Comparison sensitivity of the differential item function (DIF) detection method

Ulfa Khoeruroh¹, Heri Retnawati²

¹Education Research and Evaluation Master’s Program Graduate School, Yogyakarta State University, Yogyakarta 55281, Indonesia
²Mathematics Education Faculty of Mathematics and Natural Science, Yogyakarta State University, Yogyakarta 55281, Indonesia

¹ulfakhoeruroh.2018@student.uny.ac.id

Abstract: Differential Item Functioning (DIF) or often referred to as bias is a term that is not good, has a racial meaning, suppresses or is too fanatical about the object to be measured. Bias in a test is a test condition that is unfair, inconsistent, contaminated by factors outside the aspects to be tested and misuse of the test. The purpose of this study is to find out which method is the most sensitive in detecting the presence of DIF on items. The methodology used is to study literature from books, journals and study the results of research that has been done. There are many methods that can be used to detect DIF. These methods are the Mantel-Haenszel Method, the Chi-Squared Method of the Lord, the Probability Comparison Test Method, Item Characteristic Curve (ICC) and Extent Method According to Raju.

1. Introduction
The term test is no stranger among educators and other people, almost everyone has heard, talked about or even taken a test. The test itself is a measuring tool for recording and observing student achievement in accordance with the target assessment [1]. The test can also be interpreted as a systematic procedure to measure a sample of a person’s behaviour, meaning that the test represents aspects that will be measured as a tool to judge someone [2].

The instrument used in the test must have good validity and reliability. Quality tests are tests that can provide information in accordance with the desired objectives, including being able to determine which students have or have not thoroughly mastered the material [3]. The test must also be fair, meaning that here the items in the test can be done by students who have the same competence. If, the items presented are in favour of a particular student then those items are incorrect, in other words that the test items must be neither beneficial nor detrimental to the student. If a test set is influenced by other factors, other than what is to be measured, it is called bias in a test [4].

The term item bias or known as Differential Item Function (DIF) according to psychometric experts is the difference in item level performance between groups of students or respondents who fit the latent nature [5]. Ibrahim (2016) [6] argues DIF occurs when the responses of students who have the same ability show systematic differences in certain groups. According to Guilara, Gomez-Benito, & Hidalgo (2009) [7], a test is considered biased when students from different groups (e.g. ethnicity, culture or gender) have different probabilities, when this is matched to the attributes measured by the item. So that a test is said to be biased if two test participants with the same ability, from different groups do not get the same opportunity to answer correctly. Thus, the unbiased test items are test items
that are expected to provide the same opportunities in answering correctly to test participants who have the same ability from different groups.

Determination of whether an item is indicated by DIF or not, a DIF index is needed to know an item indicated DIF or not, that is an index that shows how strong the indication DIF is on that item. To detect the presence of DIF on a test, there are several methods that can be used to detect the presence of DIF. Camilli & Shepard (1994) [8] suggested that there are eight methods to detect DIF, but there are only three recommended methods, namely the Haenszel-Mantle method, logistic regression method, and the characteristic curve method based on item response theory. Meanwhile, according to Stark and Chernyhenko [4] there are several methods to detect DIF, both parametric statistics and non-parametric statistics. When using parametric statistics, the methods that can be used are the chi-square of the lord (Lord's chi-square), the likelihood ratio test and the signed and unsigned area methods. As for non-parametric statistics, the methods that can be used are Sibtest and Coat-Haenszel methods.

According to the Item Response Theory, DIF occurs when the Item Characteristic Curve (ICC) of two groups is not identical or does not have the same item parameters after the resizer (Ozlem Yesim Ozbek Bastug, 2016) [9]. Also, DIF might occur when the latent capability space is not specifically specified. When groups have different multidimensional abilities distribution and test items are chosen to differentiate between these abilities, using unidimensional assessment, might mark items as DIF items. Unidimensional namely each test item only measures one ability (Retnawati, 2013) [10]. Unidimensional assumptions can be demonstrated only if the test contains only one dominant component that measures the achievement of a subject.

