International Journal on Advanced Science Engineering Information Technology

Vol.6 (2016) No. 4
ISSN: 2088-5334

Detecting Differential Item Functioning and Differential Test Functioning on Math School Final-exam

Mansyur[#1], Muliana[#2]

[#] *Graduate Program, Universitas Negeri Makassar, Makassar, Indonesia*
E-mail: [1]mansyurunm@gmail.com; [2]mulianamat01@gmail.com

***Abstract*— This study aims at finding out the characteristics of Differential Item Functioning (DIF) and Differential Test Functioning (DTF) on school final-exam for Math subject based on Item Response Theory (ITR). The subjects of this study were questions and all of the students' answer sheets chosen by using convenience sampling method and obtained 286 responses consisted of 147 male and 149 female students' responses. The data of this study collected using documentation technique by quoting the response of Math school final-exam participants. The data analysis of this study was Item Response Theory approach with model 2P of Lord's chi-square DIF method. This study showed that from 40 question items analysed theoretically using Item Response Theory (ITR), affected Differential Item Functioning (DIF) gender was ten items and affected DIF location (area) was 13 items. Meanwhile, Differential Test Functioning (DTF) was benefitted for female and least profitable to citizen.**

***Keywords*— Differential Item Functioning (DIF); Differential Test Functioning (DTF), Item Response Theory (ITR).**

I. INTRODUCTION

School final-exam is one of the learning assessment processes that has important role and purpose in the education field. The outcomes of school final-exam based on score are expected to be useful in describing learners' ability as the successful indicator in the educational process. Because of the importance of school final-exam score, thus the questions should be arranged appropriately to assess the learning outcomes. The accuracy of assessment is necessary for validity. In this case, the different score between one student and other students is caused by their own ability, not caused by another factor like bias to the item test.

In education assessment, the bias term on items is recognized as differential item functioning (DIF) and differential test functioning (DTF) [1]–[9]. Various techniques or methods of DIF detection are found and used. Besides differential item functioning, differential test functioning also can be used to indicate whether or not a test is a fair for each level.

An evaluation of education is an activity or assessment process to see quality and results and systematic process to determine or make a decision, the extent to which the program objectives have been achieved [10]. Regarding the assessment aspects of math achievement test especially in learning math, a test is required as the instrument. The test consists of questions that have no right or wrong answers. The test is also defined as some issues that need answers, or some issues that require a response to measure the level of a person's ability or to reveal certain aspects of the test [11].

Item response theory is a mathematical model that relates the potential ability of examinees respond to a particular item as a whole [11]–[14], [7], [15]–[17]. This theory, in general, can be characterized as follows: 1) the characteristics of item do not depend on the examinees, 2) the score described by examinees does not rely on the test, 3) the model is more emphasized on the level of item than the degree of test, 4) the model does not require strictly parallel tests to assess the reliability, and 5) the model describes a size of each capability score that has no functional relationship between examinees and the level of capabilities.

Mathematical models in IRT say that the probability of a subject to answer an item accurately depends on the ability of the subject and grain characteristics. Thus, there is an assumption that IRT can be indirectly measured and proven. The assumptions underlying item response theory is unidimensional, local independence, and the function of item characteristics or item characteristic curve. Unidimensional means that a single test measured the dimensions of the character of the participants. In the context of achievement tests, dimensionalities referred as the number of abilities measured by tests or by a collection of items [12].

Parameters of item response theory are the level of difficulties that is symbolized as b, different power as a, and prediction as c. "Information of response pattern to a test or other instruments is used to estimate the magnitude of one's

ability, estimation of capabilities and item parameter used Maximum Likelihood (MLE) and Bayesian [18].

In addition to these three characteristics of item response theory, there is still one thing to note is the item information function. It is a method to describe the strength of an item on the questions that declared a latent capability or latent trait and is measured by the test. By recognizing the function of item information thus, the item can be matched with the model to assist in the selection of items.

At the end of school exam test requires a test that is not for particular groups of learners, both regarding tribal religion, gender and so on. Therefore, a test or question with a great information function is required to guarantee the quality of learners. The procedure in detecting bias item would determine whether or not the items would provide valid information. In this case, several ways are used to detect item and test bias so that the test device is fair to all participants of the test.

