

IDENTIFYING ITEM BIAS USING THE SIMPLE VOLUME INDICES AND ITEM RESPONSE THEORY LIKELIHOOD RATIO (IRT-LR) TEST ON POLYTOMOUS DATA

Heri Retnawati (retnawati_heriuny@yahoo.co.id)

Mathematics Education Department
Mathematics & Science Faculty Yogyakarta State University Indonesia

Abstract. In the item response theory (IRT), the relation of probability of testees' to answer correctly an item of test with items' and abilities' parameter is expressed in a linear logistic equation. When an item of a test gives different opportunity to answer correctly at two groups of testees' at the same abilities, the item loads bias or differential item functioning (DIF). The measurement of DIF can be estimated by difference of probability to answer correctly at two groups, called with focal group and reference group. In the polytomous data, this difference can be known using simple volume indices. The significance of DIF measurement between two the groups is tested by using likelihood ratio test, between compact model that loads the item is studied and overall model that loads all of item. This article is about identifying DIF using the simple volume indices in IRT and tests its significance using IRT likelihood ratio test.

Key word: MIRT, DIF, simple volume indices, likelihood ratio test

1. Acknowledgements

The ideal condition of the administration of test is objective, transparency, fair and not discriminative. If a test contains any items favoring a group of testees', the test is unfair, containing item bias or differential item functioning, DIF. There are many methods for identifying DIF have been developed by psychometric researcher using item response theory. The theory has two assumption 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 fulfill 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 one-dimensional approach has been inappropriate again, and will result a systematic mistake 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).

$$P_1(\theta) = c_i + (1-c_i) \frac{e^{Da_i(\theta-b_i)}}{1 + e^{Da_i(\theta-b_i)}} \dots\dots\dots (1)$$

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 item in test
- D : scaling factor (= 1, 7).

Item difficulty index for item-I (b_i parameter) is a point at the scale of ability on characteristics 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 $\theta = b_i$. Pseudo guessing index is a probability of testee in low ability to answer item correctly. The tested 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 the higher ability in another dimension (Spray, at all., 1990), related with probability to answer item correctly. On the contrary, in the no compensatory model, tested doesn't enable to have high ability at one dimension get compensation at low ability in other dimension. In the two dimension compensatory model, a tested who has very low ability in one dimension and very height 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 model MIRT linear (Spray, et all., 1990; Bolt and Lall, 2003), which is a multivariate logistic regression. Model MIRT linear logistics can be written as:

$$P_i(\theta) = c_i + (1-c_i) \frac{e^{\left[\sum_{m=1}^k f_{ijm}\right]+d_i}}{\left(1 + e^{\left[\sum_{m=1}^k f_{ijm}\right]+d_i}\right)} \dots\dots\dots(2)$$

where $f_{ijm} = a_{jm} \theta_{jm}$, c_i is *pseudo-guessing* parameter of item-i, a_{jm} is item discriminating index for i-item at m-dimension, d_i is item difficulty index of i-item, and θ_{jm} is m-element of j-testee's ability vector (θ_j). Like 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

$$P_i(\theta) = c_i + (1-c_i) \prod_{m=1}^k \frac{e^{(\theta_j - b_{ij})}}{\left(1 + e^{(\theta_j - b_{ij})}\right)} \dots\dots\dots(3)$$

where b_{ij} is the difficulty parameter of i-item at m-dimensions. 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 and the Focal group or probability in the Reference is subtracted by probability in the Focal group. Because the measurement of DIF is the scale of "how difference" between the two group, in the characteristic curves of items in the two groups is signed by

the brown area, at the Figure 1. The area is called by SIGNED-AREA, and the measure can be estimated mathematically using integration methods, expressed in equation 4. The DIF measurement is related with a simple area in the characteristic curves, and then Camilli and Shepard (1994) called the methods as *Simple Area Indices*.

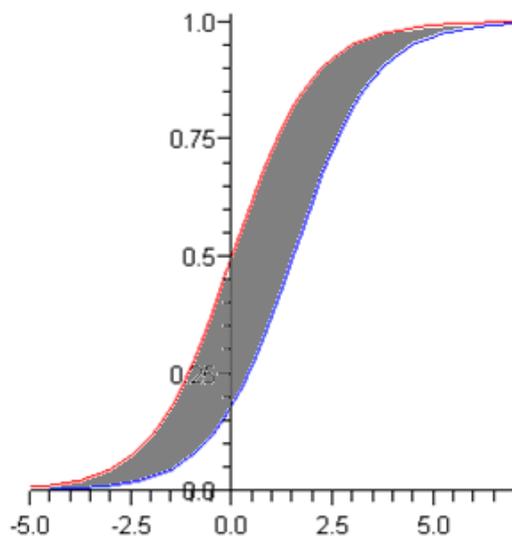


Figure 1. The characteristic curves of item in the two groups (without cross)

$$\text{SIGNED-AREA} = \int [P_R(\theta) - P_F(\theta)] d\theta \dots\dots\dots(4)$$

At the picture 1 above, the two characteristic curves are consistent, not across each other. Because the area between the two characteristic curves is the integration of the difference of the probability to answer item correctly between the Focal group and Reference group, if the performance different favor the Focal Group, this area measure will be negative. On the contrary, if the performance different favor the Reference group, this area measure will be positive.