Research on various methods of detecting the presence of DIF has been done. Research conducted by Kartowagiran (2004) [11] concerning the comparison of the sensitivity of the Likelihood Ratio Test method, grain characteristic curves, Chi-square Lord and Extensive Raju in detecting the presence of DIF. Likewise, the research conducted by Sudaryono (2012) [3] about the differences between the Chi-square Scheuneman method, the Mantel Haenszel method, and the Rassi model grain responsiveness theory method which is the most sensitive in DIF research.

Therefore, the problem in this paper is the most sensitive to detect the presence of DIF, namely the Mantel-Haenszel Method, the Chi-Squared Lord Method, the Expansion Method according to Raju, the Likelihood Ratio Test Method, or Item Characteristics Curve (ICC)? Based on the problem formulation, the purpose of this study is to obtain differences in the sensitivity of the Differential Item Function (DIF) detection method, namely between the Mantel-Haenszel Method, the Chi-Squared Lord Method, the Expansion Method according to Raju, the Likelihood Ratio Test Method, or the Item Characteristic Curve (ICC).

2. Research Method
The research method used in this article is library research or literature review, with a series of collection of library data or research objects which are explored through various literary information (books, journals, reviews of research results that have been conducted, reports / conclusions of the seminar, notes / records of scientific discussions and documentation). Literature research or literature review is research that describes or reviews knowledge, findings or ideas that are relevant to the topic of discussion. The data used in this research is secondary data. Secondary data is data obtained by researchers from existing sources, not from direct observation. Secondary data sources are referred to from original books or scientific reports in the form of articles or journals.

3. Result and Discussion

3.1. Concept of Differential Item Function
The item bias or Differential Item Function is a threat to the validity of the measurement because the score results will be tainted by something that is not planned to be measured. Systematically, item bias can be stated in the form of probability [3]. Meaning, students who have the same ability, but will not have the same opportunity in answering questions correctly. If an item is relatively more difficult for
students who have the same ability background, indicating that the item is biased. Item bias in measurements indicates an error in the measurement. In obtaining information on whether or not there is DIF, it can be done with several methods of detecting DIF. Here are some DIF detection methods.

3.2. Mantel-Haenszel Method

The Mantel-Haenszel (MH) method was originally used to match retrospective subjects to risk factors for developing cancer. The MH procedure was first developed by Mantel and Haenszel in 1959, and was used to detect DIF by Holland and Thayer in 1988 which is still used to analyze the presence of DIF. The MH technique is ideal because it does not rely solely on statistics $\chi^2$, which can be overly sensitive when using large samples, which is common in DIF analysis. Statistics not only complete statistics $\chi^2$, but also the possibility of the assessment of the behavior of the DIF to be made. Analysis of the existence of DIF by the Mantel-Haenszel method using the following equation:

$$
\alpha_{MH} = \frac{\sum_{k=1}^{K} M_{Rbk}M_{Fsk}}{M_k} \frac{\sum_{k=1}^{K} M_{Rbk}M_{Fsk}}{M_k}
$$

If $\alpha_{MH} > 1$ then the item under investigation is influenced by the DIF which benefits the reference group. If $\alpha_{MH} > 1$ then the item under investigation is influenced by the DIF in favor of the focus group. The bias calculation of the Mantel-Haenszel grain with chi-square distribution is done using the formula:

$$
\chi^2_{MH} = \left[\frac{\sum_{k=1}^{K} M_{Rbk} - \sum_{k=1}^{K} \frac{M_{bk}M_{Rk}}{M_k}}{\sum_{k=1}^{K} M_{bk}M_{sk}M_{Rk}M_{fk}}\right]^2
$$

with the Chi-square distribution, the test condition is if $\chi^2_{MH} \geq \chi^2_{table}$, then the item is indicated by DIF (Rustam et al, 2019) [12].

Following are the steps to detect DIF by the Mantel-Haenszel method:

- Divide subgroups into reference groups and focus groups.
- The focus group is the most important in the analysis and compared with the reference group after matching it on $\theta$
- Reference and focus groups are compared at unit intervals $\theta$ to the number of examinees
- The results of this comparison are used to estimate odds ratios so that they can be calculated, and $\chi^2$ significant tests can be performed to assess the presence of DIF.