The detection of test bias (DTF) is based on the item response theory by looking at the graph of the value opportunity. As in DTF, the detection item bias (DIF) based on item response theory can be done with a variety of approaches. The first is difference test of item difficulty parameter. The second is the method of item deviation. The third is Lord chi-square test. The fourth approach is the empirical distribution sampling for DIF index. The last or the fifth is the comparison model of item response theory [3]. While, the model of comparison approach of item response theory is divided into four categories. The first category is a general introduction about the likelihood. The second is the ratio of likelihood based on inferential statistics. The third is the model approach of DIF comparative analysis. The fourth is the three parameters of item response theory"

The difference of parameter causes DIF occurs in two general categories: (1) consistent or uniform DIF that occurs when the characteristic curve item is different and indirectly intersect or cross, and (2) inconsistent or non-uniform DIF happens if the characteristic item curve is different but intersect at a scale of θ [3]. It can be viewed directly from the graph of the opportunity value to find out which group is better.

Question bank can be simply defined as a set of test items. However, the question bank is not just a collection of questions only. Those items in it are items that have been selected through a procedure or accurate information[19] .

The criteria of good quality items according to item response theory refers to each parameter item. Criteria for difficulty level (b), those items which have a value of more than 2 or b> 2 were items that were considered very difficult [12]. The item that is very complex cannot measure the function properly because the test taker tends to answer by using an educated prediction. The value of parameters for good difficulty level ranges from 0 to 2. Those items that have a value less than -2 parameters are items that are very easy and should be revised.

As for the question differentiator power criteria (a), Hanlbleton [12] described when the item was revised or discarded, while differentiator >2 is rare happened. So as a matter of differentiator power ranges from 0 to 2 indicates that the item was able to distinguish between high-ability test takers and test-takers that are less capable.

To detect differential item functioning (bias point) based on the Theory of Item Response logistic model 2, the parameter is the Lord's chi-square method. While in detecting the differential test functioning based on item response theory model of logistic 2, the parameter is to see a graph of the probability of each item depending on which groups are advantaged and disadvantaged. In this study, there are two types of DIF namely DIF by gender differences and DIF based on different locations. How do the characteristics of DIF and DTF on questions of school final-exam based on item response theory were the focus of this study.

II. METHOD

This study aims at finding out the bias of items (differential item functioning) and bias of test (differential test functioning) based on item response theory. This study used convenience sampling method based on the reason of ease access of population inclusion [20]. This study was conducted at Junior High School 20 (SMP Negeri 20), Bulukumba Regency, South Sulawesi, Indonesia for three months. Using the criteria of sample size ranging from 50, 200, and 1000 with extended test 10, 20, 80 [12], [17].

The data in this study was the response from math questions by the students in academic year 2013/2014 as the examinees. Sources of the data in the form of answer sheets of students who documented. There were three math teachers created the questions. This study involved 286 responses consisted of 137 responses from male student and 149 from the female.

Collecting data in this study was done by using documentation technique, by quoting the participants' responses of school final-exam on Math subject. The technique of data analysis in this study was item response theory approach using DIF detection approach of Lord's chi-square DIF method with 2P category [21].

III. RESULTS AND DISCUSSIONS

The analysis results of the validated item of knowledge (math achievement test) begin with the instrument designed by the validator. Validation of the contents by experts involved two experts in math. Both are the lecturers at Faculty of Math and Science, Universitas Negeri Makassar, Indonesia. From the assessment given by the validator showed that one item less relevant (cell A) was clause 14, in item 9 first validator gave a relevant assessment but the second validator considered as less relevant then this item was incorporated into cell B, in item 6 first validator gave irrelevant assessment but the second validator considered less relevant then this item has been integrated into the cell C and 37 items including very relevant (cells D), thus degree of validation can be calculated based on the formula of Gregory internal consistency model as follows.

