

and alternate hypothesis given as,

$$H_0 = \left[\pi_{Rj}/\left(1-\pi_{Rj}\right)\right] = \left[\pi_{Fj}/\left(1-\pi_{Fj}\right)\right] j = 1,2,....,k, \quad (1)$$

$$H_A = \left[\pi_{Rj}/\left(1-\pi_{Rj}\right)\right] = \alpha\left[\pi_{Fj}/\left(1-\pi_{Fj}\right)\right] j = 1,2,....,k, \alpha \neq 1, \quad (2)$$

are tested against each other, where α does not equal 1 and π is the probability of a correct response. The odds (α) and weighted odds (α_{MH}) given as:

$$\alpha = \frac{\frac{C}{D}}{\frac{A}{B}} = \frac{CB}{AD}, \quad (3)$$

$$\alpha_{MH} = \frac{\sum A_j B_j / N_{..j}}{\sum B_j C_j / N_{..j}}. \quad (4)$$

are computed. The weighted odds ratio takes on values between 0 and infinity. A general guide for interpretation of a α_{MH} result might be: α_{MH}= 1.0 indicating no DIF, α_{MH} >1 indicating the item favors the reference group and α_{MH} <1 indicating the item favors the focal group (Clauser and Mazor, 1998; Hambleton and Swaminatan, 1985; Millsap and Everson, 1993; Narayan and Swaminathan, 1996). Because the interpretation of these values is difficult, a logistic transformation is used.

$$\Delta_{MH} = -\frac{4}{1.7}\ln\left(\alpha_{MH}\right) = -2.35\ln\left(\alpha_{MH}\right), \quad (5)$$

where Δ_{MH} takes values between negative infinity and positive infinity and can be interpreted as: Δ_{MH} >0 which indicates item favors the reference group and Δ_{MH} <0 indicates item favors the focal group (Millsap and Howard, 1993). The Mantel Haenszel procedure also provides a significance test. This test is given as

$$\chi^2_{MH} = \frac{[\,|\sum A_j - \sum E\left(A_j\right)| - .5]^2}{\sum Var\left(A_j\right)}, \quad (6)$$

where Aj corresponds to the number of examinees in the reference group responding correctly to J^{th} item,

$$E\left(A_j\right)=\frac{N_{Rj}N_{1.j}}{N_{..j}}, \quad (7)$$

$$Var\left(A_j\right)=\frac{N_{Rj}N_{Fi}N_{1.j}N_{0.j}}{\left(N_{..j}\right)^2\left(N_{..j}-1\right)}, \quad (8)$$

and the ratio has a chi-square distribution with one degree of freedom. Chi-square statistics are affected by sample size; therefore, testing for both statistical significance and effect size might be useful to avoid detecting items with small practical significance erroneously, such as DIF items (Clauser and Mazor, 1998; Millsap and Everson, 1993). Although the MH procedure is one of the most utalized DIF methods due to its simplicity and practicality, it also has some major drawbacks. The MH procedures are successful in detecting uniform DIF but it might yield misleading results in nonuniform DIF or when using more complex models (DeMars, 2009; Güler ve Penfield, 2009; Millsap and Everson, 1993; Narayan and Swaminathan, 1996). Nowadays, a version of MH procedures for polytomous items and a software program called the Mantel Haenszel is available.

Simultaneous item bias test (SIBTEST)

The SIBTEST, generated by Shealy and Stout (1993), provides a DIF procedure that can do a set of DIF analyses at the same time. In SIBTEST, items suspected to be functioning differentially are called “suspected subsets” and remaining items are called “valid item subsets”. The SIBTEST matches reference and focal group according to their estimated latent ability based upon the observed score on what the practitioner considers to be the valid items. First, examinee scores are calculated on a valid subset, and then the proportion of correct responses is calculated for suspected items. The SIBTEST works iteratively until all suspected items are removed from the valid subset. The final subsets of items that are DIF free are used as the matching criterion (Clauser and Mazor, 1998). The SIBTEST can detect both uniform and non-uniform DIF. The hypotheses for testing uniform and nonuniform DIF are

$$H_0:\beta_U=0, \quad (9)$$

$$H_A:\left|\beta_U\right|>0, \quad (10)$$

$$H_0:\beta_C=0, \quad (11)$$

$$H_A:\left|\beta_C\right|>0, \quad (12)$$

where β denotes the amount of DIF and a β value of zero indicates no DIF. Sometimes β values larger than zero, due to systematic differences between the groups require a regression correction. The regression correction is used to compute a true score for each examinee. This latent score is then used to match examinees. Uniform and non-uniform DIF hypotheses are tested simultaneously in order to control type I error. The total score for a valid subset and for a suspected subset, respectively, is given as:

$$X=\sum_{i-1}^{n}U_i, \quad (13)$$

$$Y=\sum_{i=N+1}^{N}U_i, \quad (14)$$

where U_i is one and zero for correct and incorrect answers, respectively. An estimation of β_u is

$$\hat{\beta}_U=\sum_{k=0}^{n}\hat{P}_k\left(\overline{Y}_{Rk}-\overline{Y}_{Fk}\right), \quad (15)$$

where $\hat{P}_k$ denotes the proportion of focal group examinees who get a score of k on the valid subset

$$\hat{P}_k=\frac{\left(G_{Rk}+G_{Fk}\right)}{\sum_{j=0}^{n}\left(G_{Rk}+G_{Fk}\right)}. \quad (16)$$

