Differential Item Functioning National Examination on Device Test Mathematics High School in Central Java

Samsul Hadi[1], Basukiyatno[2], Purwo Susongko[3]
Pancasakti University Tegal, Central Java, Indonesia[1,2,3]

{samsul_hadi@upstegal.ac.id[1], basukiyatno@upstegal.ac.id[2], purwosusongko@upstegal.ac.id[3]}

Abstract. This research aims to determine The Differential Item Functioning that occurs between participants from the district and the test takers from the city area of the item from the National Math Exam Test in the academic year 2014 / 2015 in the Central Java in terms of item response theory. The study population was a student's response participant for National Examination in Mathematics High School Scince High School in 1103 Packs 2014 / 2015 school year in Central Java. Samples were taken by using purposive random sampling systematic spread over 6 districts and 5 municipalty with total participants of 4.107 students. Sampling is reached through two stages. The first stage to determine the area. The second phase to determine the answers to the test taker UN Mathematics and Natural science packets about 1103 school year 2014 / 2015. The first stage of sampling taken based on certain considerations or purposive sample.the consideration is to notice differences in the characteristics of the territory UN organizers. Differences, social, economic, cultural, and political in each rayon organizer UN cause differences in student background so the potential to produce grains *DIF* on the test. To detect the *DIF*, the territory divided into two groups of reference (A) and the study group (B). A group in this study is rayons which area in districts thath Brebes, Tegal, Pekalongan, Kendal, Semarang regency and district, Magelang with total sample size of 2.121 responses answers. Group B is determined that all five regions, namely the municipal area of Semarang, Tegal, Pekalongan, Salatiga, and Magelang City with a total sample size of 1.986 response answers. The result showed using the Rasch model of item response theory to estimate the level of difficulty grain. Waldtest for detecting Differential Item Functioning (DIF) which is operated by using a software program R version 3.2.3. There are 22 test items were detected between participantsexperienced DIF test of the district and of the city of test participants.

Keywords : Differential Item Functioning National Examination on Device Test Mathematics High Shcool In Central Java.

1 Introduction

ICONS 2020, October 13-14, Tegal, Indonesia

DOI 10.4108/eai.13-10-2020.2303726

The initial procedure conducted to determine whether there is item bias in a test item carried out using the*Differential Item functioning*analysis(*DIF)*. *DIF*analysis is used to identifying all items which have different functions for different groups. A test item indicates *DIF* if students have the same ability but come from different groups, cannot have the same chance to answer correctly (Hambleton, et. *al*., 1991). The next procedure for determining whether an item is biased or not is using logical analysis as to why test items appear more difficult for one group than for another. If an item is relatively more difficult for one group and the source of this difficulty is irrelevant to the test construct, then the item is said to be biased.Thus, a test item containing DIF is not automatically said to be biased because there are many other procedures used to determine whether a test item is biased or not, including logical analysis from experts in the field of study.

National Math exam constructsare theoretically designed to measure one dimension which is the dimension of mathematics ability. However, the mathematic contstructs may contain two dimensions. If these two dimensions related to constructare called auxiliaryand called nuisance if they are not related to construct (Roussos and Stout inGierlet al., 2003). Items containing*DIF* are not detrimental because there are auxiliary dimensionsin it.(Douglas et al., 1996).

The items containing two dimensions and being disruptive must be revised or discarded, the items containing two dimensions do not interfere with will still be used. The additional two dimensions may benefit certain groups.The groups mentioned including regional groups, gender, religion, races and others. The presence of these two dimensions must be discarded.

The National Examination in Central Java is held by the areas which include 29 districts and 6 municipalities (Depdikbud, 1999). Obtaining a different UN average for each area can cause problems. The problem is whether the difference is due to ability differences or because of the item bias which benefits students in a certain area. The item bias is very possible because the students' background in municipal districts (urban areas) will be different from the students' background in district areas which are mostly rural areas.

To find out whether there is item bias on the items of National Math Exam test in Central Java for the 2014/2015 academic year, it is necessary to conduct a DIF study based on differences in the National Exam regions.

2 National Exam Test

2.1 Evaluation

Ahman&Glock (1981) stated that educational assessment is a systematic process in order to obtain clear evidence about the effectiveness of educational activities. Furthermore, it is argued that the evaluation can be conducted during the program implementation process and the end of the program implementation. The evaluation conducted in the implementation phases is called formative *evaluation*, while the one conducted at the end of the program implementation is called*summative evaluation*. The goal of formative evaluation is to monitor the achievement of the learning objectives or the effectiveness of a predetermined program until a specified point in time, while summative evaluation aims to determine the students' achievement level of learning objectives as a whole from a topic.

In a broad sense, evaluation is a process of planning, obtaining, and providing information that is crucial for making a decision (Mehrens& Lehmann, 1978:5).