In the DIF analysis, it could be happened the two characteristic curves across and it is shown the two characteristic curves are not consistent performance. In this case, the positive and negative areas in different regions of the graph will cancel each other (Figure 2). When an index of total ICC discrepancy is desired, the integral can be evaluated using the squared probability difference is expresses as UNSIGNED-AREA in equation 5.

$$\text{UNSIGNED-AREA} = \sqrt{\int [P_R(\theta) - P_F(\theta)]^2 d\theta} \dots\dots\dots(5)$$

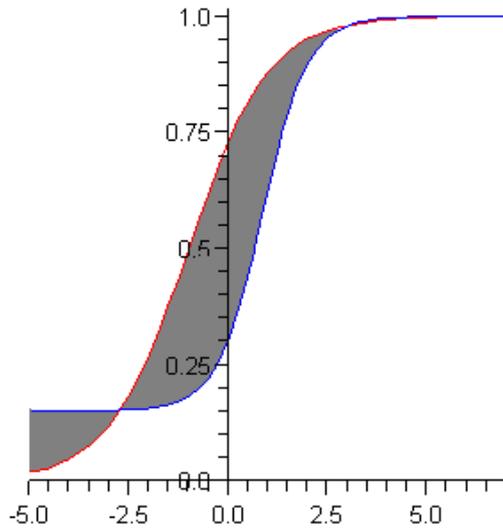


Figure 2. The characteristic curves of item in the two groups (with cross)

Using the definition of DIF, this concept can be used to identifying DIF in multidimensional IRT. The area which is the probability difference of the two characteristic curves called Simple Area Indices in one-dimensional item response theory is developed to Simple Volume Indices in Multidimensional Item Response Theory for a multidimensional item. If an item measures two ability dimension, for example θ_1 and θ_2 , the relation between the parameters' item, abilities and probability can be drawn as item characteristics surfaces, like in the Figure 3.

The area between the two surfaces is called by SIGNED-VOLUME. The measurement of the volume can be estimated using multiple integration.

$$\text{SIGNED-VOLUME} = \int \int [P_R(\theta_1, \theta_2) - P_F(\theta_1, \theta_2)] d\theta_1 d\theta_2 \dots\dots\dots (6)$$

From the Figure 3, the two characteristic surfaces are consistent, not across each other. Like in the simple area indices, if the performance different favor the Focal Group, this area measure will be negative. On the contrary, if the performance different favor the Reference group, this area measure will be positive. When the two characteristic surfaces across and it is shown the two characteristic surfaces are not consistent performance. The integral to estimate the measurement of DIF can be evaluated using the squared probability difference is expresses as UNSIGNED-VOLUME in equation 7.

$$\text{UNSIGNED-VOLUME} = \sqrt{\int \int [P_R(\theta_1, \theta_2) - P_F(\theta_1, \theta_2)]^2 d\theta_1 d\theta_2} \dots\dots\dots (7)$$

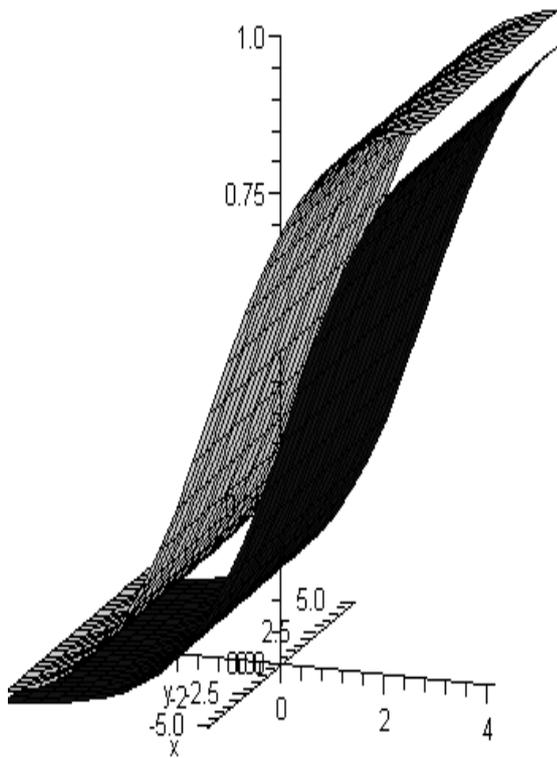


Figure 3. The characteristic surfaces of item in the two groups (without cross)

