International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-8 Issue-5C, May 2019

Sensitivity Of Differential Item Functioning Detection Methods On National Mathematics Examination In North Sumatera Province, Indonesia

Kerdid Simbolon, Yetti Supriyati, Dali S. Naga

ABSTRACT--- The purpose of this study was to examine the differences in sensitivity of three methods: IRT-Likelihood Ratio (IRT-LR), Mantel-Haenszel (MH) and Logistics Regression (LR), in detecting gender differential item functioning (DIF) on National Mathematics Examination (Ujian Nasional: UN) for 2014/2015 academic year in North Sumatera Province of Indonesia. DIF item shows the unfairness. It advantages the test takers of certain groups and disadvantages other group test takers, in the case they have the same ability. The presence of DIF was reviewed in grouping by gender: men as reference groups (R) and women as focus groups (F). This study used the experimental method, 3x1 design, with one factor (i.e. method) with three treatments, in the form of 3 different DIF detection methods. There are 5 types of UN Mathematics Year 2015 packages (codes: 1107, 2207, 3307, 4407 and 5507). The 2207 package code was taken as the sample data, consisting of 5000 participants (3067 women, 1933 men; for 40 UN items). Item selection was carried out based on the classical test theory (CTT) on 40 UN items, producing 32 items that fulfilled, and item response theory selection (IRT) produced 18 items that fulfilled. With program R 3.333 and IRTLRDIF 2.0, it was found 5 items were detected as DIF by the IRT-Likelihood Ratio-method (IRT-LR), 4 items were detected as DIF by the Logistic Regression method (LR), and 3 items were detected as DIF by the Mantel-Haenszel method (MH). To test the sensitivity of the three methods, it is not enough with just one time DIF detection, but formed six groups of data analysis: (4400,40),(4400,32), (4400,18), (3000,40), (3000,32), (3000,18), and generate 40 random data sets (without repetitions) in each group, and conduct detecting DIF on the items in each data set. Although the data lacks model fit, the 3 parameter logistic model (3PL) is chosen as the most suitable model. With the Tukey's HSD post hoc test, the IRT-LR method is known to be more sensitive than the MH and LR methods in the group (4400,40) and (3000,40). The IRT-LR method is not longer more sensitive than LR in the group (4400,32) and (3000,32), but still more sensitive than MH. In the groups (4400,18) and (3000,18) the IRT-LR method is more sensitive than LR, but not significantly more sensitive than MH. The LR method is consistently tested to be more sensitive than the MH method in the entire analysis groups.

Keywords: Differential item functioning (DIF), IRT Likelihood Ratio, Mantel-Haenszel, Logistic Regression.

I. INTRODUCTION

Written tests are currently widely used as educational evaluation tools in Indonesia, as well as abroad. Written tests, such as the *Ujian Nasional* (UN) or National Examination received much public attention because this written test caused broad social impacts, such as its use to determine the examinee's success or failure and some other social consequences that appeared around the test. Such a test, by experts, is called "high stake exams" or "high stake test", which is an exam (test) whose measurement results are a basis for decission-making that can change lives, or tests are very risky [1]-[3]. Through such high stake tests evaluation, important decisions are made, e.g. student graduation and rewards and punishments giving to teachers or schools. The evaluation pattern of high stake exams has an element of strength as well as weakness. Regardless of the pro and contra of the high test exams, this test model is used in Indonesia, so inevitably the thought of written tests that are free of various errors, including being free of the pressure arising around test takers, is a must to think by evaluation and measurement experts.

Realizing importance and urgency of the results of the national examination (UN), the compiled National Examination instruments must really be able to measure what should be measured. It must be able to provide reliable measurement results and reflect the true abilities of students. Feasibility of decisions taken based on UN scores is largely determined by the quality of the UN instrument. Therefore it is understandable why measurement experts demand the fulfillment of the requirements for validity, reliability, objectivity, and fairness in the test instrument as a measuring instrument. The validity requirements of the test confirm the accuracy of the measure of what the test wants to measure. Reliability confirms the constant results of the several times instrument being tested. And the terms of objectivity confirm the absence of subjective factors that influence [4]. Objective assessment, based on clear procedures and criteria, is not influenced by the subjectivity of the assessor [5]. Likewise, a fair assessment will ensure that no participant from a particular group gains certain anvantage from the test, while the other groups are disadvantaged.

Manuscript published on 30 May 2019.
***** Correspondence Author (s)
Kerdid Simbolon, Educational Research and Evaluation Program, Jakarta State University, & Mathematics Education Study Program, Teacher Training and Education Faculty, Indonesian Christian University, Jakarta, Indonesia
Yetty Supriyati, Educational Research and Evaluation Program, Jakarta State University, Jakarta, Indonesia
Dali S. Naga, Educational Research and Evaluation Program, Jakarta State University, Jakarta, Indonesia

Unfairness on a test instrument arises when one group of test participants gets certain advantages or disadvantages over other groups. The earliest term used for this purpose is "item bias", used in a social context and has a negative meaning. In terms of item bias, it includes both social meaning and statistical meaning. In determining whether a test is biased or not, it is often done without considering statistical meaning. These results give an ambiguity about the criteria for evaluating item bias. The research methodology then developed methods for analyzing the index of items bias with a focus on differences between groups of participants who answered the same test questions and then determined matching criteria as a basis for comparison. Furthermore, statistical information can be used as a tool to determine item bias, and the terminology changes to "*Differential Item Functioning* (DIF)". DIF is a more neutral term and more suitable for use [6]. Differential Item Functioning (DIF) compared test results between at least two subgroups of participants with the same level of ability.

In fact, the instrument test as expected above is not always available, so it is often for the tests used in education evaluation have low validity and reliability, and contain elements of unfairness. Items that are unfair, often encountered, and items like this are called bias items. This item tends to disadvantages one group of test participants and advantages the other group. The research results of Budiyono (2005) show that the high school UN Mathematics in academic year 2003/2004 in Surakarta contains 4 items DIF [7]. Badrun Kartowagiran found that there were 9 items bias on the junior high school mathematics questions used in the 2003 UN detected using IRT-Likelihood Ratio [8]. The items infected DIF were also found in the 2003 National Mathematics Examination by Heri Retnawati [9], proved to contain 2 dimensions and showed that from 28 items analyzed, 2 uniform DIF and 26 non-uniform DIF items. Research by Retnawati and Hidayati on the National Junior High School Mathematics Examination in the Province of DIY in the academic year 2004/2005 contains uniform bias test (*Differential test Functioning*: DTF) that benefits women students. There are 6 items detected contain DIF based on gender differences.[10] Triyatno, Sriyono, and Ngazizah found DIF items on the even semester physics exam in Porworejo District Public High School 2012/2013 academic year. Found 14 items that were DIF gender [11]. In addition to domestically, the bias of the test items has actually been realized and has been investigated earlier in foreign countries.

Along with the emergence of item bias problems, on the other hand, it also urges the need for a good DIF detection method, a sensitive method of detecting the presence of bias items, and explaining the causes of bias. Lately, there have been many DIF detection methods developed, both those who based on classical test theory (CTT) and those who based on modern test theory or item response theory (IRT). Various DIF detection methods that are more suitable for detecting the presence of DIF are needed in an effort to make free test instruments from cases of bias or unfair. If such methods do not exist, then the problem of item bias will remain on the various written tests that are used, or this problem will be more slowly resolved. Such a situation will reduce confidence and lead to a presumption of guilt, can even expand into a social problem, because the element of unfair contained in the test. Thus the assessment of the goodness of the current DIF detection methods, as well as the development of better new methods, is very important and urgent to do. The fundamental objective in this study was to determine the differences in sensitivity of the three DIF detection methods, namely Mantel Haenszel (MH), Logistic Regression (RL) and IRT-Likelihood Ratio (IRT-LR).