2.2 Definition of Test, Classification Test, and Achievement test

2.2.1 Definition of test

A test is a set of questions which have a correct or incorrect answer. The test is also defined as a number of questions requiring answers, or a number of questions that must be responded to with the aim of measuring a person's ability level or revealing certain aspects of the person being tested (Mardapi, 2008:67). Anne Anastasi (1976) in Azwar (2007:3) stated that the test is basically an objective and standardized measurement of behavior samples.

2.2.2 Classification Test

CronbachinAzwar (2007:5) divides test into two big groups, namely tests of maximum performance and tests of typical performance. This study will examine the test groups included in the classification of achievement test. In this case, the achievement test which is given in groups at the national level is usually called the National Exam.

2.2.3 Achievement Test

In relation to learning achievement, (1996:226) stated that learning achievement is evidence of the success that has been achieved by a person. Then, learning achievement is the maximum result achieved by a person after doinglearning efforts. Learning achievement can be measured through test called achievement test. The goal of achievement test is to reveal a person's success in learning. Testing basically digs up information that can be used as a basis for decision making.The achievement test is a planned test to reveal the maximum performance of a subject in mastering materials that have been taught. In formal education activities, achievement test is in the form of formative test, summative test, even and college entrance exams (Azwar, 2005: 8 - 9). Students' learning achievement can be seen after the evaluation is conducted. The results of evaluation can show the category of students' achievement (high or low).

2.2.4 Types of achievement tests

The types of tests used in educational institutions can be classified into two. The first one is objective test which can be seen through the scoring system, it means that whoever checks the test answer sheet, it will result the same score. The second one is non-objective tests of which scoring system influenced by the examiners. In other words, it can be concluded that the objective test is the one with objective scoring, while the non-objective is influenced by the examiners' subjectivity. Based on the National Exam, the test instruments used are usually objective in the form of multiple choice with more than two alternatives. Thus, what is examined in this study is the empirical analysis of test items on achievement test items in the form of multiple choice.

2.2.5 Senior High School National Exam (UN SMA)

Regulation of the Minister of Education and Culture of the Republic of Indonesia number 5 of 2015 article 1 paragraph (5) describes that the National Exam (UN) is an activity to measure and achieve the graduates' competence in national on certain subjects. The regulation

of the Minister of Education and Culture of the Republic of Indonesia on the execution, implementation and supervision of the National Examination Article 21 paragraphs (1) and (2) states that the results of the UN are used for:

1) Mapping the program quality and/or education unit;
2) Selection consideration for the next level of education; and
3) Consideration in fostering and providing assistance to educational units to improve the quality of education. The Ministry maps the UN results at the educational units, district/city, provincial, and national levels.

3 The Differential Item Functioning

3.1 Test Theories

3.1.1 Item Response Theory

In evaluations conducted in education, students get 1 score if they answer multiple choice items correctly and 0 if it is incorrect.In classical theory, students' abilities are expressed by the total score they get. This approach pays less attention to the interaction between each student and the items. The item response theory is an alternative approach that can be used in analyzing a test. There are two principles used, namely the principle of relativity and the principle of probability. In the principle of relativity, the basic unit of measurement is not students or items, but rather students' abilities relative to items. If$\beta_n$is the index of students' nth ability on the measured trait, and$\vartheta_i$is the index difficulty level of the ith item relative to the measured trait, then β_n atau $\vartheta_i$is not the unit of measurement, but rather the difference between abilities and from students relative to the item difficulty or (β_n - ϑ_i) needs to be evaluated. As an alternative, a comparison between ability and difficulty level can be used. If the students'abilities exceed the item difficulty level, then the students' response are expected to be correct, and if the students' ability are less than the item difficulty level, then the students' response are expected to be incorrect(Keeves and Alagumalai, 1999:24).

In the item response theory, the principle of probability becomes a concern. Suppose that the nth student's ability is expressed by θ_n and the item difficulty level isΔ_i, then according to the principle of relativity, if θ_n>Δ_i students are expected to answer correctly, and θ_n<Δ_i students are expected to answer incorrectly. The probability of a correct response is in the range 0 to 1.0 and this prevents the data from being expressed as an interval scale. The raw scores resulted from this method are difficult to represent as a scale. To solve this problem, logistic regression can be used, so that the relationship between the item difficulty and the chance of getting it right is not a linear. Item response theory, the mathematical model has the meaning that the probability of a subject to answer an item correctly depends on the subject's ability and item characteristics. This means that test takers with high abilities will have a greater probability of answering a question correctly when compared to participants with low abilities.