The test statistic for testing uniform DIF (B_U) and standard error of beta $\hat{\sigma}\left(\hat{\beta}_U\right)$ are

$$B_U=\hat{\beta}_U/\hat{\sigma}\left(\hat{\beta}_U\right), \quad (17)$$

$$\hat{\sigma}\left(\hat{\beta}_U\right)=\left\{\sum_{k=0}^{n}\hat{P}^2{}_k\left[\frac{1}{G_{Rk}}\sigma^2\left(Y|k,R\right)+\frac{1}{G_{Fk}}\sigma^2\left(Y|k,F\right)\right]\right\}^{\frac{1}{2}}, \quad (18)$$

where $k=0,...,n$. The estimator of β_c for non-uniform DIF and the test statistic are

$$\hat{\beta}_c=\sum_{k=0}^{k_0}\hat{P}k\left(\overline{Y}_{fk}-\overline{Y}_{Rk}\right)+\left(\overline{Y}_{Rk}-\overline{Y}_{Fk}\right), \quad (19)$$

$$B_C=\hat{\beta}_C/\hat{\sigma}\left(\hat{\beta}_c\right) \quad (20)$$

(Narrayan and Swaminathan, 1996). When there is no DIF, βu and βc are distributed as normal distributions with a mean of 0 mean and a variance of 1. The alternate hypothesis is accepted if the test statistic exceeds 100 (1-α) percentile (Narayan and Swaminathan, 1996). Positive β values indicate that an item is favoring the focal group. Also, the SIBTEST can be used for bundle DIF analysis.

Item response theory (IRT)

Although there is no single IRT method that can be used to detect DIF, all IRT procedures compare item characteristic curves (ICC) that are assumed to be invariant across groups after they have been rescaled. A general framework includes: (a) matching examines, (b) selecting an appropriate IRT model, (c) estimating item and examinee parameters for each group, (d) transforming estimates to a common scale, and (e) finding the DIF area by subtracting the reference and focal group's ICC from each other. Because item parameters are estimated separately for the focal and reference groups, they share different scales and cannot be compared directly. A common scale is needed. Scaling is possible on both item and ability parameters. Scaling is performed on the item difficulty parameter by constraining the mean and standard deviation to 0 and 1, respectively. This methodology is convenient for three unidimensional logistic models and the normal ogive model. This process puts estimates on a common scale; however, they are constrained separately. Scaling on ability parameters by constraining the mean and variance to 0 and 1, respectively, does not provide a common scale for comparison and an additional transformation is required. Discrimination and difficulty estimates can be transformed as follows:

$$\hat{b}_{\lg} = k\hat{b}_{\lg^*} = m, \tag{21}$$

$$\hat{a}_{\lg} = \frac{\hat{a}_{\lg^*}}{k}, \tag{22}$$

where $\hat{k} = \dfrac{\left(\hat{\sigma}_2^2 - \hat{\sigma}_1^2\right) + \sqrt{\left(\hat{\sigma}_2^2 - \hat{\sigma}_1^2\right)^2 + 4\hat{\rho}_{12}}\hat{\sigma}_1\hat{\sigma}_2}{2\hat{\rho}_{12}\hat{\sigma}_1\hat{\sigma}_2}$, (23)

and $\hat{m} = \bar{b}_2 - \hat{k}\bar{b}_1$. (24)

The pseudo guessing parameter is not transformed because it is invariant by scale (Crocker and Algina, 1986).

Among the various IRT procedures, the area method is perhaps the easiest and also provides a test of significance. Raju's (1990) procedure will be examined in here as an example of the AREA method. According to Raju, the area between two ICCs can be found by subtracting the two ICCs from each other. Also, item mean and variance can be calculated for each item and later they can be used in hypothesis testing. Raju formulated mean and variance for both signed and unsigned areas for the one, two and three parameter logistic models. In this paper, only formulas for the one and the two parameter models are demonstrated (further detail can be obtained from Raju, 1990). The test statistic for signed area (SA) and unsigned area (US) when using one parameter models, respectively, are

$$SA_{10} = \hat{b}_2 - \hat{b}_1, \tag{25}$$

$$Ua_{11} = \left|\hat{b}_2 - \hat{b}_1\right|. \tag{26}$$

Also, the mean and variance for one parameter models are

$$\mu\left(SA_{10}\right) = E\left(\hat{b}_2\right) - E\left(\hat{b}_1\right) = b_2 - b_1, \tag{27}$$

$$\sigma^2\left(SA_{10}\right) = Var\left(\hat{b}_2\right) - Var(\hat{b}_1), \tag{28}$$

where $Var(\hat{b}_i) = \left[\sum_{J=1}^{Nli} P_i(\theta_j)\vartheta_i(\theta_j)\right]^{-1}$. (29)

