THE PROBABILITY DIFFERENCE INDICES AND EMPIRICAL SAMPLING DISTRIBUTION FOR DIF INDICES FOR IDENTIFYING ITEM BIAS IN MULTIDIMENSIONAL ITEM RESPONSE THEORY

Badrun Kartowagiran (badrunkw@yahoo.com)
Heri Retnawati (retnawati_heriuny@yahoo.co.id)

(Yogyakarta State University Indonesia)

Abstract. One condition of the test administration is fair. A test has the fair character if there isn't difference of probability of testee from different group in the same ability to answer the item of the test correctly. If there is difference of probability of the two groups to answer the item correctly, the item loads bias or differential item functioning (DIF). In Multidimensional Item Response Theory (MIRT), the existence of DIF can be known by probability difference indices. The significance of this measurement can be held using the empirical sampling distribution for DIF indices, by dividing randomly the focal group (F) and the reference group (R) becomes 2 group, for example F1 and F2, and R1 and R2, then calculates the probability difference indices between F1 and F2, and between R1 and R2. This article studied about identifying DIF using the probability difference indices in MIRT and tests its significance using the empirical sampling distribution for DIF indices.

Key word: MIRT, DIF, probability difference indices, empirical sampling distribution for DIF indices

1. Acknowledgements

In an assessment process, ideally there should not be any error, either random or systematic error, more specifically, there should not be any error coming from the test takers, test administrasion, and test items. The instrument which is used for measuring should be valid, reliable, stable and impartial. This means that no person or group gets disadvantaged by the presence of impartial test items. If there is difference of probability of the two groups to answer the item correctly, the item loads bias or differential item functioning (DIF).

Unfortunately, things do not happen the way they are expected to happen. Often times, test scores fail to give the correct information about test takers' abilities. This may be due to the fact that the information does not touch into the quantity or dimension to be measured. It may also be that other quantities or dimensions overlap with the intended quantities or dimensions so that the test results are misleading. It may also be due to the fact that the test is not properly administered so that it does not produce the correct information.

There are many methods for identifying DIF have been developed by psychometric researcher using item response theory. The theory has two assumptions that are local independent and unidimension.

The definition of unidimensional test is a test is measuring only single ability. It can be shown by test only measures the dominant component of testees' ability. Practically the assumption is difficult to be fulfilled tightly. The most educational and psychological tests is multidimensional, because the tests are not only measuring the dominant component, but also other component (Bolt and Lall, 2003; Ackerman, et. al., 2003). In this situation, the item analysis using unidimensional approach has been inappropriate again, and will result a systematic error and the informations obtained will mislead.

By paying attention to the characteristics of the tests are multidimensional, researcher can use multidimensional item response theory (MIRT). This theory can be used for analysis items of tests, including identifying differential item functioning. This paper is studied about detecting DIF using Simple Area Indices in unidimensional item response theory that is developed to Simple Volume Indices in Multidimensional Item

Response Theory for a multidimensional test. The significance of DIF in an item of test can be estimated by likelihood ratio test.

2. Solution

In the unidimensional item response theory, the relation between items parameters that are item difficulty index, item discriminating index, and *pseudo guessing* index and ability is expressed by equation of probability to answer the item correctly. Mathematically, the three parameters logistic model can be expressed as follows (Hambleton and Swaminathan, 1985: 49; Hambleton, Swaminathan, and Rogers, 1991: 17).

$$P_i(\theta) = c_i + (1-c_i)\,\frac{e^{Da_i(\theta-b_i)}}{1+e^{Da_i(\theta-b_i)}} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ (1)$$

Where:

θ : Testee's ability

$P_i(\theta)$: the testee probability at θ to answer i item correctly

a_i : item discriminating index for item-i

b_i : item difficulty index for item-i

c_i : pseudo guessing index for item-i

n : the number of item in the test

D : scaling factor (= 1.7).

Item difficulty index for item-i (b_i parameter) is a point at the scale of ability on characteristic curve when the testee probability to answer correctly is 50%. Item discriminating index for item-i (a_i) is slope of a tangent line at $\theta = b$. Pseudo guessing index is a probability of testee in low ability to answer item correctly. The testee ability (θ) is usually located in (- 3.00, +3.00), according to the area of normal distribution.

The two parameters logistic model and the one parameters logistic model are cases of the three parameters logistic model. When the pseudo-guessing index equals with 0 (c=0), the three parameters logistic model is become the two parameters logistic model. In the two parameters logistic model, when the item discriminating index is 1, the model become the one parameter logistic model or it's called with the Rasch model.