3.1.2 Assumptions of Item Response Theory

Hambleton, Swaminathan & Rogers (1991) stated that there are three assumptions underlying item response theory, those are unidimensionality, local indepence, and parameter invariance. These three assumptions can be explained as follows.

a) Unidimensionality

Unidimensionality means that each test item measures only a single ability. For example, in the mathematics achievement test, the items contained in it only measure students' ability in mathematics, not from other fields. In practice, the unidimensional assumptions cannot be strictly met because of cognitive, personality and test-taking factors, such as anxiety, motivation, and a tendency to guess. Therefore, the unidimensionality assumptions can be demonstrated only if the test contains only one dominant component that measures the subjects' achievement. (Hambleton, at, *al,.* 1991). The second assumption is the test measures one dimension of ability. Ideally, each item made is expected to measure one of the test takers' abilities, not two or more of the test takers' abilities. A test where all items measure the same ability is a test that is formed from items that only measure one ability.

b) Local independence

Local independence does not influence trait of a test item and a test taker. If the test takers respond to the test items arranged based on the item difficulty level or arranged randomly, then in that test if all the items are local independence, then the position of the items on the test does not affect the test takers' answer score.Differences in position occur when interpreting response patterns. Tests arranged according to the item difficulty level will be easier to use to see the test takers' response patterns. Local independence occurs when test takers and items in the sub-population are statistically independent. In other words, the number of items that the test takers respond to in a homogeneous or heterogeneous sub-population are independent of each other. However, there is a more lenient rule which does not use the term statistical independence but rather the term correlation. According to Hambleton, Swaminathan, & Rogers (1991: 10), local independence is mathematically expressed as:

$$P(u1,u2,\ldots,un|\theta) = P(u1|\theta).P(u2|\theta)\ldots. P(un|\theta) = \prod_{i=1}^{n} P(U_i|\theta) \quad (1)$$

Annotation :
I : 1, 2, 3, …n
n : number of test items
P(ui|θ) : the probability of a test taker who has ability θcan answer i items correctly.
P(u1,u2,…,un|θ) : the probability of a test taker who has the ability θcan answer items 1 to n correctly

c) Parameter Invariance

According to Hambleton, Swaminathan, & Rogers (1991: 18), the invariance of ability parameters can be investigated by proposing two or more test sets that have different levels of difficulty in a group of test takers. The invariance of the ability parameters will be proven if the results of the test takers' ability estimation are not different even though the tests conducted have different levels of difficulty. In item response theory, apart from the assumptions previously described, the important thing to note is the selection of the right model. Selection of the right model will reveal the true condition of the test data as a result of

measurement. There are 3 models of the relationship between ability and item parameters, namely 1 parameter model (Rasch model), 2 parameter model, and 3 parameter model.

$$\text{Pi}(\theta) = \left(\frac{e^{(\theta - b_i)}}{1+e^{(\theta - b_1)}}\right), \text{where i : 1,2,3, …,n} \quad (2)$$

Pi (θ) : probability of a randomly selected test taker who has the ability θ can answer item I correctly.
θ : ability level (as independent variable)
bi : theindexdifficulty of item i
e : a natural number whose value is close to 2.718
n : number of items in the test

The bi parameter is a point on the ability scale for a 50% chance of answering correctly. For example, a test item with bi parameter = 0.3, meaning that it requires a minimum ability of 0.3 on a 50% chance to be able to answer correctly. The greater the value of the bi parameter, the greater the ability needed to answer correctly with a 50% chance. In other words, the greater the value of the bi parameter, the more difficult the item is. The relationship between the probability of answering correctly Pi (θ) and ability level (θ) can be illustrated as an item characteristic curve (ICC). A characteristic of item parametersinvariance means that the test item parameters obtained from different subject groups will always be the same (Hambleton, at, al, 1991). For example, subject groups are based on race, ethnic origin, gender, ability level, and so on.

The explanation of item parameter invariance is based on groups. In one population, the participants of UN Math SMA IPA Package 1103 for the 2014/2015 academic year in Central Java were identified into two groups, namely, the regency group as group A and the city group as group B, each of which had abilities that were normally distributed with mean of μ_θ, and standard deviation of σ_θ, which are different. The proportion or probability of answering correctly will not be affected by which group it belongs to, but only by the ability level θ.

d) Item Response Theory Model

Item response theory constructs a model which relates item characteristics to participant characteristics. With a number of certain conditions, this relational model is made to apply spontaneously to any group of items and groups of participants who meet the requirements. Item characteristics and participant characteristics are linked by a model in the form of a function or graphical environment. The requirements mentioned are stated by a number of parameters. Those are divided into two, namely item parameters and participant parameters. There are several models of response items or item characteristics, including the logistic regression model. Furthermore, according to the limitations and formulation of research problems, which will be discussed further is the logistic regression model.