The test statistic, mean and variance for the signed area for two-parameter models are

$$SA_{20} = \hat{b}_2 - \hat{b}_1. \tag{30}$$

$$\mu\left(SA_{20}\right) = E\left(\hat{b}_2\right) - E(\hat{b}_1) = b_2 - b_1, \tag{31}$$

$$\sigma^2(SA_{20}) = Var(\hat{b}_2 - \hat{b}_1) = Var(\hat{b}_2) - Var(\hat{b}_1), \tag{32}$$

where $Var(\hat{b}_i) = \dfrac{I_{ai}}{I_{ai}I_{bi} - I_{aibi}}$, (33)

and item information (I) is

$$I_{ai} = D^2 \sum_{j=1}^{Ni}\left(\theta - b_i\right)^2 P_i\left(\theta_j\right)Q_i\left(\theta_j\right), \tag{34}$$

$$I_{bi} = D^2a^2 \sum_{j=1}^{Ni} P_i(\theta_j)Q_i(\theta_j), \tag{35}$$

$$I_{aibi} = D^2a_i\left(\theta_j - b_i\right)P_i(\theta_j)Q_i(\theta_j). \tag{36}$$

Assuming the difference between the reference and focal group ICC is normally distributed, a significance test for signed area is given by Raju in Equation 37. If the obtained test statistics is between -z and +z, the null hypothesis is accepted and the item is considered DIF free. Because large samples tend to give more significant results, using small alpha values can protect erroneously detecting items as DIF.

$$Z = \frac{SA - 0}{\sigma(SA)}. \quad (37)$$

When the normality assumption is not tenable for unsigned areas (US), a different formula for US is given in Equation 38.

$$Z = \frac{H - 0}{\sigma(H)}, \quad (38)$$

A more simplistic approach without a significance test given by Linn et al. (1981) is as follows

$$A_{Li} = \sum_{\theta=-3}^{=3} \left| P_{i1}(\theta_k) - P_{i2}(\theta_k) \right| \Delta\theta, \quad (39)$$

where $\Delta\theta$ is at intervals of .005. Because the error of estimating ICCs differs at each ab ility level, weighting the ICC by its standard error can solve this problem. This is performed as follows:

$$A_{Li} = \sum_{\theta=-3}^{=3} \left\{ \left[\sum_{\theta=-3}^{+3} P_{i1}(\theta_k) - P_{i2}(\theta_k) \right]^2 \Delta\theta \right\}^{\frac{1}{2}}. \quad (40)$$

Interpretation of the value A is a little bit vague. Large A values indicate large bias, while small A values indicate small bias. Although IRT provides a general framework for DIF analyses, it has some major drawbacks. All IRT methods require a large sample size and this increases the number of parameters that have to be estimated. IRT procedures, unlike other methods, also require a considerable knowledge of IRT theory. Compared to other methods, IRT is less practical and much more complex (Crocker and Algina, 1986; Hambleton and Swaminathan, 1985).

Logistic regression (LR)

A logistic regression model detecting DIF items between the focal and the reference groups was introduced by Swaminathan and Rogers (1991). Although the logistic regression model is sensitive to both uniform DIF and non-uniform DIF, it has mainly been developed for detecting non-uniform DIF. The standard LR model for predicting the probability of a dichotomous dependent variable is (Bock, 1975):

$$P(u = 1|\theta) = \frac{e^{(\beta_0 + \beta_1\theta)}}{1 + e^{(\beta_0 + \beta_1\theta)}}, \quad (41)$$