In the *multidimensional item response theory, MIRT*, there are two models, *compensatory* and *noncompensatory*. In the compensatory model, a testee who has lower ability in one dimension get compensation as higher ability in another dimension (Spray, at all., 1990), related with probability to answer item correctly. On the contrary, in the noncompensatory model, testee doesn't enable to have lower ability at one dimension get compensation as higher ability in other dimension. In the two dimension compensatory model as an example, a testee who has very low ability in one dimension and very hight ability in other dimension can answer an item of test correctly.

There are two type *compensatory model*, they are logistic MIRT (Reckase, 1997) and normal *ogive* model from Samejima, by expressing linear combination from multidimensional ability in the probability formula to answer item correctly. This model is also called with linear MIRT model (Spray, et. al., 1990; Bolt and Lall, 2003), that is a multivariate logistic regression. Model MIRT linear logistics can be written as:

$$P_i(\boldsymbol{\theta}_j) = c_i + (1-c_i)\,\frac{e^{[\sum_{m=1}^{k} f_{ijm}]+d_i}}{(1+e^{[\sum_{m=1}^{k} f_{ijm}]+d_i})} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ (2)$$

where $f_{ijm} = \mathbf{a}_{jm}{}' \, \boldsymbol{\theta}_{im}$, c_i is *pseudo-guessing* parameter of item-i, $\mathbf{a}_{jm}$ is item discriminating index for i-item at m-dimension, d_i is item difficulty index of i-item, and $\boldsymbol{\theta}_{jm}$ is m-element of j-testee's ability vector ($\boldsymbol{\theta}_j$). Like as in the unidimensional IRT, in the compensatory MIRT model, there are 3 parameters of item, called item discriminating index, item difficulty index, and pseudo-guessing index.

On the other hand, noncompensatory MIRT model is expressed as

$$P_i(\boldsymbol{\theta}_j) = c_i + (1-c_i)\prod_{m=1}^{k}\frac{e^{(\theta_j - b_{ij})}}{(1+e^{(\theta_j - b_{ij})})} \text{ ....(3)}$$

where b_{ij} is the difficulty parameter of i-item at m-dimension. Because of its form is the multiplication result, this model is also called as multiplicative model. This paper will only discuss the application of the compensatory MIRT model for identifying differential item functioning.

Item bias or differential item functioning is defined as the difference of the probability to answer item correctly between two groups of testees named as Focal group and Reference group (Angoff, 1993; Lawrence, 1994; Hambleton & Rogers, 1995). In unidimensional IRT, DIF is expressed as difference of the probability to answer item correctly between the Reference (F) and the Focal (F) group or probability in the Reference is subtracted by probability in the Focal group. Using the probability difference, the DIF measurement can be expressed by index.

This index characterizes item bias as a difference in the probability of giving a correct answer to a test item for individuals who have the same abilities but who come from different groups. Camilli and Shepard (1994) mathematically represent a probability difference indices as

$\Delta P_j = P_k(\theta_j) - P_p(\theta_j)$ .... (4)

where ΔP_j is the probability difference of the $P_k(\theta_j)$ j th testee from group P; $P_k(\theta_j)$ stands for the probability of the j th testee from k-group with ability θ to give the correct answer to an item based on the item parameter on the k-group scale; and $P_p(\theta_j)$ stands for the probability of the j th testee from group P with ability θ to give the correct answer to an item based on the item parameter on the group P scale.

The probability difference indices, Signed Probability Difference controlling for θ can be expressed as:

$$SPD-\theta = \frac{\sum_{j=1}^{n_p}\Delta P(\theta_j)}{n_p} \text{ ....(5)}$$

The equation is used to determine the difference probability of two groups that the item characteristic curves don't cross each other, like in the Figure 1 below.