The logistic regression model consists of a one-parameter logistic model (1PL), a two-parameter logistic model (2PL), and a three-parameter logistic model (3PL). All three apply to items with a dichotomous response, namely items whose scores are true or false. As the name implies, the three-parameter model has three item parameters, namely item difficulty, distinguishing power, and pseudo-guessing. The two-parameter logistic model has two item parameters, namely, the itemdifficulty level and the distinguishing power, while the pseudo estimation parameter is considered zero. The one-parameter logistic model has one item

parameter, namely, the level difficulty, while the distinguishing power parameter is considered the same, and the pseudo-guessing parameter is equal to zero.

f) One-Parameter (1PL) Logistic Model

In the one-parameter logistic model, the probability of a test taker is determined by one item characteristic, namely the item difficulty index, so that the only problem level being tested. According to Hamelton, Swaminathan, and Rogers (1991: 12) mathematically the one parameter logistic model can be stated as follows

$$P_i(\theta) = \frac{e^{D(\theta - b_i)}}{1+e^{(\theta - b_i)}} \; where \; i: 1,2,3, \dots, n \qquad (3)$$

Pi (θ) : the probability of a randomly selected test taker who has the ability ɵ can answer item i correctly.
a_i :the distinguishing power of item i,
b_i :difficulty level parameter, which is one point on the ability scale where the probability of answering correctly is 0.5.
c_i :chance of correct guessing of item i.
θ :parameters of students' abilities,
D :The scale factor was included to make the logistic function as similar possible to the normal ogive function (D = 1,7).
E :transcendental number which has a value of 2.718.

g) Parameter Estimation

The test items analysis using the logistic model item response theory method, in addition to obtaining characteristics or in the form of one to three item parameters, namely: difficulty level (b), distinguishing power (a) and guessing factor (c), one participant parameter is obtained, namely the ability parameter. (θ). Although there seems to be a similarity, in the form of difficulty level and distinguishing power, in item characteristics between classical test theory and item response, they are obtained in very different ways. In item response theory, the probability of answering correctly depends on the test takers'abilities and item parameters. Since these parameters were not known in advance, it was necessary to estimate them based on the test takers' responses to the test items. There are several ways to estimate the parameters, namely (1) joint maximum probability, (2) marginal maximum probability, (3) conditional maximum probability, (4) joint and marginal bayesian, (5) heuristic, and (6) nonlinear factor analysis(Hambleton, at. al., 1985). The first three methods are the most commonly used.

In the first method, the ability parameters and item parameters are estimated together or simultaneously. The second method, the integrated ability parameters and item parameters were estimated. By determining the parameter estimation, the next item is used to estimate the ability parameters. The third method of estimation is adjusted on the number of correct answers. The fourth method is to use the mean or mode of the distribution of the pre-existing ability parameters and item parameters, removing some of the existing problems such as incorrect and non-convergent parameter estimates, followed by estimating the joint and marginal maximum probability. The fifth method is mainly used in the logistic model and three parameters. The sixth method is to use least squares in the factor analysis.

The parameter estimation uses a fairly complicated mathematical calculation formula, especially if it involves a test with a large number of items which is also responded to by a large number of participants. For instance, if there are n test participants who take the test consisting of m items, then n number of ability parameters will be estimated and at least m item parameters according to the model used. The one-parameter logistic model only estimates m item parameters, the two-parameter logistic model estimates 2m item parameters, and the three-parameter logistic model estimates 3m item parameters. You can imagine how long it will take to complete all these calculations.

Ways to overcome these obstacles, since 1979 several computer programs have been developed including: Bical, Logist, Bilog, Nohram, Microscale, Rascal, Ascal and Rida (Hambleton et. al., 1991). There is also an R program performed to complete the parameter estimation of the item response theory. The packages contained in the R program are eRm, ltm, lme4, pIRasch, mirt, mokken, KernSmoothIRT, MCMCpack, pscl, Dppackage, mRm, psychomix, mixRasch, psychotree, difR, lordif, kequate, EstCRM, catR.

The ltm package can analyze dichotomous and polytomous data with aitem response theory approach including: Rasch model, 2 parameter model, 3 parameter model, graded response (GRM) and generalized partial credit model (GPCM). The parameter estimation in the ltm package uses the marginal maximum likelihood (MML) method. As for the estimation of test takers' ability parameters, three methods are provided, namely: empirical bayes (EB), multiple imputation (MI) and expected a posteriori scores (EAP).

The eRm package can analyze, including: Rasch Model (RM), linear logistic test models (LLTM), rating scale models (LRSM), partial credit model (PCM), and linear partial credit model (LPCM). Other facilities obtained from the eRm package are the bias test (Waldtes), the ability estimation of test takers, the Anderson Likelihood Ratio test, and so on. According to Sudaryono (2013), the Rasch model can be considered as a very simple form of IRT, namely one parameter for the students' abilities and one parameter for the difficulty levels of the questions. The item characteristic curves for the one-parameter model are expressed as follows:

$$P_i(\theta) = \frac{e^{(\theta - b_i)}}{1 + e^{(\theta - b_i)}} \quad (4)$$

Annotation:

$P_i(\theta)$: the chance of test taker with the ability to answer item icorrectly,

a_i :the parameter of distinguishing power of item i,

b_i :difficulty level parameter, which is one point on the ability scale where the probability of answering correctly is 0.5

c_i :chance of correct guessing item i.