Swaminathan and Rogers (1991) specified a LR model for DIF by creating separate equations for focal and reference group. This equation is as follows:

$$P(u_{ij} = 1|\theta_{ij}) = \frac{e^{(\beta_{0j} + \beta_{1j}\theta_{pj})}}{1 + e^{(\beta_{0j} + \beta_{1j}\theta_{pj})}}, \quad (42)$$

where $P(u_{ij}$=1) is the response of person p in group j, β_{0j} is the intercept parameter for group j, β_{1j} is the slope parameter for group j, and θ_{pj} is the ability of person p in group j. According to this model, an item is unbiased if intercept and slope terms are the same across the groups; that is, their logistic regression curves are exactly the same ($\beta_{01} = \beta_{02}$, $\beta_{11} = \beta_{12}$). On the other hand, an item is biased if logistic regression curves for the two groups are not exactly the same and differ across the groups. Uniform DIF occurs when logistic regression curves are parallel but not coincident. That is, when $\beta_{11} = \beta_{12}$ but $\beta_{01} \neq \beta_{02}$. Non-uniform DIF occurs when logistic regression curves cross each other. Because the model in Equation 42 does not capture the non-uniform DIF case, Swaminathan and Rogers (1991) reparameterized the LR model to capture uniform DIF and non-uniform DIF as

$$P(u_{ij} = 1|\theta_{ij}) = \frac{e^{z_{pj}}}{1 + e^{z_{pj}}}, \quad (43)$$

where $Zpj = T_0 + T_1\theta_{pj} + T_2 g_j + T_3(\theta_{pj} g_j)$, (44)

$P(u_{ij} = 1)$ is the probability of a correct response for person p in group j, T_0 is the intercept, T_1 is the coefficient of ability, $T_2 (= \beta_{01} - \beta_{02})$ is the group difference, and $T_3 (= \beta_{11} - \beta_{12})$ is the interaction between groups and ability while g is the group membership variable. In this model, an item is identified as exhibiting uniform DIF item when $T_3 = 0$ but $T_2 \neq 0$, and an item is identified as nonuniform DIF item if $T_3 \neq 0$ (whether or not $T_2 = 0$). The parameters of the model can be estimated by using the maximum likelihood method. The likelihood function for any item are given as:

$$L(u_{pj}|\theta) = \prod_{p=1}^{N} \prod_{j=1}^{n} (P(u_{pj}))^{u_{pj}} \left[1 + P(u_{pj}) \right]^{1-u_{pj}}, \quad (45)$$

where N is sample size, n is the test length, and u_{pj} =1, and $P(u_{pj})$ is the probability of a correct response for person p in group j as in Equation 43. Generally, the statistical significance of a coefficient is determined by using either likelihood ratio test or Wald statistic (Swaminathan and Roger, 1991). The Wald test is:

$$Z^2 = \frac{\hat{\beta}}{ASE}, \quad (46)$$

where Z^2 is the Wald statistic, $\hat{\beta}$ is the parameter estimate and ASE is the standard error of the estimate. Z^2 has a chi-square distribution with df =1 (Agresti, 1990). The likelihood ratio test (G) compares the likelihood ratio of a full and reduced model. The full and reduced models for uniform DIF case are as follows:

$$Z_{Full} = T_0 + T_1\theta + T_2 g, \quad = \quad (47)$$

$$Z_{\mathrm{Re}duced} = T_0 + T_1\theta. \quad (48)$$

Also, the full and reduced models for a non-uniform DIF case might be given as:

$$Z_{Full} = T_0 + T_1\theta + T_2 g + T_3\theta g, \quad (49)$$

$$Z_{\mathrm{Re}duced} = T_0 + T_1\theta + T_2 g. \quad (50)$$

Log likelihood ratios of these models can be calculated by using Equation 45. Then, the likelihood ratio test statistic is

$$G = -2(L_{\mathrm{Re}duced} - L_{full}), \quad (51)$$

which has a chi-square distribution with two degrees of freedom (Whitmore and Schumacker, 1999).

The logistic regression procedure can be used with multiple examinee groups and with polytomous item scores (Agresti, 1990). Another advantage of using logistic regression is that estimates of the regression coefficients can be plotted. This plot can then be used to detect where along the scale the DIF is becoming problematic (Miller et al., 1993). The LR procedure might give clear perspective on the possible causes of DIF by inclusion of a curvilinear term and other relevant examinee characteristics such as text anxiety. LR procedures use total score as a proxy for latent trait and this feature might cause some problems when items have a multiparameter IRT model. The MH and the SIBTEST also share the same problem. IRT procedures have calibration methods but if the underlying trait is not unidimensional, calibration will not put the groups on the same scale. In these conditions, practitioners should use caution when interpreting the results. If items are capable of measuring more than one ability, equal correct response number may not have the same meaning in the reference and the focal groups (Ackerman, 1992).

METHODS

To illustrate the effect of assuming a homogenous ability distribution for the groups while they differ in terms of their underlying multidimensional ability levels on the DIF procedures, two different data sets were generated, one set in which DIF occurs, and one set in which no DIF occurs by using the following 2PL model:

$$P(x=1|\theta, a_i, b_i) = \frac{e^{1.7ai(\theta - b_i)}}{1 + e^{1.7ai(\theta - b_i)}}. \quad (52)$$

The UNIGEN program was used to generate the data. Each of the data sets contained 1000 examinees and 25 items. Item parameters where chosen to be capable of measuring a two dimensional ability distribution of the two groups. Item parameters which are used to generate data can be found in Table 2.