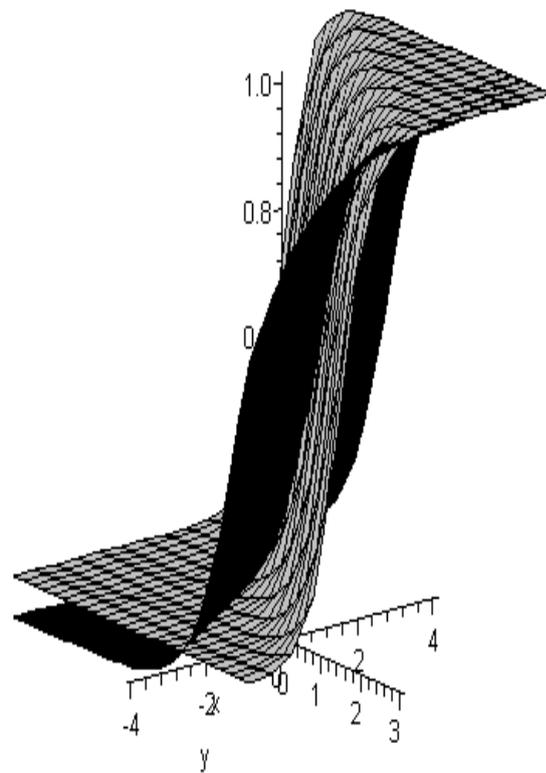


Figure 4. The characteristic surfaces of item in the two groups (with cross)

The resemblance of these characters is applicable to measuring DIF more than 2 dimensions, but the characteristic surfaces cannot be drawn again. If an item is measuring k-ability, SIGNED-VOLUME and UNSIGNED-VOLUME can be determined using equation:

$$\text{SIGNED-VOLUME} = \int \int \dots \int [P_R(\theta_1, \theta_2, \dots, \theta_k) - P_F(\theta_1, \theta_2, \dots, \theta_k)] d\theta_1 d\theta_2 \dots d\theta_k \dots (8)$$

$$\text{UNSIGNED-VOLUME} = \sqrt{\int \int \dots \int [P_R(\theta_1, \theta_2, \dots, \theta_k) - P_F(\theta_1, \theta_2, \dots, \theta_k)]^2 d\theta_1 d\theta_2 \dots d\theta_k} \dots (9)$$

The significance of DIF in an item of test can be estimated by likelihood ratio test or the model comparison approach (Camilli and Shepard, 1994: 76-97; Thissen et. al., 1993: 72). The model comparison approach is implemented by comparing the relative fit of two models. The first is called the compact model (C) and the second is the augmented model (A). The augmented model is an elaboration of the compact model; it has all of the parameters of model C plus a set of additional parameters. The comparison is to see whether the additional parameters in set A are really necessary. A simpler model with a single ICC for the Reference and Focal Group (for a particular item) is always preferable to a more complex model with each group has its own ICC. Only if a more complex model demonstrates significantly better fit to the data, its additional features is deemed necessary.

For illustration, L^* is value of L likelihood function. There are two models will be compared, C model (*compact model*) and A model (*augmented model*). The C Model is simpler model. The hypothesis is

$$H_o : \Gamma = \text{Set}_C \text{ (Set}_C \text{ contains N parameters) } \dots\dots\dots (10)$$

$$H_a : \Gamma = \text{Set}_A \text{ (Set}_A \text{ contains N+M parameters) } \dots\dots\dots (11)$$

Γ Stands for the true set parameters. The C model has M fewer parameters than A model. The *Likelihood Ratio for the two models is:*

$$LR = \frac{L_{(C)}^*}{L_{(A)}^*} \dots\dots\dots (12)$$

Where $L_{(C)}^*$ is value of L likelihood function of C model and $L_{(A)}^*$ is value of L likelihood function of A model .

Then it is transformed in the natural logarithm:

$$\begin{aligned} \chi_{(M)}^2 &= -2 \ln(LR) \\ &= [-2 \ln L_{(C)}^*] - [-2 \ln L_{(A)}^*] \dots\dots\dots (13) \end{aligned}$$

For simplicity notation, $G(C) = [-2 \ln L_{(C)}^*]$ and $G(A) = [-2 \ln L_{(A)}^*]$, then the ratio likelihood becomes

$$\chi_{(M)}^2 = -2 \ln(LR) = G(C) - G(A) \dots\dots\dots (14)$$

$\chi_{(M)}^2$ Is approximately distributed as a chi-square with M degrees of freedom.