II. LITERATURE REVIEW

A. Impact Item, DIF and Item Bias

In the assessment of fair items, three important terms appear that have different meanings. *First*, the impact of the item, *second*, differential item functioning (DIF) and *the third*, item bias. Item impact is evidence that occurs when test participants from different groups have different opportunities to correctly answer an item, which is due to differences in the actual abilities of the two groups that the item wants to measure. DIF appears when test participants from different groups show different opportunities to answer the item correctly, but their abilities have been matched, or they have the same ability. Meanwhile, item bias arises when participants from one group are less likely to answer the items correctly than the other group because a number of characteristics of test items or situations are not relevant to the test objectives. DIF becomes a necessary condition for the occurrence of item bias, but not enough conditions. Item impact and item bias, both differ in group situations based on relevant characteristics or irrelevant characteristics of the test. However, if DIF appears, then this occurrence is not enough to prove the occurrence of item bias; but furthermore, the item must be analyzed (for example by content analysis, field evaluation) to assess the presence of item bias in it [12].

DIF appears when an item is substantially more difficult for one group than another group, after all the differences in subject matter tested have been taken into account. Thus, DIF refers to the ways in which item function differently for individuals or groups of test takers who are equally capable. The DIF analysis is based on the principle of comparing focus group performance (e.g. women) to an item with a reference group (e.g. men), by controlling the knowledge being tested. DIF does not only means that an item is more difficult for one group than for another group but also if participants in one group tend to know more test subjects than other groups, they will perform better on all test items. Therefore, once a DIF is identified on an item, it can be related to the appearance of item bias or item impact [13]. Item bias, a challenge to the validity of the test, leads to systematic errors that can give a misinterpretation of conclusions made for certain group members. In other words, when an item unfairly benefits one group over another, then an item bias appears. The item is biased

because the item itself contains certain sources of difficulties other than the construct being tested, and this difficulty factor is detrimental to the performance of the test takers [14]. However, differences in performance or ability to answer items are not automatically evidence of item bias. Differences in group performance can represent differences in pure experience and knowledge with respect to the purpose of the test [15]. This result is referred to as item impact. As biased, the impact is constant on certain group members, but this effect describes the performance differences that the test wants to measure [16].

Abedalaziz, Ismail, and Hussin noted that test items with bias content could be: (1) contain content that is familiarly different to participants in the matching group; (2) contains sources of difficulty that are not relevant to the test construct that affect performance; (3) loading material that may be offensive, demeaning, or emotionally debilitating participants' motivation and attention to the test, thereby reducing the performance; and (4) asking for information which participants did not have the same opportunity to study it [17]. Test items with gender bias can include: (1) tasks that perpetuate the type of unwanted role, type of race or gender (gender); (2) material or references that can attack members of one sex; and (3) references to objects and ideas that may be more familiar to men or women. A biased test can be caused by the presence of various irrelevant factors, which are not the target construct of the test, which are related to gender, ethnicity, race, linguistic background, economic status or inhibiting conditions, differences in the environment, culture, and dayly life experience of participants.

In general, DIF analysis is considered as the first step, the statistical step, to decide whether the item is biased towards a particular group. The emergence of the DIF must first be seen as an impact, namely the real difference in the ability of the two groups. This is important because if items are detected as DIF, it does not always mean that the item is biased. In this case, it is important to consider whether the reason for the difference in group scores on the item is relevant or not, which depends on the object or purpose of measurement. The first case, DIF is caused by actual differences, and the second case is caused by bias [18].

In the DIF analysis, the population is divided into two subgroups called reference groups and focus groups. Reference groups are made as a basis (referring to the majority or the beneficiaries) and focus groups as the center of attention for fairness (referring to minority or disadvantaged parties). Two types of DIF can be identified as uniform DIF and nonuniform DIF. Uniform DIF arises when one group's advantage over another group is evenly distributed, advantages only one group consistently along the ability scale. Nonuniform DIF occurs when conditional dependence on group membership and item performance changes in size. Advantage or disadvantage does not occur along the continuum of the ability scale.

In notation, the understanding of DIF on Penfield and Camilli, cited from Kondratek and Grudniewska, is stated as follows.[19] If U_i states the answer or response to item i, θ is the level of participant's ability and G group membership variable, then the general equation that defines DIF by considering group membership, is written: $(U_i|\theta, G) \neq (U_i|\theta)$. In the case of dichotomous scores are written: $P(U_i = 1|\theta, G) \neq P(U_i|\theta)$, which means the opportunity to correctly answer an item depends not only on ability θ, but also on group G membership. G with two values (F, R), then DIF in item i can be denoted also as: $P(U_i = 1|\theta, G = F) \neq P(U_i = 1|\theta, G = R)$. Which means, the probability to answer the item correctly by participant with the ability θ in group F is different from the probability to answer correctly in group R. Conversely, an item is said not to contain DIF if the probability of answering the item correctly is the same for test takers with the same ability, regardless of their membership group.

B. Gender Differences in Mathematical Abilities

In recent decades, research has repeatedly reported gender differences in mathematical abilities in a number of standard mathematical tests such as SAT-M (Scholastic Assessment Test-Mathematics). Scores on standardized tests have been considered an important measure of the ability to work on mathematical problems. But the results of the study were inconsistent: some found that men outperformed women's abilities in mathematical tasks; a number of others show different sizes of gender differences that are appropriate for the type of mathematical assignment. T. B. Caplan and P. J. Caplan argue that the relationship between gender and mathematical abilities is very weak [20]. Battista conducted a study of 145 high school geometry students from middle-class society [21]. This study investigates the role of spatial visualization and verbal-logical thinking played in gender differences in geometric problem-solving in high school. The findings resulted in different men and women in space abilities and verbal abilities. Gallagher, De lisi, Holst, McGillicuddy-De Lisi, Morely, and Cahalan state that men are more flexible than women in implementing solution strategies [22].

Research has largely focused on the relationship between three cognitive abilities (verbal, quantitative, and visual-spatial abilities) and gender differences in mathematical abilities. However, the evidence from this study is inconsistent and sometimes contradictory. Spatial ability generally refers to skills in representing, transforming, producing and remembering symbols, nonlinguistic information. Spatial skills involve the ability to think and reason using mental images rather than words. This is believed to be one of the important components of mathematical thinking in mathematical problem-solving [21].

Studies that report gender differences in mathematical abilities that benefit men, generally have consistent conclusions. Linn and Hyde concluded that women were superior in calculations, at all ages and differences that favored men were in problem solving, in middle school. Benbow and Stanley show that gender differences in mathematical reasoning abilities that benefit boys have been observed before girls and boys began to differ in the mathematics courses taken. This gender difference even increased during the high school years. Benbow and Stanley

also suggested that men dominated in the mathematical reasoning ability before they entered adolescence [23].