Among the 25 items, the first ten were loaded on the first ability and the last ten items were loaded on the second ability. Items from 11-15 were loaded on both abilities and they indicated valid direction. The data for the no-DIF case were generated by using the same means and variances for both groups $(\mu_1 = 1, \mu_1 = 1, \sigma_1^2 = 1, \sigma_2^2 = 1)$. For the DIF data set, means were different but variances were kept the same $(\mu_1 = 1, \mu_2 = 0, \sigma_1^2 = 1, \sigma_2^2 = 1)$ so that the reference group was successful on the first ability level and the focal group was successful on the second ability level. Generation of data gave control over expectation of DIF items and no DIF items. Because item 11, 12, 13, 14, 15 were loaded on both ability levels they were expected to be detected as DIF items. Other items expected to favor either focal or reference group depending on its conditioning. In the no-DIF case, because both groups had the same underlying ability level and had the same item parameters, none of the items were expected to be detected as DIF item. Thus, all items were in valid direction.

The MH, the SIBTEST, the AREA and the LR procedures were applied to the data both with DIF and without DIF. The MH statistic was obtained by using the computer programs MANTEL and SIBTEST. The SIB statistic was obtained from the SIBTEST program. For the IRT statistic, estimates of item parameters were obtained separately for each group using the BILOG program. Because estimates were obtained independently, they did not share a common scale and the result had to be converted on a common scale in order to achieve possible comparison between the focal group and the reference group ICCs. The focal group's item estimates were rescaled onto the reference group's item estimates by using the RESCAL program. Then the areas between two groups' ICCs were obtained from the AREA program. A separate LR analysis was performed for each item by using the SAS program. The model included a dependent variable, dichotomous item response, and independent variables, total score, group and the interaction between group and total score.

RESULT

Four results from the DIF procedures, the MH, the

Table 2. Item Parameters Used to Generate 2PL Items for DIF Case and no DIF Case.

Item no	a1	a2	d	Item no	a1	a2	d
1	1.500	0.000	0	14	1.143	0.971	0
2	1.498	0.081	0	15	1.089	1.032	0
3	1.491	0.162	0	16	1.031	1.089	0
4	1.480	0.243	0	17	0.971	1.143	0
5	1.465	0.323	0	18	0.907	1.194	0
6	1.445	0.402	0	19	0.842	1.242	0
7	1.421	0.479	0	20	0.773	1.285	0
8	1.393	0.555	0	21	0.703	1.325	0
9	1.360	0.630	0	22	0.630	1.361	0
10	1.325	0.703	0	23	0.555	1.393	0
11	1.285	0.773	0	24	0.479	1.421	0
12	1.241	0.842	0	25	0.400	1.445	0
13	1.194	0.908	0				

SIBTEST, the item response (IRT), and the LR, were included. The number of items which flagged as DIF items in the DIF case for the MH, the SIBTEST, the AREA and the LR methods were, respectively, 22, 22, 20 and 19. The result for the SIBTEST and the MH were exactly the same, and they detected item 1-11 in favor of the focal group and item 15-25 in favor of the reference group (Tables 3 and 4 show the MH and the SIBTEST results for no DIF case and DIF case, respectively). The result for the AREA and LR were also similar (Tables 5 and 6 show the AREA and the LR results, respectively). The AREA method detected item 1-10 in favor of the reference group and item 16 - 25 in favor of the focal group. The LR method detected item 1- 9 in favor of the reference group and item 16 - 25 in favor of the focal group. In the no-DIF case, both the MH and the SIBTEST detected item 15 in favor of the focal group while the AREA method detected item 5 in favor of the reference group. The LR procedure did not flag any item.

DISCUSSION

Validation can be thought of as hypothesizing a certain construct as a potential source of plausible explanations of scores on a particular test. However, recognizing or foreseeing other constructs as potential sources of explanations of scores from the test and investigating the tenability of these alternative hypotheses are invaluable part of the validation process. Recently, DIF analysis shows up as a promising method for the validity investigation; to study construct relevant as well as irrelevant sources (Roussos and Stout, 1996; Walker and Beretvas, 2001). Although DIF studies have been undertaken since the early 1960s, still the underlying causes of DIF are not known (Messick, 1989; Walker and Beretvas, 2001). Given the apparent failure of cumulated DIF studies, researchers have emphasized the need to become familiar with the underlying latent ability distribution of the test before performing any DIF analysis (Messick, 1989).