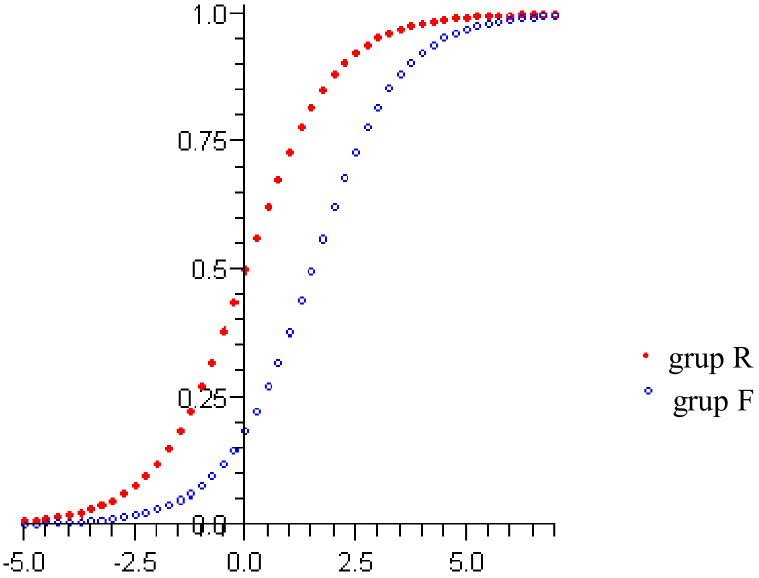

Figure 1. Item characteristics curves to determine the difference probablity of two groups (uniform DIF)

For item that the item characteristic curves of the two group cross each other, the DIF measurement is expressed by Unsigned Probability Difference controlling for θ in equation 6.

$$UPD-\theta=\sqrt{\frac{\sum_{j=1}^{n_p}\left(\Delta P(\theta_j)\right)^2}{n_p}} \quad \text{..} (6)$$

Where n_P is the number of testees in *P* group, Δ $P_p(\theta_j)$ is difference of probability to answer correctly of *j-testee* from *Pgroup*, *SPD-θ* stands for *signed probability difference* controlled by θ for *j-testee*, and *UPD-θ* stands for *unsigned probability difference* controlled by θ for *j-testee* (Camilli dan Shepard, 1994). The *SPD-θ* measure can be positive and negative that can be canceled each other. *UPD-θ is* comulative measure, then this is not cancelled each other to indicate the difference of the item characteristic curves. The positive *DIF's measure* is indicated that *P group in unfavored condition in the item. The indication wether there is any cross in the item characteristic curves is UPD-θ index is exceeded* the *SPD-θ index*. The small difference betwee *UPD-θ* and *SPD-θ* practically show the cross between the two item characteristic curves is not significant.

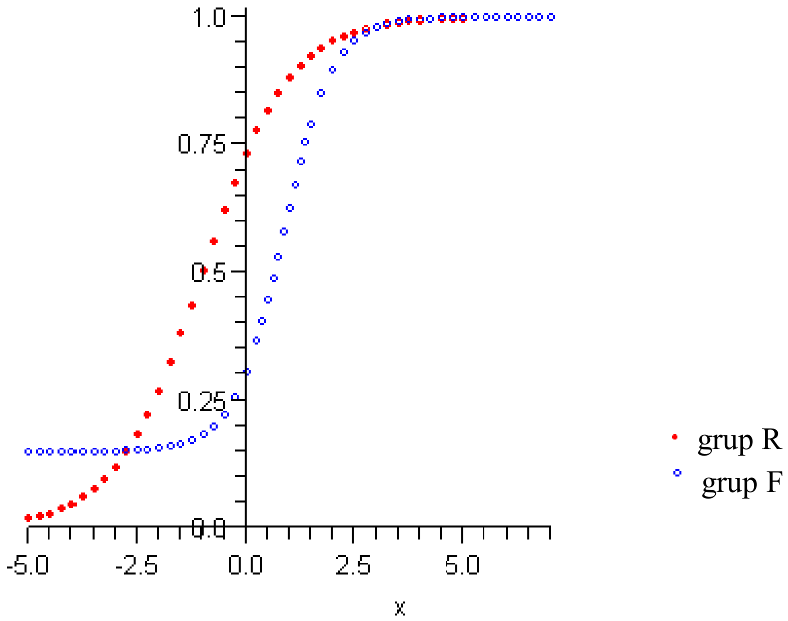

Figure 2. Item characteristics curve to determine the difference probablity of two groups (nonuniform DIF)

Using the definition of DIF, the difference of probability between Reference group and Focal group to answer item correctly, this concept can be used to identifying DIF in multidimensional IRT. The probability difference indices for identifying DIF in an item measuring k-dimension, expressed as

$$\Delta P_j = P_k(\theta_{1j},\ \theta_{2j},\ \ldots,\theta_{kj}) - P_p(\theta_{1j},\ \theta_{2j},\ \ldots,\theta_{kj}) \quad \text{...} (7)$$

$$SPD-\theta=\frac{\sum_{j=1}^{n_p}\Delta P(\theta_{1j},\theta_{2j},\ldots,\theta_{kj})}{n_p} \quad \text{..} (8)$$

$$UPD-\theta=\sqrt{\frac{\sum_{j=1}^{n_p}(\Delta P(\theta_{1j},\theta_{2j},\ldots,\theta_{kj}))^2}{n_p}}\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (9)$$