3.3. Lord’s Chi-Squared Method

The Lord's $\chi^2$ method is to compare the chances of answering questions by both groups on an item by checking the difference in item parameters depending on the IRT model used. Price (2014) [13] states that the lord method, comparing focus group parameters and reference group parameters, looks at whether there are significant differences between the two groups with statistics. This test is for ICC equality between reference groups and focus groups. To detect DIF using the Chi-Square method from Lord, use the following equation:

$$
(\chi^2_i) = v_i \sum_{i}^{-1} v_i
$$

where $v_i$ is a vector of differences in the $i$-th item parameter estimation between the focus group and the reference group, while $\Sigma_{i}^{-1}$ is a variance-covariance matrix for differences in item parameter estimation [4].

Following are the steps to detect DIF by the Chi-Square method from Lord:

- Divide subgroups into reference groups and focus groups.
- Estimates item parameters and variance-covariance from focus groups and reference groups separately.
- Determine the parameter equivalent constant between the focal group and the reference group.
- Calculate the value of \( \chi^2 \) and compare it with a table of criticism values \( \chi^2 \) with degrees of freedom 3 or can use the desired level of significance.

### 3.4. Extensive Method According to Raju

The Extensive Method According to Raju is a method developed by Raju in 1988 then refined in 1990. In this method, one approach that can be used to compare the function characteristics of items is to compare the ICC itself rather than its parameters. If the item parameters have been scaled and ICC is identical then the area between the ICC will be zero. If the area between ICC is not zero, it means that the test item contains DIF. To detect DIF use the area index method according to Raju to calculate the area between two ICCs for 3 parameters, using the following equation:

\[
A = (1 - c) \left[ \frac{2(a_2 - a_1)}{D a_1 a_2} \right] \left[ \ln \left[ 1 + e^{D a_1 a_2(b_2 - b_1)/(a_2 - a_1)} \right] - (b_2 - b_1) \right]
\]

Following are the steps for detecting DIF by the Luasan method according to Raju:
- Divide subgroups into reference groups and focus groups.
- Estimating item parameters for individual reference groups and focus groups.
- Determine parameter equalization constants to position the focus and reference parameters on a general metric (here is also used to transform covariant matrices).
- Compare the results of calculations with \( \chi^2 \) tables.
- If \( \chi^2 \) counts are smaller than \( \chi^2 \) tables, then the item does not contain DIF.

### 3.5. Likelihood Ratio Test

This method is a method for detecting DIF using Item Response Theory, by Camili and Shepard [8]. Likelihood Ratio Test is one of the comparative test methods between two data distributions, aiming to find out which distribution is more appropriate to be applied in a particular case. The strength in this method is influenced by the sample size, ability distribution, the power of distinguishing the item studied and the DIF pattern [14]. The test statistics used in the likelihood ratio test are as follows:

\[
X^2_{(M)} = -2 \ln (LR)
\]

with the LR formula, which is:

\[
LR = \frac{L^*_1}{L^*_2}
\]

Information:
- \( L^*_1 \) = the value of the likelihood function of model 1
- \( L^*_2 \) = the value of the likelihood function of model 2

This test is done by comparing the value of the Likelihood ratio when the model only enters the value of the constant or -2 Log L in step 1 with the model after entering the independent variable or -2 Log L in step 2.

Following are the steps for detecting DIF by the Likelihood Ratio Test method:
- Estimating item parameters with the item response theory of the 3p model, obtaining a match statistic that detects the \( x^2 \) distribution of probability transformations g (c) on a test consisting of k items.
- Determine one test item, for example the k-th item, to be evaluated.
- The test items are made as if they were two items.
• Estimates the parameter, and we get the value of $x^2$ probability transformation $g(a)$ for a test consisting of $k + 1$ item.
• Determine the value of $x^2_{(M)}$, which can be used to indicate the existence of DIF on item.