C. DIF Detection Method

Various types of DIF detection methods have been developed. Where these methods basically assume that if the test takers predictably have the same knowledge or abilities (for example with a total score), then they should show the same performance (though not identical) on the items in the entire group [24]. Among the popular methods of detecting DIF for dichotomous scores, are Mantel-Haenszel [25], Logistic Regression [26], Standardization [27][28], SIBTEST [29], and Item Response Theory [30]. In particular, a number of IRT-based DIF methods are based on a comparison of item parameter estimation values or a comparison of the goodness of fit between item response models and data [31]; A number of them developed a statistical significance test or measured the difference between the curves found from the two groups analyzed [32], such as estimating the area between the curve or the size of the difference in squares, or weighting the area and measuring the difference in measurement. The IRT method has been seen in many studies to be superior to other methods. However, the IRT-based approach does not seem to fit the small sample size. A large number of individuals are needed to match the curve of IRT. This is estimated by Embretson and Riese that between 250 and 500 individuals are needed for stable IRT item parameter estimates [33]. While the bias assessment methods based CTT is fundamentally limited, especially in the presence appraisal approach the group average difference in total score in all groups of participants. These methods cannot distinguish between situations (a) subgroups have different averages, and tests biased, versus (b) different averages, but tests are not biased.

Hidalgo and Lopez-Pina with simulation data studying DIF and size of influence by comparing Logistic Regression, Mantel-Haenszel, and the Mantel-Haenszel revision method [34]. The results show Logistic Regression assessing non-uniform DIF better than MH and MH revisions; although logistic regression analysis is sensitive, it is still inadequate to detect DIF sizes under special conditions. Moses, Miao and Dorans carried out a comparative study of conditional DIF estimates on scientific and historical learning outcomes tests, using simulation data and real data [35]. They used four DIF detection methods: Mantel-Haenszel, Logistic Regression, Log-linear model, and Kernel Smoothing. The results show logistic regression as the best bias method and variance estimation method. Vaughn and Wang conducted a study using a classification tree to investigate DIF by comparing type I errors and power tests, with the MH and RL methods [36]. The study used simulation data of 40 items in dichotomous scores. The results show that the classification tree is an alternative to DIF detection of traditional methods. Type I error and power, equivalent to the MH and RL methods.

Based on the research related to the DIF study for a number of dichotomy items, it can be concluded that (1). the study of the effects of DIF can be classified according to various factors, including gender, race, test difficulties, distribution of test participant abilities, sample size, length of test and many tests infected DIF; (2). commonly used methods comparing the efficacy of different methods of DIF (using CTT and IRT analysis) are: MH method and Logistic Regression, used to compare power tests and level type I errors as shown by the research of Hidalgo and Lopez-Pina and Moses, Miao, and Dorans and Vaughn and Wang [34]. Thus, researchers who want to choose a biased detection method are faced with many methods and there are no clear guidelines for choosing between them. Comparison of bias detection methods is an important practical problem, to help someone choosing the appropriate DIF method for the nature of certain data.

1. IRT-Likelihood Ratio

IRT-Likelihood Ratio-method (IRT-LR) is a type of model-based parametric method for detecting DIF [37]. Uniform DIF type and nonuniform DIF can be tested by this method [38]. IRT-LR is the best measurement method for statistical significance, but is not an index of good effect sizes.[39] The problem with using an IRT-LR is when the sample size is small, especially in any small focus group. The IRT-LR method itself is designed for large amounts of data. In detecting DIF, IRT-Likelihood Ratio (IRT-LR) tested the null hypothesis: "the parameters between the reference and the focus group are not different". The IRT-LR method has a striking advantage because it directly tests the hypothesis of the item response model parameter. Can detect DIF that arises from differences in the level of difficulty, differences in item constructs (discriminate power), or from differences in the level of guessing [40]. An item is detected as DIF, if the likelihood ratio (LR) differs between *compact models* (with few parameters) and augmented models (with all variables involved). In the *compact model*, the parameters of all the items in the focus and reference groups are assumed to be the same. In other words, not one item is assumed to be affected DIF (null hypothesis). While in the *augmented model*, it is assumed that only the item-i (item investigated) is assumed to be different in the whole group, and for other items, the parameters are assumed to be the same. So, only the parameter item i (which is being analyzed) is estimated separately in the reference and focus group (item i is not restricted). For example, in the *augmented model*, item 1 is being analyzed. The parameter estimation of item 1 is not limited to the reference and focus group. Remaining items forms an anchor set for an *augmented model*, each of which is limited, so that the parameter estimates are the same in the two groups. The comparison of the two models is done to see whether additional parameters in set A (*augmented model*) are important or significant. The purpose of this model comparison approach is to test whether additional parameters to augmented models differ significantly in improving the goodness of the model (model's goodness of fit). This means that the addition of parameters in the *augmented model* improves the goodness of fit of the model

The comparison of the likelihood ratio for the two models is expressed by the equation: LR $= L_C / L_A$, where $\boldsymbol{L_C}$ is the

likelihood value of the *compact model*, and L_A is the likelihood value of the *augmented model*. This LR statistic by Thissen and friends is symbolized as G^2, so LR = G^2. Then it is transformed with natural logarithms and multiplied by -2 obtained:

$G^2 = -2\ln(L_C/L_A)$ or $G^2 = [-2lnL_C] - [-2lnL_A]$

In a large number of samples, G^2 has a *chi-square* distribution with degrees of freedom equal to the difference in the number of parameters estimated in the *compact* and *augmented models*. If the G^2 statistic exceeds the critical value χ^2 (table value), or exceeds 3.84 (χ^2 (db = 1; $\propto$ = 0.05)) at the confidence level α, then the null hypothesis is rejected. Means, DIF is present in the item under investigation.

2. Mantel-Haenszel

The Mantel-Haenszel (MH) procedure is a common method in detecting DIF. Holland and Thayer adapted MH procedures to detect DIF [41]. The MH procedure was developed to detect uniform DIF. The MH method works first by dividing the test group into reference groups (e.g. men) and focus groups (e.g. women). The performance of the reference group participants (R) and focus group (F) was compared, in units of ability intervals (total scores), which were weighted by the number of participants at each level of ability. The MH method compares the probability of a correct answer to the reference group and focus group for participants with similar abilities [42]. In comparing the chances of a correct answer, the answers to the reference group items and the focus groups are arranged in a series of 2x2 contingency tables. One table is constructed for each test item at each level of the total score. The entire 2x2xK contingency table is constructed to test the independence of ability variables and group membership [36]. K is the number of total scores (matching groups). The contingency table for item i for level j total score is shown in Table I.

Table I. 2x2 Contingency table item i for level j of ability

Groups	Item score		Total
	1	0	
Reference (R_j)	A_j	B_j	N_{Rj}
Focus (F_j)	C_j	D_j	N_{Fj}
Total	M_{1j}	M_{0j}	T_j

Where, A_j stated total number of reference group participants who answered the item correctly; B_j: total number of reference group participants who answered the item incorrectly; C_j: total number of focus group participants who answered the item correctly; D_j: total number of focus group participants who answered the item incorrectly; M_{1j}: total number test participants who answered the item correctly; M_{0j}: total number of test participants who answered the item wrong; N_{Rj}: total number of reference group participants; N_{Fj}: total number of focus group participants; T_j: *the* total number of participants in j level total score; j index which refers to the jth ability group; j = 1, ..., K.