Internal consistency coefficient:

$$= \frac{D}{(A+B+C+d)} \quad (1)$$

$$= \frac{37}{(1+1+1+37)} \quad (2)$$

$$= 0.93 \quad (3)$$

It is concluded that the validity obtained was 0.93 or V = 93%. It means that the results of the second assessment validator > 75%, so the criteria was strong relevancy [22].

A. Item Characteristic with 2P Logistics Model

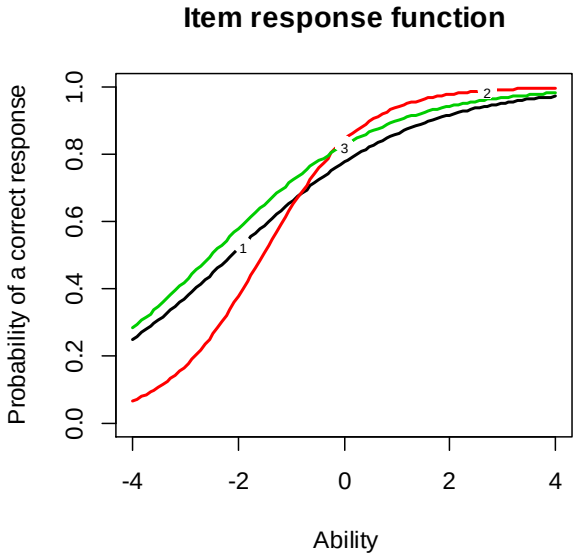

Fig. 1 Curve of item characteristic with 2P Logistics Model.

Based on the data obtained from the Figure 1 above, it is concluded that from 40 items that, they were analysed based on item response theory models 2P model there were 7.5% items with very good category and 30% as well as 62.5% of the poor category.

B. Item Difficulty Level with 2P Logistics Model

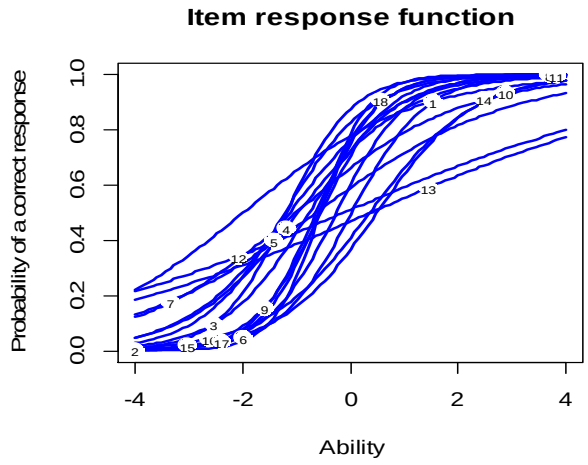

Fig. 2 Curve of item difficulty level with 2P logistics Model

Based on data obtained on the difficulty level of multiple choice questions based on the Figure 2 above showed that there were 3 items or 7.5% on the difficulty level of the test is easy categories, 3 items or 7.5% were in the moderate category, 12 items or 30% category is very easy, 5 items or 12.5% categorized as difficult, and 17 items or 42.5% categorized as extremely difficult.

C. Item Bias Based on Gender

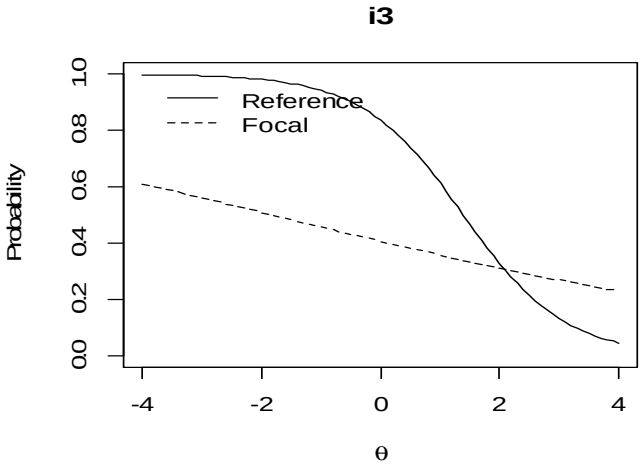

Fig. 3 Bias Chart of Item 3.