3.6. Item Characteristic Curve Method (ICC)
Item Characteristics Curves (ICC) describe the relationship of subject performance on an item and the underlying latent device. ICC is a mathematical function that states the relationship between the opportunity for someone to answer the item correctly with the ability measured by the item [15]. This relationship is a nonlinear regression function of the item trait score or ability measured by the test. Whether or not a DIF occurs on an item lies in the item response function for that item in the group in question. The curve that describes the response function is called the Item Characteristic Curve. At ICC obtained a description of the relationship of student ability on an item and the parameters on the item, namely the level of difficulty, different power and guess parameters. ICC is another name for the item response theory (IRT).

The sum of all grain characteristic curves is a test characteristic curve, shown in the following figure:

![Figure 1. Item Characteristic Curve](image)

The parameters of each test item are slope (a), threshold (b) and asymptote (c) of the item characteristic curve. The sum of all item characteristic curves is a test characteristic curve. Each item has different characteristics and shapes depending on the level of difficulty, the difference between the power and guesses of each item. The greater the guessing value of an item, the more it will go up the ICC image. The steepness of the curves of each item also varies depending on the different grain power indexes. A gentle curve shows that the item has a low power difference.

The steps to detect DIF with the ICC technique are:
• Estimates parameters a, b, and c of the N items of a test and obtains a statistical match that approximates the Chi-square distribution, $G(1)$ for example by using the BILOG program.
• Determine the item to detect the DIF.
• Analyzing these items is specific to group F only and to R only.
• Re-estimate the parameters of the items so that the Chi-square price of the Transformation of Probability Ratio, $G(2)$, for $N + 1$ items.
• Calculate the difference in Chi-square values using the formula:
\[
\chi^2 = G(1) - G(2)
\]
and compare the results with the price of Chi-square criticism with a certain $\alpha$ and degree of freedom $= 3$. Make a test decision, where if Chi-square Calculate $>$ Chi-square Table can be concluded that the existence of DIF is proven significantly.
3.7. **Strengths and Weaknesses of DIF Detection Methods**

| No | Methods | Strengths | Weaknesses |
|----|---------|-----------|------------|
| 1  | Mantel-Haenszel | Statistically efficient and easy to calculate | Lack of power to detect non-uniform DIF in the range of scores θ |
| 2  | Chi-Squared Lord | The use of this method can detect DIF consistent (Uniform) and inconsistent (Non-Uniform) | Sensitive to the number of samples used. |
| 3  | Expansion Method according to Raju | Sensitive to non-uniform or presence of DIF. Statistical calculations made point by point. | Difficult calculations that must be done. Item parameters must be carried out in two groups if the number of items and participants is large enough that it takes a long time to count even though using computer aids. |
| 4  | Likelihood Ratio Test | The strength in this method is influenced by the sample size, ability distribution, the power of differentiating the studied item, and the DIF pattern. | The calculations are difficult to do manually, especially for the censored data and the software needed to be compatible |
| 5  | Item Characteristics Curve | Not based on dependent, students' scores are described as not test dependent, emphasizing the level of items not the test, a method that requires a measure of determination for each ability level score | The method works through a difficult process, given that one of the advantages of this method is the accuracy in local measurements |

Research on the sensitivity of the DIF detection model has been widely carried out, many previous researchers have conducted the study. Rustam et al. [12] conducted a study of the comparison of the mantle-haenszel method and the standardization method: detecting differences in function of test items. Sensitivity is seen from the number of DIF items. The results show that the Mantel Haenszel method is more sensitive than the standardization method in the detection of DIF for both 400 and 2000 samples. This is consistent with the results of Kabasakal, Arsan, Gok & Kelecioglu [16] research on Comparing Performance (Type I error and Power) of IRT Likelihood, SIBTEST and Mantel-Haenszel Methods in the Determination of Differential Item Function. The conclusion of his research is that the most powerful method to detect DIF is the Mantel Haenszel method, and the Type I error rate found in this study with the SIBTEST method.