θ :parameters of students' abilities,

D : The scale factor was included to make the logistic function as similar as possible to the normal ogive function (D = 1).

e :transcendental number which has a value of 2.718.

There are similarities between the one-parameter logistic model and the Rasch model. The difference lies in the value of D. In the logistic model the value of D is 1.7, while in the Rasch model the value of D is 1.

3.1.3 Test Bias

Camili and Shepard (1994) stated that the aim of biased research is an attempt to distinct whether the reasons for group differences are real or just artifacts (caused by the measurement process itself). For example, women's scoresin mathtests were lower than the most men, but women also received less math lessons than men in both school and academic, implying that the differences in tests indicate differences in acquired math knowledge rather than bias. Thus, differences in observations caused by bias or distortion in the measurement can be reasonable only if men and women were in the same level and lesson preparation show differences in test results

The main problem about bias in psychological tests is theconstruct validity, namely the scope of which the test items can be assumed to measure a construct or trait that can theoretically be defined. If the test items have the same construct validity for all test takers in the population, test takers with comparable abilities must have an equal chance of being able to answer the items correctly called the test takers' success probability of different groups drawn from the same population. A test item is supposed to be unbiased if the probability of success on that test item is the same for test takers who have the same ability from the same population regardless of group membership in their subgroup (Mazor et. Al., (1995), Shepard et. al., (1981) and Piene (1977) in Osterlind (1983)).

Group differences in test performance or on test items are not automatically indicative of bias because differences in scores may be a reflection of differences in group knowledge and experience. Therefore, the concept of relative difficulty is put forward. Then, bias is operationalized as the relative difficulty of excessive or deviant different items for a particular group or fixed group. Only if an item is relatively difficult for one group and the difficulty is not relevant to the test construct, then the item is said to be biased. Conceptually, bias is the interaction between item performance and group membership (Camili and Shepard, 1994).

The detection of bias in this study was only limited to the internal bias of the items in the UN Math SMA IPAPackage 1103 for the 2014/2015 academic year in Central Java.

3.1.4 DIF

a) Definition of DIF

To keep the difference between relative difficulty and bias, Holland and Thayer in Camilil and Shepard (1994) and Adams in Keeves (1992) named relative difficulty a differential item functioning abbreviated as DIF. Ideally, DIF statistics should be used to identify all items that have different functions for different groups, then after a logical analysis as to why items seem relatively more difficult, a group of DIF items should be identified as biased and should be excluded from the test. Henceforth, the term bias in this study refers specifically to the occurrence of DIF where items appear invalid for members of a particular group.

The focus of this research is the differential item functioning (DIF). As described above, DIF is part of the internal bias of a test, especially for achievement tests.

This research is more specifically focused on DIF which refers to the differences in the implementation area of the UN. This difference is in accordance with the division of local government in Central Java Province, namely 29 districts and 6 municipalities.

b) Types of DIF

The types of DIF can be divided into two, namely uniform DIF and non-uniform DIF. It is called uniform if along the ability scale, the probability of answering one group correctly is always higher than other groups (Camili and Shepard, 1994). The second type is non-uniform DIF. In the non-uniform, the probability of answering correctly will be higher for one group in a certain ability range, and higher for one group in another ability range.

4 Results and Discussion

4.1 DIF Detection Method

Osterlind (1983) suggested five techniques for detecting the presence or absence of DIF on test items, namely: (1) analysis of variance, (2) transformed item difficulty index, (3) chi-square, (4) distractor response analysis, and (5) item characteristic curves. Camili and Shepard (1994), proposed seven techniques for detecting bias, however, five of them based on classical test theory are not recommended. These five techniques: (1) transformed item difficulty, (2) adjustment of the transformed item difficulty index, (3) golden rule procedures, (4) analysis of variance, and (5) differences in point-biserial correlation. Two other techniques recommended are: (1) item characteristic curves, based on item response theory, and (2) contingency tables. Both Osterlind (1983), and Camili and Shepard (1994) agreed that the characteristic curve technique is the best for detecting DIF. Therefore, only this technique which will be discussed further and used for the evaluation of DIF in this study.

The item characteristic curves technique for detecting DIF is to compare the differences in the item characteristic curves of the two groups studied. The difference in the item characteristic curve between the two groups shows that at the same ability level the test takers from the reference group (A) and the other group (B) do not have the same probability of answering items a, b, and c, therefore the differences in item characteristic curves for two groups can be mathematically represented as the differences in the parameters a, b, or c, or a combination of the three.