The figure 3 above showed that the level of ability of the two graphs intersect, this indicates that the degree of these capabilities as the theory described that DIF consistently occurs if the graph does not intersect, and inconsistent DIF appears if the graph intersected at one point and did not happen at any level or capability scale (ability). Therefore, in the figure above shows the DIF inconsistent (non-uniform) categorized bias based on gender.

D. Item Bias Based on Location (Region)

On bias based on location, it indicates the groups of examinees coming from the city as a focus group (focal) and examinees from the village as the reference group (reference). For more details, see each image bias following locations.

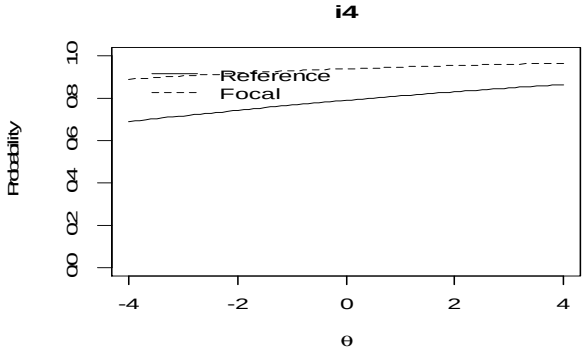

Fig. 4 Graph of bias based on location of item 4.

The Figure 4 above showed that the graph does not intersect at every level of ability. It indicates that the DIF occurs at every level of ability. Therefore, in the figure

above shows a consistent DIF (uniform). Based on the results of data analysis according to Item Response Theory 2 parameter logistic model by using the Lord's chi-square method 386.3.1.2 version showed that from 40 items, 3 items with the excellent category , 12 items with good categories, and 25 items with bad category. Thus, 15 items can be selected in the question bank development activities i.e., items 1, 2, 5, 6, 7, 9, 11, 12, 16, 22, 23, 28, 29. However, the differential item functioning and differential test functioning turns of 40 items that are detected as bias based on gender and item bias based on location (region).

From 40 items contained 10 items (25%) indicated to contain gender bias so that when seen from the graph the value of opportunity turns 6 items contain consistent biases and 4 items contain inconsistent bias which these items tend to benefit women's groups and disadvantaged groups of men. In addition to the question of gender bias was also detected bias locations. from 40 items contained 13 items (32.5%). It is indicated that gender bias when seen from the graph the value of opportunity the item 6 and 7 are consistent bias and inconsistent bias in which these items tend to benefit group of test participants located in cities, and disadvantaged groups of test takers are located in the village.

IV. CONCLUSIONS

Characteristics of item on school final-exam at Yunior High School level for Math subject detected ten items indicating item bias (DIF) based on gender and 13 item bias indicated by location (region). Characteristics of Math Exams questions detected test bias (DTF) 6 items inconsistent and consistently biased tests as much as six items based on gender biased for women, while based on the location contain consistent test bias six items and seven items contain inconsistent test bias based on the city.

Based on the results obtained in the study, there are several items which need to be considered related to bias detection point based on item response theory with 2P models on the method of Lord's chi-square statistic. To the authors math SMP either in local or national scale should use problems are empirically proven good quality and do not contain DIF and DTF. Empirically proving the quality of an item is needed to be further developed and disseminated to all practitioners of education for example with training for teachers both mathematic and other subject areas. To detect DIF and DTF a test can be carried out based on item response theory. To further expand the study of DIF and DTF is need for further research using other methods.

REFERENCES

[1] P. W. Holland and H. Wainer, Differential item functioning. Routledge, 2012.

[2] K. E. Ryan, “Methods for identifying biased test items,” *Eval. Pract.*, vol. 18, pp. 73–76, 1997.

[3] G. Camilli and L. A. Shepard, *Methods for identifying biased test items*. Sage, 1994.

[4] Y. Liu, B. E. Magnus, and D. Thissen, “Modeling and Testing Differential Item Functioning in Unidimensional Binary Item Response Models with a Single Continuous Covariate: A Functional Data Analysis Approach,” *Psychometrika*, vol. 81, no. 2, pp. 371–398, 2016.