To estimate the probability of the focus group and reference groups members who answer the items correctly, the MH method produces two statistics. First, *chi-square* statistics to estimate the significance of statistical differences, and second, odds ratios (α) to estimate the size or size of the difference [43]. The odds ratio of the correct answer is the probability for the correct answer to be divided by the probability for the wrong answer. So that the odds ratio of the test participants in the R and F groups is given:

$$\propto_{MH} = \sum_{j=1}^{K}[(A_jD_j)/T_j] \Big/ \sum_{j=1}^{K}[(B_jC_j)/T_j]$$

The null hypothesis and the alternative hypothesis tested in MH are expressed as $H_0: \alpha_{MH} = 1$ dan $H_1: \alpha_{MH} = \alpha \neq 1$, where j $\epsilon \{1, \dots, K\}$ [41]. The null hypothesis states that the odds of correctly answer the items at the j-level score (total score) are the same in the reference and focus group, at all K score levels as matching criteria. In a symbolic way, the null hypothesis for testing a DIF item is defined as:

H_0: $P(U_i = 1|m, f) = P(U_i = 1|m, r)$,m $\epsilon\{1, \dots, K\}$. For each item in the j-ability level, DIF detection uses the Mantel-Haenszel statistical test:

$\chi^2_{MH} = (|\sum_{j=1}^{K} A_j - \sum_{j=1}^{K} E(A_j)| - 1/2)^2 / \sum_{j=1}^{K} var(A_j)$. Where $E(A_j) = (N_{Rj}M_{1j})/T_j$ is the expectation value; and $var(A_j) = (N_{Rj}N_{Fj}M_{1j}M_{0j})/T_j^2(T_j - 1)$ is the value of variance. According to H_0, the asymptotically χ^2_{MH} statistic is distributed as χ^2 (*Chi-square*) with the degree of freedom 1. If $MH_{\chi^2_{obs}} > \chi^2_{\alpha,1}$ then the null hypothesis is rejected, meaning that the item is significantly affected by DIF. Means there is a relationship between ability and group membership.

In determining the large size of DIF, the odds ratio is transformed into delta metrics: $MH_{\Delta DIF} = -2.35\ln(\alpha_{MH})$. The degree of DIF on an item is based on the size of the delta Δ_{MH} stated as follows: item with $|MH_{\Delta DIF}|<1$ is called negligible DIF (type A DIF); item with $1 \leq |MH_{\Delta DIF}| < 1.5$ is called moderate DIF (type B DIF); and items with $|MH_{\Delta DIF}| \geq 1.5$ is called large DIF (type C DIF).[44] The size of $MH_{\Delta DIF}$ can be from -∞ to + ∞. A negative value of $MH_{\Delta DIF}$ indicates the DIF item is detrimental to the focus group, while a positive value indicates the DIF item is detrimental to the reference group. The zero value of $MH_{\Delta DIF}$ means that the item does not indicate DIF.

3. Logistic Regression

The logistic regression method (LR) first proposed by Swaminathan and Rogers became the DIF detection method. Basically, the LR method is used when the dependent variable is binary or dichotomy [45]. LR method has become a widely used method of detecting DIF over the past two decades. The ability of LR to detect uniform DIF and nonuniform DIF, both in the dichotomy and polytomy score item, makes LR a powerful method of detecting DIF.

In the LR method, the probability for someone to answer the item correctly for the ability score (θ) and group membership (G), follows the logistical function: $(Y = 1|\theta, G) = e^Z/(1 + e^Z)$. Where Z = $Ln(P/1-P)$ is the natural logarithm of logit (log odds) the probability to answer the item correctly. Because what want to look for is

differences between groups (DIF uniform) and group membership interactions and participant abilities (nonuniform DIF), then the Z function is expressed in the form of: $z = \tau_0 + \tau_1\theta + \tau_2 G + \tau_3(\theta G)$. So the logistic function becomes: $P/(1-P) = e^{\tau_0+\tau_1\theta+\tau_2 G+\tau_3(\theta G)}$. Taking natural logarithms (Ln) on both sides, is obtained: $z = Ln(P/(1-P) = \tau_0 + \tau_1\theta + \tau_2 G + \tau_3(\theta G)$. This last function is called a logistic function that looks linear, so that it can be solved by means of multiple linear regression in general. Group membership (G) is generally defined as a focus group (F) coded 1, and reference group (R) is coded 2. While θ represents the participant's ability score, which is generally a total score. θG states the interaction between participants' abilities and group membership. Parameters $\tau_0, \tau_1, \tau_2,$ and τ_3 are logistic regression coefficients, respectively representing intercepts, ability coefficients (total score), coefficient of difference in group ability, and coefficient of interaction ability with group membership.

The LR method for detecting DIF is done by entering a new variable at each step to see if the new model provides improvements to the previous model with the presence of the new variable.[46] First, enter the total score variable (θ) into the model, namely: $z = \tau_0 + \tau_1\theta$ (baseline, model 1). Second, enter the group variable (G), namely: $z = \tau_0 + \tau_1\theta + \tau_2 G$ (model 2). Third, enter the interaction variable ability with group membership (θG), namely: $z = \tau_0 + \tau_1\theta + \tau_2 G + \tau_3(\theta G)$ (full model, model 3) [47].

The LR test is used to compare the likelihood of two models. Models with a smaller value of -2LogL have better fit values for data.[48] The LR statistical value is calculated by the formula: G^2 = (–2 log likelihood null model) – (–2 log likelihood estimated model). The G^2 statistic follows the *Chi-square* distribution with degrees of freedom (k), equal to the difference in the parameters of the two models compared. If the value of $G^2 > \chi^2(\alpha, k)$ or if the p-value statistic is smaller than α (5%) then the null hypothesis (H_0) is rejected. The G^2 value for the uniform DIF test is calculated by taking the difference from -2 log likelihood model 1 and -2 log likelihood model 2, with degrees of freedom 1. The value of G^2 for the nonuniform DIF test is calculated by taking the difference from the -2 log likelihood model 2 and the value of -2 log likelihood model 3, with degrees of freedom 1 [49].

In general, a significance test of the *Chi-square* likelihood ratio was carried out, estimating the coefficients τ_2 and τ_3 to see whether DIF was present or not. The null hypothesis is: $\tau_2 = \tau_3 = 0$. The items show uniform DIF if the coefficients τ_2 are significant ($\tau_2 \neq 0$) and $\tau_3 = 0$, with degrees of freedom 1. Item show nonuniform DIF, if the coefficient τ_3 is significant ($\tau_3 \neq 0$) with freedom degrees 1, and without considering τ_2. If τ_3 and τ_2 are not significant ($\tau_2 = 0, \tau_3 = 0$), this indicates that the item in question does not indicate DIF.[50] Some experts suggest testing the presence of uniform and nonuniform DIF simultaneously, with the null hypothesis test: $\tau_2 = \tau_3 = 0$ [47]. The difference from -2 log likelihood (-2LL) from the models was tested by *Chi-square* distribution with degrees of freedom 2 (df = 2). If this step provides significant results, the presence of uniform DIF alone is tested for significance τ_2 with a *Chi-square* distribution with degrees of freedom 1 (df = 1). Simulation tests have shown that this method leads to improved performance.

With 1 as the focus group code and 2 reference group code, when there is a uniform DIF (τ_2 significant) and the value of $\tau_2 > 0$, then the DIF item advantages the reference group. If the value is $\tau_2 < 0$, then the DIF item advantages the focus group. In the case of nonuniform DIF (τ_3 significant) and the value of $\tau_3 > 0$, then items affected by DIF benefit high-ability group reference participants and low-ability focus group participants. Conversely, if the value of $\tau_3 < 0$, then the DIF item advantages low-ability reference group participants and high-ability focus group participants [51].

III. RESEARCH METHODS

A. *Population and Sampling Techniques*

The population of this study was the whole package of participant answers to the 2015 UN Mathematics in North Sumatera Province. Consisting of 25000 participant responses units, in five types of UN packages codes: 1107, 2207, 3307, 4407 and 5507. The 2207code package was taken randomly as a study sample data, which consisted of 5000 units of participant responses to 40 items in the UN. Of the 5000 UN participants, consist of 3067 women (61.34%), and 1933 men (38.66%). The total numbers of National Examination items consist of 40.