The difference in these parameters induces DIF to occur in general categories: first, DIF is consistent or uniform, occurs when the item characteristic curves are different and do not cross each other. This happens because of bothitem characteristic curves have the same parameter a, so, they differ only in parameter b. Second, DIF is inconsistent or non-uniform, occurs when the item characteristic curves are different but intersect at a point on the scale θ. Therefore, the DIF for and to certain groups is balanced or to a certain extent mutually excludes one another. Positive DIFs may completely or partially cancel each other out depending on the pair of the two item characteristic curves (Camili and Shepard, 1994).

The area between the two item characteristic curves gives a notion of the DIF levels. If two item characteristic curves intersect, one part of the area is said to be positive DIF and another part as negative DIF. In such a special case, the two regions between these characteristic curves are considered as positive DIF and they are added together to form the overall index.

4.2 DIF Statistical Hypothesis Testing

The DIF statistical hypothesis testing in this study used the Waldtest method available in the R program. In this approach, the item response theory of one parameter logistic model was chosen, namely the Rasch model. According to Agresti (1996),Waldtest is used to test the significance of each coefficient (β) in the model. The hypothesis in the Wald Test is H0: $\beta = 0$,

which states the chance of success is independent of variable X. The test statistic used in the Wald Test, namely the Z square value follows a chi-square distribution with df = 1. Jika $Z^2 \geq \chi^2_{(1)}$, then reject H_0,, accept in other cases..

In the case of the univariate one-parameter test, the Wald statistic can be definedas$\frac{(\hat{\theta}-\theta_0)^2}{var(\hat{\theta})}$ compared to the Chi-square distribution. Apart from that the differences, it can be also compared with the normal distribution. In this case, the statistical test is ($\frac{\hat{\theta}-\theta_0}{se(\hat{\theta})}$where $se(\hat{\theta})$ is the standard error of the maximum likelihood estimate (MLE). A reasonable estimate of the standard error for MLE can be described by $\frac{1}{\sqrt{I_n(MLE)}}$where I_n is the n information of the parameter. The Wald Test is a parametric statistical test named after Abraham Wald's Transylvanian statistic with a variety of uses. Each time a relationship within or between data items can be expressed as a statistical model with the parameters to be estimated from the samples. The Wald test can be used in a variety of different models including models for dichotomous variables and continuous variables. For example, a researcher has response data from male and female students on the Mathematics National Examination. To find out whether there is a difference in the chance of answering correctly between male and female test takerswith the same ability. Then, the Wald test can be used.

The method for testing DIF as the difference between the item characteristic curves of two groups on a test consisting of a number of items. In this research, it can be described as follows: first, inputting the test takers' response data into notepad format. The data is divided into two, namely upper group data for regencies and lower group data for urban areas. Data for district areas are given number 1 at the end of the row and data for city areas are given number 0 at the end of each row. Then the data was entered into a folder and analyzed with the help of the R program using the Waldtest method in the eRm package. The results of the DIF analysis using the Waldtest method will obtain a z-value consisting of a z-statistic and a p-value. The presence of DIF in each test item can be seen from the opportunity value (p-value) at a significance level of 0.01. The test item is saidto have DIF if the result of opportunity value (p-value) is less than or equal to 0.01 and if the opportunity value (p-value) is more than 0.01 then the item is said to be not DIF.

4.3 DIF Analysis Between District and Urban Areas

There are 22 test items that statistically the chance of answering correctly (p - value) is DIF. 11 items of which benefit participants from the regency area, namely the test items number 3, 4, 9, 12, 14, 17, 21, 28, 30, 33 and 36.This was indicated by the item characteristic curve for the higher district area and the opportunity value of answering correctly which is greater for regency areas when compared to urban areas. This means that with the same abilities, participants from the districts have a greater probability than participants from urban areas to correctly answer the test items 3, 4, 9, 12, 14, 17, 21, 28, 30, 33 and 36.

Eleven other test items benefited participants from urban areas, namely the test items number 7, 8, 10, 11, 16, 19, 27, 29, 35, 37 and 40. This was indicated by a higher item characteristic curve and the probability of answering correctly which is greater for urban areas when compared to districts. This means that with the same abilities, participants from urban areas have a greater probability than participants from regencies to correctly answer the test items 7, 8, 10, 11, 16, 19, 27, 29, 35, 37 and 40.

The occurrence of DIF on an item can become an initialsymptom of bias in an item. However, to determine whether an item is biased or not, it still requires further and more in-

depth research. This is because an item can be caused by the learning process, such as theteachers'abilities, facilities and infrastructure for teaching methods. The implementation of the UN test is also the cause of the occurence of DIF, especially in relation to the supervision of tests and the rules for administering tests. Researchers cannot confirm whether the DIF is caused by the learning process because to utter the cause of DIF requires more in-depth research. But at least this explanation can help in further research on the occurrence of DIF on the test items.

