http://www.ijate.net
e-ISSN: 2148-7456 © IJATE
INTERNATIONAL JOURNAL OF ASSESSMENT TOOLS IN EDUCATION

Research Article

The Dif Identification in Constructed Response Items Using Partial Credit Model

Heri Retnawati*[1]

[1]Mathematics and Science Faculty, Universitas Negeri Yogyakarta, Yogyakarta, Indonesia

Abstract: The study was to identify the load, the type and the significance of differential item functioning (DIF) in constructed response item using the partial credit model (PCM). The data in the study were the students' instruments and the students' responses toward the PISA-like test items that had been completed by 386 ninth grade students and 460 tenth grade students who had been about 15 years old in the Province of Yogyakarta Special Region in Indonesia. The analysis toward the item characteristics through the student categorization based on their class was conducted toward the PCM using CONQUEST software. Furthermore, by applying these items characteristics, the researcher draw the category response function (CRF) graphic in order to identify whether the type of DIF content had been in uniform or non-uniform. The significance of DIF was identified by comparing the discrepancy between the difficulty level parameter and the error in the CONQUEST output results. The results of the analysis showed that from 18 items that had been analyzed there were 4 items which had not been identified load DIF, there were 5 items that had been identified containing DIF but not statistically significant and there were 9 items that had been identified containing DIF significantly. The causes of items containing DIF were discussed.

ARTICLE HISTORY

Received: 08 August *2017*
Revised: 23 October 2017
Accepted: 26 October 2017

KEYWORDS

DIF,
polytomous data,
partial credit model,

1. INTRODUCTION

In performing a measurement, there should be utilized valid and reliable instruments. By utilizing instruments that satisfy the both criteria, the measurement results will describe the aspects that should be measured without being influenced by other factors or other loads that should not be measured. An instrument that has been influenced by the other factors other that should be measured certainly contains an error. If the error caused the significance of performance of testees from many groups, it called with bias (Ogbebor & Onuka, 2013).

The bias of a test and a measurement refers to a not good condition, it has unfair meaning, gives to much pressure or becomes too fanatic toward the object under measurement (Osterlind, 1983). The bias within a test has been an unfair and inconsistent condition that has been

*Corresponding Author Phone: +628122774435, E-mail: heri_retnawati@uny.ac.id

ISSN: 2148-7456 online /© 2018
DOI: 10.21449/ijate.347956

contaminated by the factors outside the aspects under the test and by the errors in the test application. This matter shows that the bias within a test and a measurement does not support the characteristics of a valid and consistent test.

Several researchers provide their limitations regarding the item bias, namely Osterlind (1983), Shepard (Adams, 1992), Mazor et al. (1995), Budiono (2004), and Retnawati (2013). A test will be considered biased if two test participants under the same ability from two different groups do not have the same probability to get a correct response. Therefore, the unbiased test items are the ones that have been expected to provide the same probability of providing the correct response among the test participants under the same ability from two different groups (Adams, 1992; Mazor et al. (1995). There are two types of bias namely the external bias and the internal bias.

According to Osterlind (1983), the external bias has been a degree in the test score which shows the correlational relationship of independent variables within a test or an instrument. Furthermore, he states that the problem of the external bias is the social consequence within the test implementation such as the fairness in the test administration and the criteria that might be applied. In relation to this matter, the test administrator has the right to execute the test and to design the criteria that will be related to the fair decisions within the test. Therefore, the aspect that should be given attention within the external bias is the test in overall (the construct validity and predictive validity).

Adams (1992) states that the internal bias which is also known as the item bias refers to the bias within a test that has been related to the psychometric characteristics of a test item and a test in overall. The procedures of detecting the biased items are focused mainly on the investigation whether each test time has similar behaviors or not, namely the similarity in the measurement of psychometric characteristics. According to Osterlind (1983), a test will be considered biased if there is evidence from the interaction between the group members and the test performance in which the different ability or psychological condition among these groups is controlled.

Several psychometric experts have taken the steps to eliminate the lowering connotation in relation to the item bias (Holland & Thayer, 1988; Plake, Patience, & Whitney, 1988). The term that has been used in order to replace the item bias is the differential item performance (DIP) or the differential item functioning (DIF) (Adams, 1992). The new term reflectes the objective of the bias detection method in identifying the items that have different functions for different test participant groups such as the ones that have different facility, different region, different sex and alike.

Based on the results of international studies such as Programme for International Student Assessment (PISA), people can attain information that the literacy scores of Indonesian students has not been satisfying as expected. PISA measures the literacy proficiency that includes the science literacy and the mathematics literacy. These results show that within the conduct of PISA international study the Indonesian students' literacy scores has been far below the international mean (OECD, 2013). Such unsatisfying results might be explored further in relation to the development of the Indonesian students' literacy. Taking a close attention to the test that has been administered by PISA, the respondents of the test are about 15 years old students. These students are both the ones in the ninth grade or in the third grade of junior high school and the ones in tenth grade or the first grade of senior and vocational high school.