Ackerman (1992) and Sheally and Stout (1991) showed that if the groups are not homogenous in terms of their underlying ability distributions, so that they do not have the same multidimensional ability levels, and the items are capable of measuring these dimensions, using unidimensional scoring instead of multidimensional scoring, they may cause items to be flagged as biased items. Therefore, the purpose of this study was to demonstrate the effect of assuming uniform ability distribution for the groups while they differ in terms of their underlying multidimensional ability levels on the DIF methods. The result supported their view and all the four DIF procedures (MH, SIBTEST, AREA, and LR) flagged items as biased in the misspecification of latent space. Table 7 shows the percentage of items which were flagged as biased across the four procedures. Although all of the four methods detected misspecification of latent space as DIF, their results were not exactly the same. The MH and the SIBTEST flagged more items than the AREA and the LR did. It was interesting to see that even the underlying multidimensional distributions are the same for the groups. In the no-DIF case, some items have been detected as DIF items. With the exception of LR procedure, all methods flagged one item as biased. This study showed the LR procedure was the best among these methods in terms of false positives. The difference among the results might be attributed to the sensitivity of each method. Previous research showed that the MH method was the least sensitive for non-uniform DIF, while the SIBTEST, the AREA and the LR methods had the same sensitivity for nonuniform and uniform DIF (Clauser and Mazor, 1998; Erdem, 2014; Güler and Penfield, 2009; Gommez-Benito and Navas-Ara, 2000; Narayan and Swaminathan, 1996; Rogers and Swaminathan, 1992). Previous studies also showed that the IRT

Table 3. Summary of the MH and SIBTEST analysis results in no DIF case.

Item no	SIB test			Mantel Haenszel		
	Beta-uni	z-statistic	p-value	Chı sqr.	p- value	(D-DIF)
1	-0.02	-0.983	0.326	0.48	0.491	0.25
2	-0.031	-1.575	0.115	2.17	0.141	0.51
3	-0.015	-0.78	0.435	0.23	0.635	0.19
4	-0.005	-0.23	0.818	0.11	0.74	0.13
5	-0.02	-1.018	0.309	2.41	0.121	0.56
6	0.013	0.716	0.474	1.01	0.316	-0.39
7	0.002	0.107	0.914	0	0.95	-0.05
8	-0.003	-0.177	0.86	0.05	0.831	0.1
9	0.016	0.811	0.418	1.01	0.316	-0.37
10	0.013	0.68	0.496	0.19	0.663	-0.17
11	0.011	0.555	0.579	0.1	0.755	-0.14
12	0.01	0.558	0.577	0.47	0.495	-0.28
13	-0.003	-0.188	0.851	0.01	0.92	0.07
14	-0.023	-1.178	0.239	0.82	0.364	0.33
15	0.054	2.757	0.006	8.68	0.003	-1.06*
16	0.006	0.287	0.774	0.05	0.826	-0.1
17	0.008	0.399	0.69	0.01	0.943	-0.05
18	0.023	1.176	0.24	1.54	0.215	-0.44
19	-0.017	-0.837	0.402	0.28	0.594	0.2
20	-0.026	-1.232	0.218	1.4	0.237	0.4
21	0.017	0.786	0.432	0.75	0.387	-0.3
22	0	-0.003	0.997	0.08	0.78	0.11
23	0.035	1.582	0.114	1.36	0.244	-0.37
24	-0.039	-1.747	0.081	1.89	0.169	0.43
25	-0.016	-0.721	0.471	0.19	0.666	0.15

p<0.05.

Table 4. Summary of the MH and SIBTEST analysis results in DIF case.

Item no	SIB Test			Mantel Haenszel		
	Beta-uni	z-statistic	p-value	Chı sqr.	p-value	(D-DIF)
1	0.327	15.543	0	207.11	0	-4.26*
2	0.29	13.644	0	164.14	0	-3.85*
3	0.274	13.423	0	157.55	0	-4.06*
4	0.239	11.869	0	131.29	0	-3.71*
5	0.214	10.944	0	118.95	0	-3.57*
6	0.186	9.277	0	85.64	0	-3.05*
7	0.173	8.864	0	71.18	0	-2.81*
8	0.165	8.543	0	65.72	0	-2.69*
9	0.109	5.842	0	33.82	0	-1.99*
10	0.081	4.206	0	16.26	0	-1.38*
11	0.04	2.128	0.033	5.34	0.021	-0.82*
12	0.015	0.781	0.435	0.54	0.461	-0.27
13	-0.013	-0.702	0.483	0.74	0.388	0.32
14	-0.014	-0.761	0.447	0.48	0.489	0.26
15	-0.044	-2.345	0.019	4.4	0.036	0.74*
16	-0.093	-5.054	0	25.04	0	1.81*
17	-0.124	-6.549	0	42.07	0	2.24*

Table 4. Cont'd.

18	-0.146	-7.63	0	59.02	0	2.6*
19	-0.157	-8.217	0	65.25	0	2.73*
20	-0.184	-9.637	0	88.28	0	3.2*
21	-0.224	-11.534	0	116.92	0	3.66*
22	-0.268	-13.251	0	154.38	0	3.87*
23	-0.259	-13.22	0	150.5	0	4*
24	-0.27	-13.23	0	161.27	0	4.04*
25	-0.287	-14.024	0	173.9	0	4.06*

p<0.05.