Other studies, namely research conducted by Kartowagiran [11] conducted research on the comparison of sensitivity of various methods in detecting grain bias giving the conclusion that the Likelihood Ratio Test method was the most sensitive in detecting grain bias followed by grain characteristic curves, Chi-square Lord, and Raju area. The results of the study are consistent with the results of Kim and Cohen's [17] research comparing the detection methods of Chi-square Lord, area of Raju, and probability ratio. The results showed that there was a good match in detecting DIF between the three methods. The probability ratio test method is more sensitive in detecting more DIF than the Chi-square Lord method, and the extent of Raju. In line with Rahmawati's [18] research on the Comparison of the KAI-Square Lord method, the area of marked and unmarked regions, and the comparative test of the possibility to detect DIF. The results showed that the comparison test method
might be more sensitive than the other methods. Rahmawati suggested using a comparative test method to detect DIF.

Likewise with Li-An Wu & Rung-Ching Tsai's [19] research on A Comparison of Three Polytomous DIF Detection Methods which provides the results of the performance of the three procedures are the logistic regression procedure (LogR), the likelihood ratio test (LRT), and the differential functioning of items and tests procedure (DFIT) in detecting differential item functioning (DIF) is the most powerful LRT among the three in detecting DIF. DFIT was less powerful than LRT, but also useful for DIF detection, especially with groups of different ability distributions and relatively large percentage of DIF items. LogR, with mean powers lower than 0.4 in all conditions, appears to be sensitive only to items with large DIF sizes. Retnawati [10] also conducted research on detecting the function of distinguishing items with a simple volume index based on the multidimensional grain response theory that in detecting DIF, a simple volume index is used which is the development of a marked area index. Using the likelihood comparison method and the slight difference in opportunity, with the sample size being enlarged the change becomes more significant. Therefore, likelihood comparisons are very sensitive to identifying DIF when the sample is enlarged.

In most cases, the detection rate of the five procedures is influenced by sample size, distribution of abilities, strength of the items being studied. Based on several studies that have been presented, many researchers use the likelihood ratio test method which is more significant in detecting DIF. With a large number of samples will be more significant and more accurate in detecting DIF. Each method has advantages and disadvantages of each, but there will be the most sensitive method in detecting DIF. In the literature review of this article, the likelihood ratio method is the most sensitive among others and many previous researchers used this method to detect DIF. This is supported by several previous studies. The likelihood ratio method is the most sensitive among others that this method is the method of Item Response Theory or Modern Theory. While the other method is the method in the Classical Test Theory. And if you use a lot of research samples it is better to use the likelihood ratio test method, as well as easy patterns in detecting DIF. However, it does not rule out in a study that the other methods are more sensitive than the likelihood ratio test method.

4. Conclusions
A test that is used should be evaluated to see fairness in the test. Injustice in the test will harm one party. If, the items presented are in favor of a particular student then those items are incorrect, in other words that the test items must be neither beneficial nor detrimental to the student. If it harms or benefits one party, the test is biased or detected by DIF. There are several methods in detecting the presence of DIF in a test. The methods discussed in the article are the Mantel-Haenszel Method, the Chi-Squared Lord Method, the Expansion Method according to Raju, the Likelihood Ratio Test, and the Item Characteristic Curve (ICC). The detection rate of DIF is influenced by sample size, ability distribution, and distinguishing power of the items being studied. The most sensitive likelihood ratio test method among others in this literature study. Because, when viewed from the sample size this method is very sensitive if you have a lot of research samples. As well as patterns in detecting DIF that are easily understood. However, it does not rule out in a study that the other methods are more sensitive than the likelihood ratio test method.