[5] W. H. Finch, “Applied Measurement in Education Detection of Differential Item Functioning for More Than Two Groups : A Monte Carlo Comparison of Methods Detection of Differential Item Functioning for More Than Two,” vol. 7347, no. December 2015, 2016.

[6] B. Terluin, P. C. Unalan, N. Turfaner Sipahioğlu, S. Arslan Özkul, and H. W. J. Van Marwijk, “Cross-cultural validation of the Turkish Four-Dimensional Symptom Questionnaire (4DSQ) using differential item and test functioning (DIF and DTF) analysis,” *BMC Fam. Pract.*, vol. 17, no. 1, pp. 1–9, 2016.

[7] L. Tay, Q. Huang, and J. K. Vermunt, “Item Response Theory With Covariates (IRT-C): Assessing Item Recovery and Differential Item Functioning for the Three-Parameter Logistic Model,” *Educ. Psychol. Meas.*, vol. 76, no. 1, pp. 22–42, Feb. 2016.

[8] A. Cavanagh, C. J. Wilson, P. Caputi, and D. J. Kavanagh, “Symptom endorsement in men versus women with a diagnosis of depression: A differential item functioning approach,” *Int. J. Soc. Psychiatry*, 2016.

[9] Y. Cheng, C. Shao, and Q. N. Lathrop, “The mediated MIMIC model for understanding the underlying mechanism of DIF,” *Educ. Psychol. Meas.*, 2015.

[10] M. D. Miller, R. L. Linn, N. E. Gronlund, and D. Miller, *Measurement and Assessment in Teaching*, Tenth Edit. New Jersey: Pearson Education Inc., 2009.

[11] L. Crocker and J. Algina, *Introduction to classical and modern test theory*. ERIC, 1986.

[12] R. K. Hanlbleton, H. Swaminathan, and D. J. Rogers, *Fundamentals of Item Response Theory*. Sage Publications, Inc, 1991.

[13] N. K. Chadha, *Applied Psychometry*. B-42, Panchsheel Enclave, New Delhi 110 017 India: SAGE Publications India Pvt Ltd, 2009.

[14] L. Crocker and J. Alglna, *Introduction to Classical and Modern Test Theory*. 2008.

[15] R. M. van Nispen, D. L. Knol, M. Langelaan, and G. H. van Rens, “Re-evaluating a vision-related quality of life questionnaire with item response theory (IRT) and differential item functioning (DIF) analyses,” *BMC Med. Res. Methodol.*, vol. 11, no. 1, p. 125, 2011.

[16] B. P. Winterstein, T. A. Ackerman, P. J. Silvia, and T. R. Kwapil, “Psychometric properties of the wisconsin schizotypy scales in an undergraduate sample: Classical test theory, item response theory, and differential item functioning,” *J. Psychopathol. Behav. Assess.*, vol. 33, no. 4, pp. 480–490, 2011.

[17] B. P. Foley, “Improving IRT parameter estimates with small sample sizes : Evaluating the efficacy of a new data augmentation technique by,” 2010.

[18] M. E. Glickman, P. Seal, and S. V. Eisen, “A non-parametric Bayesian diagnostic for detecting differential item functioning in IRT models,” *Heal. Serv. Outcomes Res. Methodol.*, vol. 9, no. 3, pp. 145–161, 2009.

[19] J. Judge, A. Cahill, and J. Van Genabith, “Questionbank: Creating a corpus of parse-annotated questions,” in *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, 2006, pp. 497–504.

[20] C. R. Kothari, *Research Methodology: Methods and Techniques*. New Age International (P) Ltd., 2004.

[21] M. E. McLaughlin and F. Drasgow, “Lord’s Chi-Square Test of Item Bias With Estimated and With Known Person Parameters,” *Appl. Psychol. Meas.*, vol. 11, no. 2, pp. 161–173, 1987.

[22] V. L. J. Gregory, “Gregory Research Beliefs Scale: Factor structure and psychometric properties.,” *Diss. Abstr. Int. Sect. A Humanit. Soc. Sci.*, vol. 70, no. 5-A, p. 1783, 2009.