B. *DIF Detection Method*

The are three DIF detection methods used in this study: (1). IRT-Likelihood ratio (IRT-LR), (2). Mantel-Haenszel (MH), and (3). Logistic Regression (LR). Data analysis uses two statistical programs, namely the R program version 3.333 and the IRTLRDIF program version 2.0 [52]. Besides R, the IRTLRDIF program version 2.0 was also used to carry out the functions of the IRT-LR method, since the R program was less effective in carrying out the functions of the IRT-LR method for large amounts of data. Since the DIF method used consists of nonparametric methods (MH, LR) and parametric method (IRT-LR), the first two methods can be applied directly without certain conditions. The IRT-LR method itself requires the fulfillment of strict item response theory (IRT). But because the condition of the UN data does not meet the IRT requirements, such as data that is not fit the model, and many participants' response pattern are not fair (Lz <0), the use of the IRT-LR method is continued to see results in general without absolute interpretation.

C. *Research Data*

The research data amounted to 5000 responses to the 2015 National Examination participants. After being cleared of defective data, 4862 participants' response were obtained, for 40 UN items, as final data. With the selection of classical test theory (CTT), 32 items fulfilled the requirements, and from item response theory selection (IRT) was obtained 18 items were eligible. To detect DIF items in the UN items, the three methods: IRT-LR, MH, and LR, were applied to all data in three

analysis groups: (4862.40), (4862.32), and (4862.18). To test the differences in sensitivity of the three methods, 6 analysis groups were formed: (4400,40), (4400,32), (4400,18), (3000,40), (3000,32), and (3000,18). The number 4400 at (4400,40) states the number of UN participant response, and number 40 states the number of UN items. In each analysis group, 40 different data sets were formed to detect the presence of DIF items. The analysis group (4400,40) was formed by 40 times random data iterations, every time 4400 data from 4862 responses to 40 UN items (all items in the UN). Groups (4400,32), formed by taking random data 40 times, each time 4400 data from 4862 responses to 32 items UN (items selected by CTT). And so on, the analysis group (3000,18) was formed by taking random data 40 times, every time 3000 data from 4862 responses to 18 items UN (items selected by IRT). In each group of analysis, 40 times DIF detection was conducted on 40 data sets using three DIF methods (IRT-LR, MH, and LR).

D. Research Design

Research with 3x1 design, using the experimental method, with one factor (i.e. method), 3 treatments, in the form of 3 different DIF detection methods, on the score of the answers to the UN items that were responded to by the participants.

Table II. Average items detected DIF by three DIF detection methods in 3x1 design

Iteration	Average number of items detected DIF		
	MH Method	*LR Method*	*IRT-LR Method*
1	μ_{MH}	μ_{LR}	μ_{IRT-LR}
2	μ_{MH}	μ_{LR}	μ_{IRT-LR}
⋮	⋮	⋮	⋮
40	μ_{MH}	μ_{LR}	μ_{IRT-LR}

The sensitivity differences of the data in Table II, in the three columns of DIF items in 40 iterations, were tested simultaneously with the anova techniques. Then tested further with the Tukey technique, to see differences in sensitivity between pairs of DIF methods.

IV. RESULTS

The number of UN items detected as DIF by the three methods (MH, LR, IRT-LR) are listed in Table III. In general, the number of DIF items decreased, as the number of items decreased in the analysis group. In the IRT-LR and MH methods, the number and percentage of DIF items continued to decline in the three analysis groups. While the LR method, in addition to decreasing, there was a percentage of DIF items increased, namely 9.4% in the group (4862.32) to 16.7% in the group (4862.18). It can be concluded, in the number of constant of participants (4862), the sensitivity of the three methods decreases as the items in the analysis group decrease. In other words, the number of items (length of test) affects the level of sensitivity of the DIF method. The longer the test, the sensitivity of the DIF method increases, or the shorter the test, the sensitivity of the method decreases.

In the analysis group (4862,40), IRT-LT became the most sensitive method of the three methods. It detected 5 DIF items (12.5%). The second sequence of the LR method, detected 4 items of DIF (10%), and the third sequence of MH, detected 3 DIF items (7.5%). In the group (4862,32), the IRT-LR method and LR were equally strong in sensitivity, both detected 3 items of DIF (9.4%). MH only detects 2 items of DIF (6.3%). But in the analysis group (4862,18), the LR method became the most sensitive method among the three methods. It detects 3 DIF items (16.7%), while the IRT-LR and MH methods both detected only 1 DIF item (5.6%).

From Table IV, overall UN items detected as DIF were more favorable for the men group than for the women group. Four DIF items detected by the LR method, 3 items (75%) benefited men and 1 item (25%) benefited women. In the MH method, from 3 DIF items, 2 items (66%) benefit men, and 1 item (33%) benefits women. In the IRT-LR method, from 5 DIF items, 1 DIF type is uniform, benefiting men (harming women). It can be concluded, the 2015 Mathematics National Exam items tend to be DIF in women sex.

Table III. Number of DIF Items Detected by MH, LR and IRT-LR Methods

Analysis Group	Number of DIF item		*No.Item DIF*	Number of DIF item		*No.Item DIF*	Number of DIF item		*No.Item DIF*
	MH	(%)		LR	(%)		*IRT-LR*	(%)	
(4862,40)	3	7.5	*2,11,25*	4	10.0	*2, 11, 25, 32*	5	12.5	*4,10,22,24,26*
(4862,32)	2	6.3	*11, 25*	3	9.4	*11,14,25*	3	9.4	*8,20,22*
(4862,18)	1	5.6	*11*	3	16,7	*3,7,11*	1	5.6	*16*

Table IV. DIF items Advantages Reference and Focus Group

Method	Item DIF in Analysis Group (4862,40)	Amount of DIF items		
		Advantage		*Nonuniform*
		Men	*Women*	
IRT-LR	5	1	0	4

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

MH	3	2	1	0
LR	4	3	1	0

The data in Table V shows the average number of item detected as DIF in 40 times DIF detection in each group, by all three methods. In general, the sensitivity of the three DIF methods decreased as the number of items decreased in the analysis group. When the number of UN participants constant, i.e. 4400 and 3000, the average item detected as DIF decreased when the number of items decreased. So, the length of the test affects the sensitivity of the DIF method. The longer a test instrument, the sensitivity of the DIF method is increasing.

When the number of UN items remained constant in the analysis group, the reduced number of test participants from 4400 to 3000, making the average number of DIF items tended to increase. This means, when the number of items is fixed, the reduced number of participants generally increases the average number of DIF items. In this case, the number of test participants affected the sensitivity of the DIF detection method. The number of participants is getting bigger, the sensitivity of the DIF method is decreasing. At least this fact is evident in the MH and LR methods. While the IRT-LR method shows the opposite. The greater the number of test participants, the sensitivity of the IRT-LR method increases. In the group (4400,40) the average number of items DIF detected by IRT-LR method was 6.4, while in the group (3000,40) it dropped to 5.8.

In the group (4400,40), in 40 times detection of DIF, the IRT-LR method was the most sensitive to detect DIF from all methods (average 6.4). The second sequence is occupied by the LR method (average 5.0), and the third sequence is occupied by the MH method (average 3.5). In the group (4400,32), the IRT-LR and LR methods had the same sensitivity (average 3.2), and the MH method with an average of 2.2. In the group (4400,18), the LR method was the most sensitive of the three methods (average 3.1). The second sequence is filled with the MH method (mean 1.4), and the third sequence is filled with the IRT-LR method (average 1.3).