Table 5. Summary of the IRT analysis results.

Item	DIF case		No DIF case		Item	DIF case		No DIF case	
	Area	p	Area	p		Area	p	Area	p
1	-1.1196	0.5664 *	0.0333	0.021	14	0.0715	0.0596	0.061	0.0379
2	-0.9963	0.5182 *	0.117	0.0596	15	0.1322	0.0836	-0.1412	0.0954
3	-0.8865	0.4864 *	0.0059	0.0423	16	0.3171	0.2151 *	-0.0108	0.0129
4	-0.783	0.4363 *	0.0316	0.0231	17	0.3886	0.2395 *	-0.0197	0.0221
5	-0.7206	0.4270 *	-0.018	0.1115 *	18	0.5254	0.3251 *	-0.0732	0.0569
6	-0.6049	0.3436 *	-0.0265	0.0387	19	0.5222	0.3189 *	0.059	0.0381
7	-0.5148	0.3025 *	0.0005	0.0104	20	0.6116	0.3682 *	0.1021	0.0924
8	-0.5227	0.3046 *	0.0005	0.0104	21	0.6967	0.4137 *	-0.0385	0.0371
9	-0.3477	0.2199 *	-0.0561	0.0347	22	0.8633	0.4675 *	0.02	0.021
10	-0.2227	0.1359 *	-0.0454	0.0357	23	0.8257	0.4727 *	-0.0721	0.0432
11	-0.1066	0.0684	0.0058	0.0223	24	0.9027	0.4905 *	0.1067	0.0795
12	-0.0497	0.031	-0.0134	0.0383	25	0.9961	0.5426 *	0.0449	0.0583
13	0.0717	0.0455	-0.016	0.0457					

Area> 0.1 taken as significant.

Table 6. Summary of the logistic regression analysis results.

Item	DIF case			No DIF case		
	Wald Test	p-value	Odd ratio	Wald test	p-value	Odd ratio
1	46.1458	0.0001	8.249 *	1.069	0.3012	0.756
2	50.9292	0.0001	9.917 *	0.7085	0.3999	0.797
3	37.8101	0.0001	8.924 *	0.2131	0.6443	1.143
4	29.5769	0.0001	6.585 *	0.6568	0.4177	1.266
5	45.7883	0.0001	14.344 *	0.06	0.8065	1.076
6	21.4559	0.0001	5.187 *	0.6701	0.413	0.767
7	10.8767	0.001	3.391 *	1.3738	0.2412	1.47
8	15.3452	0.0001	4.028 *	0.5863	0.4438	0.775
9	13.3733	0.0003	4.184 *	0.3725	0.5417	1.233
10	3.0295	0.0818	1.913	1.2103	0.2713	1.47
11	0.3849	0.535	1.268	0.3183	0.5727	0.799
12	0.7114	0.399	1.346	3.7334	0.0533	0.473
13	1.6127	0.2041	0.62	0.0535	0.8172	1.098
14	3.6286	0.0568	0.477	1.3913	0.2382	1.618
15	0.5936	0.441	0.752	0.3962	0.5291	0.776

Table 6. Cont'd.

16	13.6453	0.0002	0.209 *	3.5754	0.0586	2.357
17	10.5413	0.0012	0.285 *	1.8267	0.1765	1.749
18	31.7051	0.0001	0.113 *	1.7814	0.182	1.8
19	24.776	0.0001	0.144 *	0.7962	0.3722	1.484
20	28.3234	0.0001	0.136 *	0.0437	0.8345	0.911
21	28.765	0.0001	0.129 *	1.0642	0.3023	1.566
22	24.4447	0.0001	0.197 *	1.2465	0.2642	0.642
23	41.649	0.0001	0.090 *	1.5655	0.2109	1.709
24	35.9381	0.0001	0.135 *	2.1945	0.1385	0.536
25	58.4206	0.0001	0.066 *	0.1121	0.7377	1.155

$p<0.05$.

Table 7. Percent of item flagged as biased.

Type of DIF	DIF case				No DIF case			
	MH	SIB	AREA	LR	MH	SIB	AREA	LR
DIF cases	0.92	0.92	0.76	0.76	0.04	0.04	0.04	0.00
False positives	0.04	0.04	0.00	0.00	0.04	0.04	0.04	0.00

methods had the largest error rate.

When establishing a test, researchers sometimes wrongly assume that subjects in the groups have the same underlying unidimensional distribution even when it is a multidimensional distribution. This assumption is likely to be violated once the numbers of examinees and items are large. This study showed that practitioners need to be aware of multidimensionality in their data and if necessary they should use multidimensional models to estimate parameters before applying any of the DIF procedure to understand the behavior of an item (Ackerman, 1992). Otherwise, in the presence of multidimensionality several valid items may have been detected by DIF procedures and can be wrongly eliminated from the test. Test construction is a very time consuming and expensive task, and false positives are challenging.