The significance of the measure DIF indices in an item can be known by Empirical sampling distributions for DIF indices (Camilli and Shepard, 1994). In the methods, the Reference and the Focal Group were both randomly split into halves, giving sub sample R1 and R2, F1 and F2 and treated these equivalent groups like Reference and Focal Groups. Because these groups were formed by random assignment, there should have been no DIF in any item. The DIF indices obtained for these two groups still showed variance owning to sampling error. The extreme values for R1-R2 and F1-F2 analysis were then chosen as critical values for signifying DIF.

For example, we analysis items of the mathematics test of National Examination for Junior High School in 2005. Using exploratory factor analysis, we can get information that the test measure two dimensions of mathematical ability. We divided testees in two group, female group as focal group and male group as reference group. After estimated the items' parameters using TESTFACT software, and trough *equating process*, the first item parameter for reference group is d= -0.777, a_1=0.800, a_2= 0.119, and c= 0.083, while for Focal group is d= -0.788, a_1=0.826, a_2=-0.027, and c=0.083 . The characteristic surfaces for the first item for the two groups are shown in Figure 3.

The result is SPD index of first number is about 0.000763, it means the first item favor the reference group or favor male group. The probability difference between focal group is Figure 4.

The sixth item parameters for reference group is d= 1.551, a_1=0.622, a_2= -0.271, and c= 0.270, while for focal group is d= 1.105, a_1=0.444, a_2=-0.162, and c=0.284. The characteristic surfaces for the first item for the two groups are shown in Figure 5.

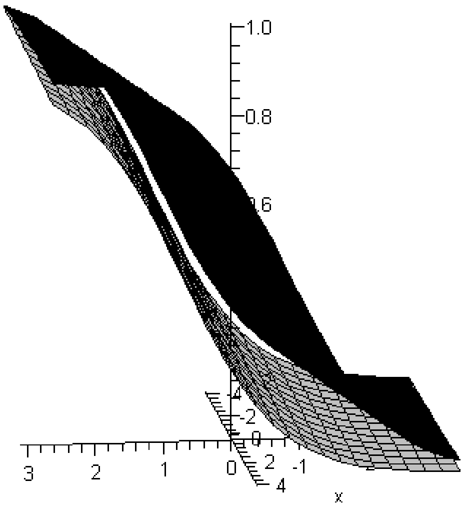

Figure 3. The characteristics surfaces for the first item for the two group

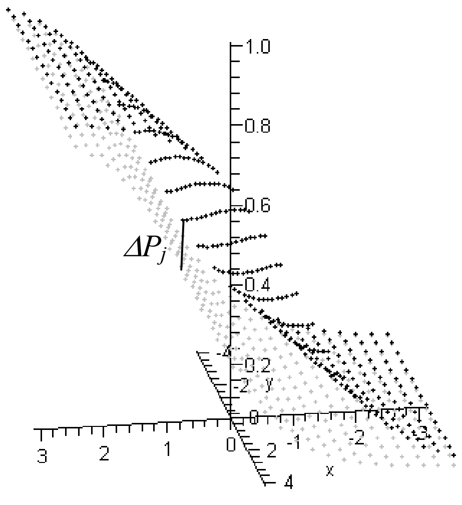

Figure 4. The probability difference between the focal group and reference group for the first item