In the group (3000,40), in 40 times the detection of DIF, the IRT-LR method was the most sensitive of the three methods (average 5.8). The second sequence was occupied by the LR method (average 5.3), and the third sequence was the MH method (mean 4.1). In the group (3000,32), the LR method was the most sensitive of the three methods (mean 3.5). The second sequence was occupied by IRT-LR (average 3.2), and the third sequence was MH method (average 2.5). In the group (3000,18), the LR method remained the most sensitive of the three methods (average 2,3). The second sequence was occupied by the MH method (average 1.5), and the IRT-LR turned out to be the least sensitive (average 1.4).

Table V. Average number of DIF items by 3 Methods at 40 Iterations

Analysis Group	Average number of DIF items (at 40 Iterations)		
	MH	LR	*IRT-LR*
(4400,40)	3.5	5.0	6.4
(4400,32)	2.2	3.2	3.2
(4400,18)	1.4	3.1	1.3
(3000,40)	4.1	5.3	5.8
(3000,32)	2.5	3.5	3.2
(3000,18)	1.5	2.3	1.4

The data in Table VI presents the results of hypothesis testing from the four research hypotheses (1 major hypothesis, 3 minor hypotheses) in 6 analysis groups. The major hypothesis was tested in each analysis group, tested 6 times. So there is a significant difference in sensitivity between the three DIF methods. Therefore, proceed to the posthoc test (Tukey test) to see pairs of different methods. The first minor hypothesis: "*the IRT-LR method is more sensitive than the MH method*", tested in 4 analysis groups: (4400,40), (4400,32), (3000,40), and (3000,32), with Pr (> | t |) <0.05; not tested in two analysis groups: (4400,18) and (3000,18), with Pr (> | t |)> 0.05. The first minor hypothesis tends to be tested in the analysis group of the number of large items (40) and (32), namely in all UN items and items selected by the CTT. There is an untested tendency in the group analyzing the smallest number of items (18), namely on the items group selected by IRT.

The second minor hypothesis: "*the IRT-LR method is more sensitive than the LR method*", tested 3 times, namely in groups: (4400.40), (4400.18), and (3000,18), with Pr (> | t |) <0.05; not tested 3 times, namely in groups: (4400,32), (3000,40), and (3000,32). This hypothesis tends to be tested in the analysis group the number of large participants (4400) and in the smallest number of items groups (18); But it tends to be untested in groups of number 32 items, namely items from CTT selection. Also, it was not tested on the group of 3000 participants with 40 (UN items) and 32 (CTT selection items). The third minor hypothesis: "*the LR method is more sensitive than the MH method*", tested in all analysis groups: (4400,40), (4400,32), (4400,18); (3000,40), (3000,32), and (3000,18), with Pr (> | t |) <0.05. This hypothesis ~~be~~ [is] the only one research hypothesis that consistently tested in all analysis groups. Significantly the LR method is more sensitive than the MH method.

Table VI: Pairing Hypothesis Test Results (Tukey). Multiple Comparisons of Means: Tukey Contrasts

Analysis Group	Null Hipotheses	Estimate	Std. Error	t value	Pr(>|t|)
(4400,40)	MH – *IRT-LR* == 0	-2.95	0.2299	-12.833	<0.000 ***
	LR – *IRT-LR* == 0	-1.475	0.2299	-6.417	<0.000 ***
	LR – MH == 0	1.475	0.2299	6.417	<0.000 ***
(4400,32)	MH – *IRT-LR* == 0	-0.975	0.1794	-5.436	<0.000 ***
	LR – *IRT-LR* == 0	0.05	0.1794	0.279	0.958
	LR – MH == 0	1.025	0.1794	5.715	<0.000 ***
(4400,18)	MH – *IRT-LR* == 0	0.1	0.1404	0.712	0.757
	LR – *IRT-LR* == 0	1.8	0.1404	12.821	<0.000 ***
	LR – MH == 0	1.7	0.1404	12.108	<0.000 ***
(3000,40)	MH – *IRT-LR* == 0	-1.65	0.4273	-3.861	<0.001 ***
	LR – *IRT-LR* == 0	-0.5	0.4273	-1.17	0.473
	LR – MH == 0	1.15	0.4273	2.691	0.022 *
(3000,32)	MH – *IRT-LR* == 0	-0.775	0.29	-2.672	0.023 *
	LR – *IRT-LR* == 0	0.3	0.29	1.034	0.557
	LR – MH == 0	1.075	0.29	3.706	0.000 ***
(3000,18)	MH – *IRT-LR* == 0	0.075	0.233	0.322	0.944
	LR – *IRT-LR* == 0	0.85	0.233	3.648	0.001 **
	LR – MH == 0	0.775	0.233	3.326	0.003 **

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1. (Adjusted p values reported -- single-step method)

V. DISCUSSION

The major hypothesis "*there are differences in sensitivity of the IRT-LR, MH, and LR methods*" was tested in all analysis groups. That is, there are significant differences in the sensitivity of the three DIF methods. This result is acceptable, considering the characteristics of the three methods are different. The Mantel-Haenszel (MH) method itself is a nonparametric type, along with the Logistic Regression (LR) method. Both methods are also different, one effectively detects uniform DIF (MH), and one effectively detects uniform and nonuniform DIF (LR). While IRT-LR is a type of parametric method, which estimates item and participants parameters in detecting DIF. Unlike the MH and LR methods, it is enough to use the observed score as the basis for matching participants' abilities.

The first minor hypothesis "*IRT-LR method is more sensitive than the MH method*", tested in 4 analysis groups: (4400,40), (4400,32), (3000,40), and (3000,32). This testing, more as an influence factor number of test participants, compared to the number of items (length of the test). When the number of participants is constant, and the number of items drops from 40 to 32, the IRT-LR remains more sensitive than MH. That means, the IRT-LR method in the condition of a large number of participants (also the number of large items), is more sensitive than the MH method. This result is in accordance with the findings of Kubra Atalay Kabasakal, et.al. where the IRT-LR method is more sensitive than the MH method [53]. In the groups: (4400,18) and (3000,18), the IRT-LR method did not prove to be more sensitive than the MH method. Despite the large number of participants, the sensitivity of the IRT-LR method is more affected or decreased compared to MH. This can be caused by the factor of the number items of 18. Where, first, it has been below the minimum number recommended by Hulin, Lissak and Drasgow, namely 1000 test participants for 60 items, or 2000 test participants for 30 items for the 3 parameter logistics model (3PL); second, the nature of items (18) is also affected, as a result of IRT selection. On this items the IRT requirements have been applied, although not strictly (data does not fit the 3PL model). This is not beneficial for IRT-LR in estimating item and participants parameters. As a result, the sensitivity of the IRT-LR method is more affected or decreased. In this position, the MH method is a nonparametric method, more suitable for the number of small items (below 20)[54], and its sensitivity begins to increase. These results are close to the results of Bartosz Kondratek's study, Magdalena Grudniewska [55], where the MH method has higher power than IRT-LR when there is uniform DIF.