Recently, Shealy and Stout (1993) developed a rigorous mathematical model for DIF (MMD; Shealy and Stout, 1993). This model explains the causes of DIF from the multidimensionality approach. Their DIF approach can be adopted to understand the underlying latent ability structure in the data and perform better DIF studies. This research was limited to examining the unidimensional DIF. Further studies might examine non-uniform DIF and other distributional features that might have lead to DIF.

Conflict of Interests

The authors have not declared any conflict of interests.

REFERENCES

Ackerman TA (1992). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. J. Educ. Measure. 29(1):67-91.

Agresti A (1990). Categorical Data Analysis. New York : Whiley.

Bock RD (1975). Multivariate Statistical Methods. New York: McGraw-Hill.

Clauser BE, Mazor KM (1998). Using statistical procedures to identify differential functioning test items. Educ. Measure. Issues Practice 31-44.

Crocker L, Algina J (1986). Introduction to Classical Modern Test Theory. Rinehard and Winston Inc. United States.

DeMars CE (2009). Modifcation of the Mantel-Haenszel and logistic regression DIF procedures to incorporate the SIBTEST regression correction. J. Educ. Behav. Statist. 34:149-170.

Erdem Keklik D (2014). Comparison of Mantel-Haenszel and Logistic Regression Techniques in Detecting Differential Item Functioning. J. Measure. Eval. Educ. Psychol. 5(2):12-25.

Güler N, Penfield RD (2009). A Comparison of Logistic Regression and Contingency Table Methods for Simultaneous Detection of Uniform and Nonuniform DIF. J. Educ. Measure. 46(3):314-329.

Gommez-Benito J, Navas-Ara MJ (2000) A Comparison of chi 2, RFA and IRT based procedures in detection of DIF. Quality Q. 34(1)17-31.

Hambleton KR, Swaminathan H (1985). Item Response Theory Principles and Application. Nijhoff Publishing.

Hambleton KR, Swaminathan H (1989). Detecting potentially biased test items: comparison of IRT area and Mantel-Haenszel methods. Appl. Measure. Educ. 2(4):313-334.

Linn RL, Levine MV, Hastings CN, Wardrop JL (1981). An investigation of item bias in test of reading comprehension. Appl. Psychol. Measure. 5:159-173.

Messick S (1989). Validity. Linn, Robert L. (ED). Educational Measurement (3rd ed.). The American council on education / Macmillan series on higher education. pp. 13-103.

Miller TR, Spray JA (1993). Logistic discrimination function analysis for DIF identification of polytomously scored Items. J. Educ. Measure. 30(2):107-122.

Millsap R, Everson HT (1993). Methodology review: Statistical approach

for assessing measurement bias. Appl. Psychol. Measure. 17(4):297-334.
Narayan P, Swaminathan H (1996) Identification of items that show nonuniform DIF. Appl. Psychol. Measure. 20(3):257-274.
Raju N (1990). Determining the significance of estimated signed and unsigned areas between two item response functions. Appl. Psychol. Measure. 14(2):197-207.
Rogers HJ, Swaminathan H (1992).A comparison of Logistic Regression and Mantel-Haenszel procedures for detecting differential item functioning. Appl. Psychol. Measure. 17(2):105-116.
Roussos LA, Stout WF (1996). A multidimensionality-based DIF analysis paradigm. Appl. Psychol. Measure. 20(4):355-371.
Sheally R, Stout W (1991). A Procedure to detect test bias present simultaneously in several Items (Tech. Rep. No, 91-3-ONR). Champaign: University of Illinois.
Shealy R, Stout WF (1993). An item response theory model for test bias. In P. W. Holland & H. Wainer (Eds.), Differential item functioning (197-239). Hillsdale NJ: Erbaum.
Stout W, Froelich AG, Gao F (2001). Using resampling methods to produce an improved DIMTEST procedure. In A. Boomsma, M.A.J. van Duijn, & T.A.B. Snijders (Eds), Essay on Item Response Theory New York: Springer-Verlag. pp. 357-375.
Swaminathan H, Rogers HJ (1991). Detecting differential item functioning using logistic regression procedures. J. Educ. Measure. 27(4):361-370.
Walker CM, Beretvas SN (2001). An empirical investigation demonstrating the multidimensional DIF paradigm: A cognitive explanation for DIF. J. Educ. Measure 38(2):147-163.
Whitmore ML, Schumacker RL (1996). A comparison of logistic regression analysis of variance differential item functioning detection methods. Educ. Psychol. Measure 59(6):910-927.