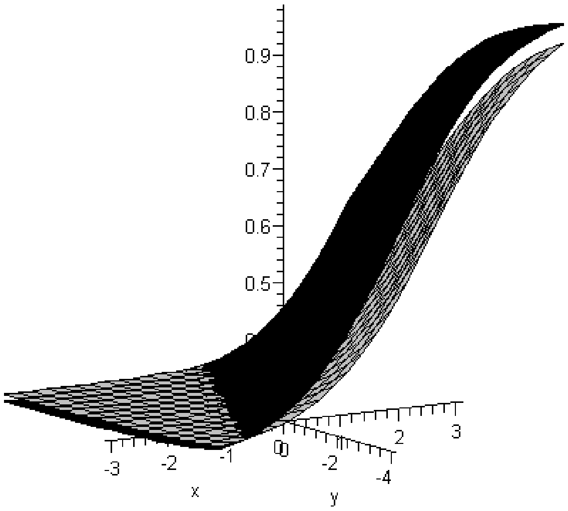

Figure 5. The characteristics surfaces for the sixth item for the two group

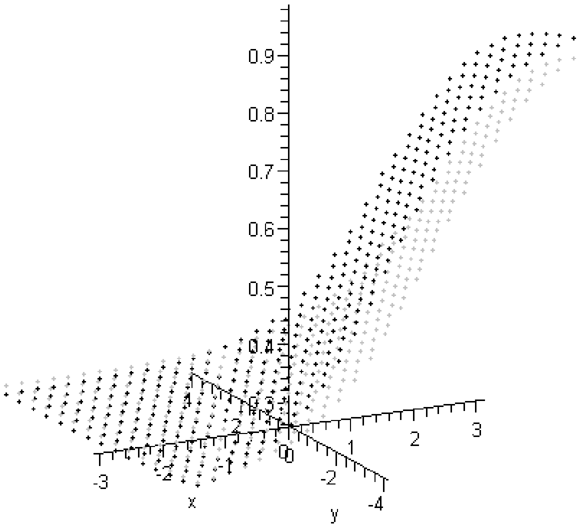

Figure 6. The probability difference between the focal group and reference group for the sixth item

The UPD index for the item number 6 is 0.045509, it means the item favor the reference group (male) in an area of ability but favor focal group in another area of ability. The probability difference of the two group is shown in Figure 6.

To test the significance of the SPD index for the first item, and UPD index for item number 6, the Reference and the Focal Group were both randomly split into halves, giving sub sample R1 and R2, F1 and F2 and treated these equivalent groups like Reference and Focal Groups. The result is expressed in the table 1 below.

Table 1. Significance test for UPD and SPD indices

No. Item	Indices Group		Indices Sub-Group		Maximum	Conclusion
1	SPD R-F	0.000763	SPD R1-R2	0.01511	0.01511	Not Significant
			SPD F1-F2	0.011593		
6	UPD R-F	0.045509	UPDR1-R2	0.090462	0.090462	Not Significant
			UPD F1-F2	0.062682		

From the table, the conclusion is there isn’t difference of probability of testee from different group in the same ability to answer the first item and the sixth item of the test correctly or there is no DIF in the items.

3. Conclusion

In Multidimensional Item Response Theory, the existence of DIF can be known by probability difference indices, they are Signed Probability Difference controlling for ability (SPD-θ) if the item characteristics curves of focal group and reference group don’t cross each other and Unsigned probability difference controlling for ability (UPD-θ) if the item characteristics curves of focal group and reference group cross each other. The significance of this measurement can be held using the empirical sampling distribution for DIF indices, by dividing randomly the focal group (F) and the reference group (R) becomes 2 group, for example F1 and F2, and R1 and R2, then calculates the probability difference indices between F1 and F2, and between R1 and R2. The extreme values for R1-R2 and F1-F2 analysis were chosen as critical values for signifying DIF.

Reference

Ackerman, T.A., et. al. (2003). Using multidimensional item response theory to evaluate educational and psychological tests. *Educational Measurement*, 22, 37-53.

Angoff, W.H. (1993). Perspectives on differential item functioning methodology. Dalam P.W. Holland dan H. Wainer (Eds.), *Differential item functioning*. Hillsdate, NJ: Erlbaum, Pp. 3 – 23.

Bolt, D.M. & Lall, V.M. (2003). Estimation of compensatory and noncompensatory multidimensional item response models using Marcov chain Monte-Carlo. *Applied Psychological Measurement*, 27, 395-414.

Camilli, G. & Shepard, L.A. (1994). *Methods for identifying biased test items*, *Vol.4*. London: Sage Publications, Inc.

Hambleton, R.K. & Rogers, H.J. (1995). Developing an item bias review form. From http://www.ericcae.net/ft/tamu/ biaspub2.htm March 10, 2007.

Hambleton, R.K., Swaminathan, H., & Rogers, H.J. (1991). *Fundamentals of item response theory*. London: Sage Publications, Inc.

Hambleton, R.K. & Swaminathan. (1985). *Item response theory*. Boston, MA: Kluwer Nijjhoff, Publisher.

Spray, J.A., et. al. (1990). Comparison of two logistic multidimensional item response theory models. *ACT Research Report Series*. United States Government.