The second minor hypothesis "*the IRT-LR method is more sensitive than the LR method*" gives some results. *(a).* in the group (4400,40), the IRT-LR method was significantly more sensitive than the LR method. But in the group (3000,40), with the same test length (40), the sensitivity of the IRT-LR method was not significantly different from the LR method. In this condition, the effect of the reduced number of participants began to emerge. Larger numbers of test takers support the IRT-LR method, which matches to the large data. Compared to the LR method which is more suitable for a small number of data. *(b).* in the group (4400,32) and (3000,32), the sensitivity of the IRT-LR method was not more sensitive than the LR method. In this case, the weakened sensitivity of the IRT-LR to the LR, not because

of the number of participants. But rather, it relates to the nature of the items of total number 32 (CTT selection results), which have not been selected in IRT. So that it supports the nonparametric LR method compared to IRT-LR. Therefore, in both of the two groups, it was plausible that IRT-LR was not significantly more sensitive than the LR method. *(c)*. Interesting things occur in analysis groups: (4400,18) and (3000,18). In both groups, the IRT-LR method again proved to be significantly more sensitive than the LR method. The causal factor can be drawn, more so to the nature of the items amounting to 18 (IRT selection results). These items have applied the IRT requirement (though not strictly), so it supports the IRT-LR method, compared to the LR method. Also the use of the 3PL logistic model, makes it more in line with the IRT-LR method, compared to the LR method that is more suited to the 1PL logistical model or the assumption of linear relationships. [54]. Including the number of items (18), the LR (nonparametric) method is more suitable for long tests (more than 20 items) [56]. Because for shorter tests, the performance of nonparametric methods is debatable.

By linking the MH and LR methods together to the IRT-LR method in the analysis group (4400,18) and (3000,18), a new result was obtained. In these groups, the MH method is relatively stronger to compensate for the sensitivity of IRT-LR, compared to the LR method for IRT-LR. In these groups, the average number of DIF items detected by the IRT-LR method is still higher than the MH method, but the difference is not significant. Meanwhile, the IRT-LR method is more sensitive than the LR method. This is due to MH method factors which are not related to the model, but the LR method requires a linear model and contains elements of the logistical model. The use of the 3PL model does not support the LR method, but does not reduce the sensitivity of the MH method, and supports the IRT-LR method. The IRT-LR and LR methods are both related to the model. The LR method is more appropriate to use a linear model or 1PL model, compared to 3PL. While the 3PL model is more suitable for the IRT-LR method. So that it is acceptable, that in both of these groups (the IRT selection item group), the IRT-LR method is no more sensitive than the MH method, but is more sensitive than the LR method. The interesting thing is the opposite, occurs in the analysis group (4400,32) and (3000,32), where the LR method is stronger to compensate for the sensitivity of the IRT-LR method, compared to the MH method.

Fourth, the hypothesis "*the LR method is more sensitive than the MH method*", is consistently significant in all analysis groups. The results are consistent with the researchers' initial expectations that the Logistic Regression (LR) method is more sensitive than the Mantel-Haenszel (MH) method of detecting differential item functioning (DIF). This is consistent with the results found by Rogers and Swaminthan concluded that the LR method is as powerful as MH in detecting uniform DIF, and is more powerful than MH procedures in detecting nonuniform DIF [99]. In other words, all items marked DIF by the MH method can also be marked DIF by the LR method.

REFERENCES

1 Azwar, S. (2010). Keputusan seleksi dalam high stake exams: wacana psikometris. Fakultas Psikologi UGM

2 Ashadi, Rice, S. (2016). High stake testing and teacher access to professional opportunity: lesson from Indonesia, Journal of education Policy, 31:6,h.728, DOI: http://dx.doi.org. /10.1080/02680939.2016.1193901

3 Au, W. (2007). High-stakes Testing and Curricular control: A Qualitative Metasynthesis. Educational Researcher, Vol. 36,No.5, h.258. DOI: 10.3102/0013189X07306523. http://er.aera.net

4 Nurjanah, & Marlianingsih, N. (2015). Analisis soal pilihan ganda dari aspek kebahasaan, Faktor jurnal Ilmu kependidikan, Volume II Nomor 1, Maret 2015, h. 70

5 Lampiran Peraturan Menteri Pendidikan Nasional Nomor 20 Tahun 2007 Tanggal 11 Juni 2007 Standar Penilaian Pendidikan

6 Holland, P.W., Wainer, H. (1993). *Differential Item Functioning*. New Jersey: Lawrence Erlbaum Associates Publisher.

7 Budiono (2005). Perbandingan Metode Mantel-Haenszel, SIBTEST, Regresi Logistik, dan Perbedaan dalam Mendeteksi Keberbedaab Funngsi Butir, Jurnal Penelitian dan Evaluasi Pendidikan, No.2,Tahun VII.

8 Kartowagiran, B. (2005) Perbandingan Berbagai Metode untuk Mendeteksi Bias Butir. Diserteasi, Yogyakarta: Pascasarjana UGM

9 Retnawati, H. (2013). Pendeteksian keberfungsian butir pembeda Dengan indeks volume sederhana Berdasarkan teori responsi butir multidimensi. Jurnal Penelitian dan Evaluasi Pendidikan, Tahun 17, Nomor 2, h.285.

10 Retnawati, H. dan Hidayati, K. (2006). Pendeteksian Bias Tes Butir Perangkat Soal Matematika Ujian Nasional SLTP Berdasarkan Teori Respons Butir.

11 Triyatno, N., Sriyono, Ngazizah, N. (2014). Bias Gender Ujian Akhir Semester Genap Fisika Kelas X SMA Negeri Kabupaten Purworejo Tahun Pelajaran 2012/2013. Jurnal Radiasi. Vol.4.No.1.

12 Zumbo,B.D. (1999). *A handbook on the theory and methods of differential item functioning* (*DIF*): Logistic regression modeling as a unitary frame work for binary and Likert-type (ordinal) item scores. Ottawa,Canada: Directorate of Human Resources Research and Evaluation, Depertment of National Defence,h.25 =[12] hal 13

13 Huang, J. ; Han. T. (2012) Revisiting Differential Item Functioning: Implications for *Fairness* Investigation. International Journal of Education ISSN 1948-5476 2012, Vol. 4, No. 2. h.75

14 Camilli, G., & Shepard, L. A. (1994). *Methods for identifying biased test items*. Sage Publication. Thousand Oaks.

15 Huang, J. ; Han. T. (2012). Revisiting Differential Item Functioning: Implications for *Fairness* Investigation. International Journal of Education ISSN 1948-5476 2012, Vol. 4, No. 2. h.75

16 Clauser B. E., & Mazor, K. M. (1998). Using statistical procedures to identify deferentially functioning test items. Educational Measurement: Issues and Practic, 17, 31-44.

17 Abedalaziz,N.,Ismail,W.,Hussin,Z. (2011). Detecting a Gender-Related *DIF* Using Logistic Regression and Transformed Item Difficulty. US-China Education Review B5 734-744.