The steps for detecting DIF are can be elaborated as follows. Firstly, we estimate item's parameter and we get $G(C)$ in test consisted of k item. Second, we determine one of item test to be evaluated. Third, item the test is made impressing becomes two items. Item firstly contains answers from Reference group, which is not responded by the Focal group. The second item contained the answers from Focal group which is not responded by the Reference group. Fourth, we estimate the parameters again and get $G(A)$ for test consisted of k+1 item. Then, we can determine $\chi_{(M)}^2$ to know the significance of DIF in one item of test.

For example, we analysis items of the mathematics national examination for Junior High School in 2005 in Indonesia. 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 the Figure 5.

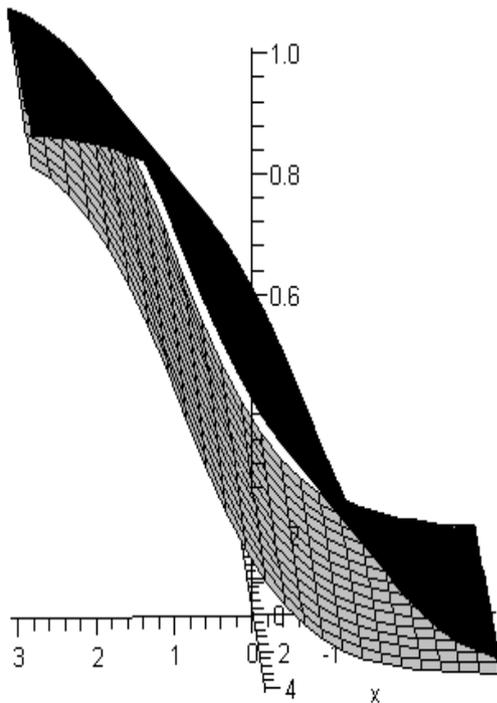


Figure 5. The item characteristics surface of the first number (the mathematics National Examination for Junior High School in 2005 in Indonesia)

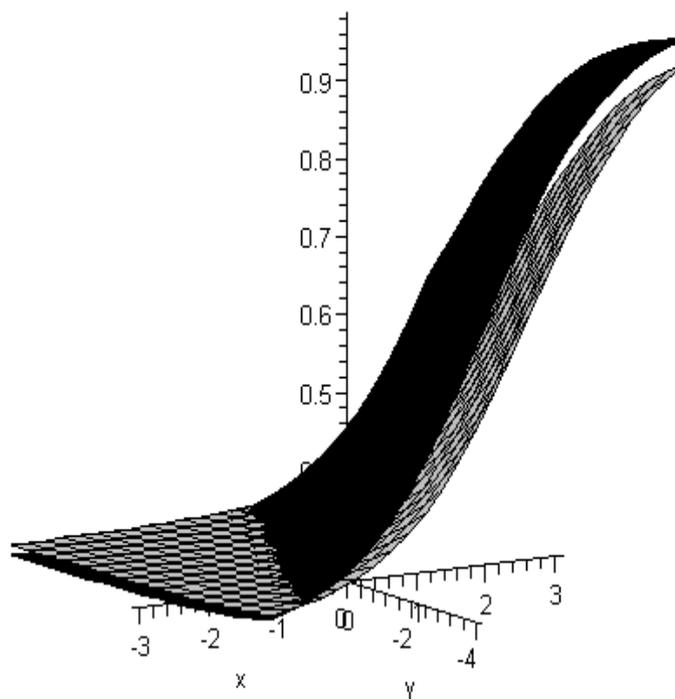


Figure 6. The item characteristics surface of the sixth number (the mathematics National Examination for Junior High School in 2005 in Indonesia)

The measurement of DIF that is SIGNED-VOLUME (the two surfaces don't cross each other) can be determined using multiple integral. For simplicity, the integration can be done using MAPLE software. The result of SIGNED-VOLUME for the first number item is 13.98723.

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 6.

The two surfaces across each other in the sixth number item, and then we used UNSIGNED-VOLUME for measure DIF. The result of determination of UNSIGNED-VOLUME is 18.06625.

Table 1. The result for determinations the significance of DIF

Item	G(A)	G(C)	$\chi^2_{(M)} = G(C) - G(A)$	$\chi^2_{(df=4, \alpha=0,010)}$	Conclusion
1	179775.5	183940.8	4165.305	13.28	Contains DIF Significantly
6	179775.5	194156.7	14381.23	13.28	Contains DIF Significantly

Likelihood of A model and C model has been estimated using TESTFACT. For the first item number and the sixth item number are presented at the following table. These results are shown that the first and sixth item of the mathematics national examination for Junior High School in 2005 contains DIF significantly.

3. Conclusion

The measurement of DIF can be estimated by the difference of probability to answer correctly at two groups, called with focal group and reference group. In the multidimensional data, this difference can be known using simple volume indices and determinate it's using multiple integral. The significance of DIF measurement between two the groups is tested by using likelihood ratio test, between compact model that loads the item is studied and overall model that loads all of items.

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