The presence of DIF on an item can also be caused by supervision and test rules which are sometimes applied differently for each region. Less strict supervision will cause test takers to be more freedom to cooperate with other test takers. Since many test takers cooperate with other test takers, it will produce different parameter estimation from the actual item parameter prices. Likewise, if the test rules are not imposed the same in all areas administering the UN. For example, the use of a calculator as a counting tool, although in the test items' scripts are stated not to use a calculator, there are certain areas that allow test takers to use a calculator. This situation is of course detrimental to test takers in other areas which in turn can lead to DIF on certain points.

5 Conclusion

Based on the findings of the research and discussion, several important things can be proposed in connection with the UN Math SMA IPA Package 1103 for the 2014/2015 Academic Year in relation to the instruments of Maths UN in Central Java as follows:

a. The results of the DIF detection analysis using the Rasch model showed that in the set of questions for the Mathematics and Natural Sciences UN Package 1103 for the Academic Year 2014/2015 in Central Java, there are 22 items found containing DIF based on regional differences, namely items 3, 4, 7, 8, 9, 10, 11, 12, 14, 16, 17, 19, 21, 27, 28, 29, 30, 33, 35, 36, 37, and 40.
b. The results of the DIF analysis with the DIF plot using the R program version 3.2.3 and the item characteristic curves using the MAPLE 18 program showed that test takers with the same ability had different chances of answering correctly. This showed that 11 items, namely, (items number 3, 4, 9, 12, 14, 17, 21, 28, 30, 33 and 36) had a greater value and a higher graph for the districts when compared to the urban areas. This showed that items 3, 4, 9, 12, 14, 17, 21, 28, 30, 33 and 36) benefit the district area. On the other hand, the other 11 items, namely, (items 7, 8, 10, 11, 16, 19, 27, 35, 37 and 40) benefit the city area. This showed that 11 items, namely, (items 7, 8, 10, 11, 16, 19, 27, 35, 37 and 40) had bigger values and higher graphs for urban areas when compared to districts.

Implications

The presence of UN Math SMA IPA Package 1103 test items in Central Java for the 2014/2015 academic year which were found to have DIF between test takers from the districts and urban areas, indicates the possibility of a number of items being biased. This item bias will cause the test objectivity requirements to be not met which in turn reduces the validity of a measuring instrument.

References

[1] Ahmann, J.S & Glock, M.D. 1981. *Evaluating Student Progress: Principles of tests and measurements.* Boston: Allyn and Bacon.
[2] Agresti, A. 1996. *An Introduction to Categorical Data Analysis.* USA: John Wiley.
[3] Allen, M.J. & Yen, W.M. 1979. *Introduction to Measurement Theory*. Belmont, California: Wadsworth, Inc.
[4] Angoff, W.H. 1993. *Perspectives on Differential Item Functioning Methodology.* New York: Lawrence Erlbaum Associates, Inc.
[5] Azwar, S. 2007. *Tes Prestasi (edisi ke-11).* Yogyakarta: Pustaka Pelajar.Camili, G. & Shepard, L.A. 1994. *Methods For Identifying Biased Test Items*. Thousand Oaks, California: Sage Publication, Inc.
[6] Cronbach, L.J. 1970. *Essential of psychological testing,* New York: Harper and Row publisher.
[7] Choi, S.W.; Gibbsons, L.E. & Crane, P.K. 2011. "An R Package for Detecting Differential Item Functioning Using Iterative Hybrid Ordinal Logistic Regression/Item Response Theory and Monte Carlo Simulations." *Journal of Education Measurement.* 39(8), 1–30.
[8] Douglas, J.A.; Roussos, L.A & Stout, W. 1996. "Item-Bundle DIF Hypothesis: Identifying Suspect Bundles and Assessing Their Differential Fungtioning." *Journal of Educational Measurement*. 33(4), 465-484.
[9] Depdiknas, 2003. *Kurikulum 2004 standar kompetensi mata pelajaran matematika sekolah menengah atas dan madrasah aliyah.* Jakarta: Depdiknas.
[10] Ebel, Robert L. 1972. *Essentials of educational measurement.* Englewood Cliffs, New Jersey: Prentice Hall Inc.
[11] Gierl, M.J.; Bizans, J.; Bizans, J.; Bizans, G. L., Boughton, & Keith, A. 2003. "identifying Content and Cognitive Skills that Produce Gender Differences in Mathematics: A demonstration of the Multidemensionalty- Based DIF Analysis." *Journal of Educational Measurement*. 40(4), 281-306.
[12] Gronlund, Norman E. 1981. *Measurement and evaluation in teaching.* New York: Macmillan.
[13] Hambleton, R.K. 1989. Principles and selected applications of item response Theory. Dalam R.L. Linn(Ed). *Education measurement hal. 147-200* New York: Macmillan.
[14] Hambleton, R.K. & Swaminathan, H. 1985. *Item response theory*. Boston: Kluwer.
[15] Hambleton, R.K.; Swaminathan, H & Rogers, H.J. 1991. *Fundamentals of Item response theory*. Newbury Park, California: Sage Publications, Inc.
[16] Hullin, C.L., et al. 1983. *Item response theory : Application to psichologycal measurement.* Homewood, IL : Dow Jones-Irwin.
[17] Klieme, E.; Baumert, J & Planck, M. 2001. "Identifying national cultures of mathematics education: Analysis of cognitive demands and differential item functioning in TIMMS." *European Journal of Psychology of Education*. Vol XVI, n° 3, 385-402.
[18] Keeves, J.P. 1992. The *IEA* technical handbook. The haque: the international association for the evaluation of the eucational achievement (IEA).
[19] Mardapi, D. 2008. *Teknik Penyusunan Instrumen Tes dan Non Tes*. Jogjakarta: Mitra Cendikia Press.