18 Cuevas, M.; Cervantes, V. H. (2012). Differential Item Functioning with Logistic Regression. Mathematics and Social Sciences: 50 année, n_ 199, 2012 (3), p. 46

19 Camilli, 2007 in Kondratek, B., Grudniewska,M. (2014) *Comparison of Mantel–Haenszel with IRT procedures for DIF detection and effect size estimation for dichotomous items*, Edukacja 2014, h.92, An interdisciplinary approach, ISSN 0239-6858
20 Caplan, J. B., & Caplan, P. J. (2005). The perseverative search for sex differences in mathematics abilities.
21 Battista, M. T. (1990) Spatial visualization and gender differences in high school geometry. Journal for Research in Mathematics Education, 21(1), 47-60.
22 Abedlaziz, Ismail, & Hussin (2011). In Gallagher, A.M., De lisi, R., Holst, P.C., McGillicuddy-De Lisi, A.V., Morely, M. & Cahalan, C. (2000). Gender Differences in Advanced Problem Solving. Journal of Experimental Child Psychology, 75, 165-190'
23 Abedalaziz, N. (2010), Detecting Gender Related DIF using Logistic Regression and Mantel-Haenszel Approaches. International Conference on Learner Diversity 2010. Faculty of Education, University of Malaya,50603 Kuala Lumpur,Malaysia. h.2
24 Dorans, N. J., & Holland, P. W. (1993). DIF detection and description: Mantel-Haenszel and standardization. In P. W. Holland and H. Wainer (Eds.) Differential Item Functioning (pp35-66). Hillsdale, NJ: Lawrence Erlbaum.
25 Holland, P.W., & Thayer, D.T.,(1998). Differential Item Performance and the Mantel-Haenszel Procedure. In Wainer, H. & Braun, H. (Eds), Test validity (pp.129-145). Hillsdale, NL: Erlbaum
26 Swaminathan, H., & Roger, H. J. (1990). Detecting Item Differential Functioning using logistic regression procedures. Journal of Educational Measurement, 27, 361-370.
27 Dorans, N. J., & Holland, P. W. (1993). DIF detection and description: Mantel-Haenszel and standardization. In P. W. Holland and H. Wainer (Eds.) Differential Item Functioning (pp 35-66). Hillsdale, NJ: Lawrence Erlbaum.
28 Dorans, N. J., & Kulick, E. (1983). Demonstrating the utility of the standardization approach to assessing unexpected differential item performance of female candidates on the Scholastic Aptitude Test. Journal of Educational Measurement, 23, 355-368.
29 Shealy, R., & Stout, W. F. (1993). A model-based standardization approach that separates true bias/DIF from group differences and detects test bias/DIF as well as item bias/DIF. Psychometrika, 58, 159-194.
30 Hambleton,Swaminathan & Rogers. (1991). Fundamentals of Item Response Theory. USA: Sage Publication, Inc.
31 Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P. W. Holland & H. Wainer (Eds.), Differential item functioning (pp. 35-113). Hillsadle, NJ: Lawrence Erlbaum.
32 Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P. W. Holland & H. Wainer (Eds.), Differential item functioning (pp. 35-113). Hillsadle, NJ: Lawrence Erlbaum.
33 Embretson, S.E, & Reise, S.P. (2000). Item Response Theory for Psychologists. Lawrence Erlbaum: Mahwah, NJ.
34 Hidalgo, M. D. & López-Pina, J. A. (2004). Differential Item Functioning Detection and Effect Size: A Comparison between Logistic Regression and Mantel- Haenszel Procedures. *Educational and Psychological Measurement*, 64, pp. 903-915.
35 Moses, T., Miao, J. & Dorans, N. J. (2010). A Comparison of Strategies for Estimating Conditional DIF. *Journal of Educational and Behavioral Statistics*, 35, pp. 726–743.
36 Vaughn, B. K., & Wang, Q. (2010). DIF trees: Using classification trees to detect differential item functioning. Educational and Psychological Measurement, 70(6), 941-952. doi: 10.1177/00131 64410379326
37 Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 35-113). Hillsadle, NJ: Lawrence Erlbaum.
38 Atar,B., Kamata, A. (2011). Comparision of IRT Likelihood Ratio test and Logistic regression DIF detection Procedures. Journal of Education.
39 Wiberg, M. (2007). *Measuring and detecting differential item functioning in criterion-refrenced licensing test: A Theoritic Comparison of Methods*. EM: UMEA Universitet, No. 60.
40 Thissen D. *IRTLRDIF* v.2.0b (2001): Software for Computation of the Statistics Involved in Item response Theory Likilihood-Ratio Tests for Differential Item Functioning. University of North Carolina at Chapel Hill: L.L. Thurstone Psychometric Laboratory; 2001. p.2
41 Dorans, N. J. and Holland, P. W. (1993). DIF detection and description: Mantel–Haenszel and standardization. In P. W. Holland and H. Wainer (eds.),*Differential item functioning* (pp. 35–66). Hillsdale,NJ: Lawrence Earlbaum
42 Holland, P.W., & Thayer, D.T.,(1998). Differential Item Performance and the Mantel-Haenszel Procedure. In Wainer, H. & Braun, H. (Eds), *Test validity* (pp.129-145). Hillsdale, NL:Erlbaum
43 Linacre,J.M. & Wright, B.D. (1989). Mantel-Haenszel DIF and PROX are equivalent! Rasch Measurement Transactions, 3(2),52-53
44 Zieky, M. (1993). Practical questions in the use of DIF statistics in test development. In P. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 337-347). Hillsdale,NJ: Erlbaum
45 Swaminathan H. & Rogers, H.J. (1990) Detecting differential item functioning using logistic regression procedures. Journal of Educational Measurement,27 (4),361-370
46 Karami, H. (2012). An Introduction to Differential Item Functioning. The International Journal of Education and Psychological Assessment , 2012, Sept,Vol 11 (2), p.63
47 Zumbo,B.D. (1999). A handbook on the theory and methods of differential item functioning (DIF): Logistic regression modeling as a unitary frame work for binary and Likert-type (ordinal)
48 Zhang, Y. (2015). Multiple Ways to detect Differential Item Functioning in SAS. Paper 2900-2015, Educational testing Service. h.3
49 Zumbo,B.D. (1999). A handbook on the theory and methods of differential item functioning (DIF): Logistic regression modeling as a unitary frame work for binary and Likert-type (ordinal)
50 Swaminathan H. & Rogers, H.J. (1990) Detecting differential item functioning using logistic regression procedures. Journal of Educational Measurement, 27(4),361-370
51 Jodoin, M.G., Gierl, M.J. (1999). Evaluating Type I Error and Power Using an Effect sixe measure with Logistic Regression Procedure for DIF detection. Alberta: University of Alberta, Edmonton,1999, h.6-12
52 Thissen D. (2001). *IRTLRDIF* v.2.0b: Software for Computation of the Statistics Involved in Item response Theory Likilihood-Ratio Tests for Differential Item Functioning. University of North Carolina at Chapel Hill: L.L. Thurstone Psychometric Laboratory; 2001. p.8
53 Kabasakal, K.A., Nihan Arsan, Bilge Gok, Hulya Kelecioglu. (2014). Comparing Performances (Type I error and Power) of IRT Likelihood ratio SIBTEST and Mantel-Haenszel Methods in the Determination of Differential Item Functioning. Educational Sciences: Theory & Practice 14(6).

54 Marie Wiberg (2007). Measuring and detecting differential item functioning in criterion-refrenced licensing test: A Theoritic Comparison of Methods (EM: UMEA Universitet, No. 60, 2007), h.15-16
55 Kondratek, B., Grudniewska, M. (2014). Comparison of Mantel–Haenszel with IRT procedures for DIF detection and effect size estimation for dichotomous items, Edukacja 2014, 5(130), 92–111. An interdisciplinary approach, ISSN 0239-6858.
56 Gabriel E. Lopez (2012). Detection and classification of DIF Types using parametric and nonparametric methods: A comparison of the IRT-Likelihood ratio Test, Crossing-SIBTEST, and Logistic regression procedures. (University of South Florida: Desertasi, 2012),h.26
57 Rezaee, A. A., Shabani, E. (2010). Gender Differential Item Functioning Analysis of the University of Tehran English Proficiency Test. Pazhuhesh-e Zabanha-ye Khareji, No. 56, Special Issue, English, Spring.