References
[1] Jacobs & Chase. 1992. Developing and Using Test Effectively: A Guide for Faculty, United States of America: Jossey-Bass Inc
[2] Yusuf, A Muri. 2015. Asesmen dan Evaluasi Pendidikan. Jakarta: Prenadamedia
[3] Sudaryono. 2012. Sensitivity of Differential Item Functioning (DIF) Detection Method. Jurnal Evaluasi Pendidikan. Vol.3, No.1, 82-94
[4] Heri Retnawati. 2014. Teori Respons Butir dan Penerapannya untuk peneliti, praktisi pengukuran dan pengujian, mahasiswa pascasarjana. Yogyakarta: Nuha Medika
[5] Abbas, A.R., & Enayatollah, S. 2010. Gender differential item functioning analysis of the
university of Tehran English proficiency test. Pazhuhesh-e Zabanha-ye Khareji, 6, 89-108

[6] Ibrahim, A. 2016. Comparative Analysis of Generalized Mantel Haenszel (GMH), Simultaneous Item Bias Test (SIBTEST), and Logistic Discriminant Function Analysis (LDFA) methods in detecting Differential Item Functioning (DIF) in Dichotomous and Ordinal tests. Internasional Journal of Education Benchmark (IJEB), 6(2), 24-43

[7] Guilera, G., Gomez, J., & Hildago, M.D. 2009. Scientific Production on the Mantel-Haenszel procedurea as a way of detecting DIF. Psicothema, 21 (3), 429-498

[8] Camilt, G & Shepard, L.A. 1994. Methods for identifying biased test items. Thousand Oaks, CA: Sage Publication

[9] Ozlem Yesim Ozbek Bastug. 2016. A Comparison of four differential item functioning procedur in the presence of multidimensionality. Academic Journals; Educational Research an Reviews. Vo. 11(13), p. 1251-1261.

[10] Heri Retnawati. 2013. Pendeteksian keberfungsian butir pembeda dengan indeks volume sederhana berdasarkan teori respons butir multidimensi. Jurnal penelitian dan evaluasi pendidikan. V. 17, No.2

[11] Kartowagiran, B. 2004. Perbandingan Berbagai Metode untuk Mendeteksi Bias Butir. Disertasi, Yogyakarta: Universitas Gadjah Mada

[12] Rustam, A., Naga, D.S, & Supriyati, Y. 2019. A Comparison of Mantel-Haenszel and Standardization Methods: Detecting Differential Item Functioning. Jurnal Matematika dan Pembelajaran, Volume 7, No 1, 16-31

[13] Price, Emily. 2014. Item Discrimination, Model Data FIT, Type I Error Rates in DIF Detection Using Lord’s $\chi^2$ the Likelihood Ratio Test, and the Mantel Haenszel Procedure. Dissertation. The Patton College Of Education of Ohio University.

[14] Burcu Atar & Akihito Kamata. 2011. Comparison of IRT Likelihood Ratio Test and Logistic Regression DIF Detection Procedures. Hacettepe Universitesi Egitim Fakultesi Dergisi (H. U. Journal of Education) 41: 36-47

[15] Djemari Mardapi. 1991. Konsep Dasar Teori Respon Butir. Perkembangan dalam bidang Pengukuran Pendidikan. Yogyakarta: Jurnal Cakrawala Pendidikan, No.3 th X.

[16] Kabasakal, Arsan, Gok & Kelecioglu. 2014. Comparing performance (Type I eror and power) of IRT likelihood ratio SIBTEST and Mantel-Haenszel in the determination of differential item functioning. ERIC; Educational Sciences: Theory and Practice, V.14, No. 6 P. 2186-2193

[17] Kim, Seock-Ho dan Allan S. Cohen. 1995. A Comparison of Lord’s Chi-square, Raju’s Ares Measures, and the Likelihood Ratio Test on Detection of Differential Item Functioning. Journal of Applied Measurement in Education. Vol 8(2).

[18] Etti Rahmawati. 2006. Perbandingan metode KAI-Kuadrat Lord, Luas daerah bertanda dan tidak bertanda Raju, dan uji perbandingan kemungkinan untuk mendeteksi Differential Item Functioning. Tesis, Yogyakarta: Universitas Gadjah Mada

[19] Li-An Wu & Rung-Ching Tsai's. 2010. A Comparison of Three Polytomous DIF Detection Methods. ResearchGate.

Acknowledgment
Thanks to the Graduate of Yogyakarta State University that has facilitated the funding of this article. Thanks to the reviewers for the suggestions and feedback.