[20] Maller, S.J. 2001. "Differential Item Fungtioning in the WISC-III: Item Parameters for Boys and Girls in the National Standardization Sample." *Educational and Psychological Measurement*. 61(5), 793-817.
[21] Mazor, K.M. *et. al.* (1985). Using Logistic Regression and Mantel-Hanzel with Multiple Ability Estimates to Detect DIF. *Journal of educational measurement*. 32(2). 131-144.
[22] Mehrens, W.A & Lehmann, I.J. 1973. *Measurement and evalution in education and psycology.* New York: Holt, Rinehart, and Winston, Inc.
[23] NYSDE, 2003. *New York State Testing Program Mathematics Grade 8 Technical Report 2002*. New York: New York state Departement of Educational McGraw-Hill.
[24] Naga, D.S. 1992. *Pengantar Teori Skor Pada Pengukuran Pendidikan*. Jakarta: Gunadarma.
[25] Osterlind, Steven J. 1983. *Test Item Bias*. Beverly Hills, CA: Stage Publication.
[26] Peraturan Menteri Pendidikan dan Kebudayaan Nomor 5 Tahun 2015 Tentang Kriteria Kelulusan Peserta Didik, Penyelenggaraan Ujian Nasional, dan Penyelenggaraan Ujian Sekolah/Madrasah/Pendidikan Kesetaraan Pada SMP/MTs atau yang Sederajat dan SMA/MA/SMK atau yang Sederajat. 2015.
[27] Retnawati, H. 2014. *Teori Respons Butir dan Penerapannya*. Yogyakarta: Nuha Medika.
[28] Retnawati, H. 2013. "Pendeteksian Keberfungsian Butir Pembeda dengan Indeks Volume Sederhana Berdasarkan Teori Respons Butir Multidimensi." *Jurnal Penelitian dan Evaluasi Pendidikan*. Volume 17 No. 2, 275-286.
[29] Retnawati, H. 2003. "Keberfungsian Butir Diferensial Pada Perangkat Tes Seleksi Masuk SLTP Mata Pelajaran Matematika." *Jurnal Penelitian dan Evaluasi.* No. 6, Tahun V, 45-58.
[30] Ridho, A. 2009. *Bias Gender Dalam Tes.* UIN-Malang Press: Malang.
[31] Snetzler, S. & Qualls, A. 2000. "Examination of Differential Item Functioning on Standardized Achievement Battery With Limited English Proficient Student." *Educational and Psycological Measurement*. 60(4), 564-577.
[32] Stone, C.A. 2003. "Empirical Power and Type I Rates for An IRT Fit Statistic That Considers The Presition of Ability Estimates." *Educational and Psychological Measurement*. 63(4), 566-583.
[33] Sudaryono. 2013. *Teori Responsi Butir.* Yogyakarta: Graha Ilmu.
[34] Susongko, P. 2000. "Keberfungsian Butir Deferensial Pada Perangkat Tes EBTANAS Kimia Sekolah Menengah Umum di Jawa Tengah". *Tesis*. Yogyakarta: Progaram Pasca Sarjana Universitas Negeri Yogyakarta.
[35] Worthen, B.R & Sanders, J. R. 1973. *Educational evalution; Theory and Practice.* Belmont, California: Wadswort Publishing Company, Inc.
[36] Zenisky, A.L.; Hambleton, R.K.; & Robin, F. 2003. "Detection of Differential Item Functioning in Large-Scale state Assessment: A study Evaluating a two Stage Aproach. "*Educational and Psychological Measurement*. 63(1), 51-64.