The ninth grade students are certainly different than the tenth grade students. The tenth grade students have been provided with the additional materials within the schools, the families and the society for one whole year. These additional materials should be investigated further in order to identify whether they provide additional literacy knowledge or not. In other words, whether there has been any DIF load or any different probability of providing the correct response toward the test items between the ninth grade students and the tenth grade students or not should be identified. Therefore, this study is to identify the load, the type and the significance of the differential item functioning (DIF) within the partial credit model (PCM) polytomous data. The data that will be manipulated in the study are the students' instrument and the students' response toward the PISA-like test items.

There are several methods that might be applied in order to identify the DIF load within the test items. These methods are classified based on the approach of their underlying theories, namely the classical test theory and the item response theory. In the approach of classical test theory, the methods that have been frequently applied are SIBTES, regression, Mantel-Haenszel (Budiono, 2004), mean covarians (Elosua and Wells, 2013), Lagrange multiplier (Khalid & Glass, 2013) and HGLM (Acara, 2011). Adams (1992) states that the methods that might be applied in order to detect the DIF are factor analysis, item discriminative index by means of point-biserial and partial correlation, item discriminative level test by means of multiple transformations, ANOVA, item response theory or latent trait, chi-square, log-linear model and Mantel-Haenszel statistical theory.

According to Bulut and Suh (2017), there are several methods that might be applied in order to detect the DIF both by means of parametric statistics and of nonparametric statistics. If one would like to apply the parametric statistics methods, then he or she might apply the Chi-Square by Lord, the Likelihood Ratio Test and the Signed and Unsigned Area Methods (Thissen, et al., 1993). On the other hand, if one would like to apply the nonparametric statistics methods then he or she might apply the SIBSTEST or the Mantel-Haenszel methods. The two statements are supported by Retnawati (2003) who performed a DIF analysis using chi-square by Lord and maximum likelihood ratio-test. The methods of both the parametric and the nonparametric statistics might only be applied on a test that measures only one ability (unidimension) and not multiple abilities (multidimension). The existing methods of DIF detection are only found in the unidimension item response theory on the dichotomous score (Camili and Shepard, 1994), the multidimension item response theory on the dichotomous score (Kartowagiran & Retnawati, 2008; Retnawati, 2013) and the likelihood maximum ratio-test (Wang, Yeh, & Yi, 2003).

In the methods of DIF detection by means of item response theory, the DIF is defined as the different probability of providing correct response between two groups that have similar ability. In order to identify the probability difference, the probability of test participants' ability should be identified first. This probability might be identified based on the item parameter, which is adjusted to the scoring type. The test participants' response toward the polytomous scoring-type test items might be analyzed by applying the partial credit model (PCM)-type unidimensional item response theory. At the beginning of the polytomous item response theory development, this model is known more as the expansion of the Rasch model which has been regarded as Partial Credit Model (PCM). The PCM is a polytomous scoring model that has been the expansion of Rasch model in the dichotomous data.

According to Muraki and Bock (1997), the general form of PCM is as follows:

$$P_{jk}(\theta)=\frac{\exp\sum_{v=0}^{k}(\theta-b_{jv})}{\sum_{h=0}^{m}\exp\sum_{v=0}^{k}(\theta-b_{jv})}\text{, k=0,1,2,...,m} \qquad (1)$$

with:

$P_{jk}(\theta)$ = the probability of θ ability test participants in attaining the *k* score category within the *j* item

θ = test participants' ability

m+1 = the number of *j* item category

b_{jk} = the *k* category difficulty index in the *j* item

$$\sum_{h=0}^{k}(\theta - b_{jh}) \equiv 0 \text{ and } \sum_{h=0}^{h}(\theta - b_{jh}) \equiv \sum_{h=1}^{h}(\theta - b_{jh}) \tag{2}$$

The score of category in the PCM displays the number of the steps that might be taken in order to complete the related test item correctly. The higher score of category resembles the greater ability than that of the lower score of category. In the PCM, if a test item has two categories then the second equation will be the Rasch model equation, like the one that has been proposed by Hambleton and Swaminathan (1985) and that has been supported by Hambleton, Swaminathan and Roger (1991). As a consequence, the PCM might also be implemented toward the polytomous and the dichotomous test items.

In the Rasch model, one of the most famous software for analysis is the QUEST or the CONQUEST by ACER. There are slight differences on the parameter symbols that should be operated. The location parameter between the two software is δ_{ij} instead of *b*. In order to easily understand the related equation and the interpretation of analysis results, the researcher would like to display a mathematical model along the item characteristic curve that is also known as the category response function (CRF).

In order to estimate the parameter along with the *n* test participants (case/person) and the *i* test item with the θ ability and the location parameter of *j* category in the *i* test item that has been equal to δ_{ij} for the 0, 1 and 2 score category, the researcher formulates the following equation (Masters, 2010):

$$P_{ni0} = \frac{1}{\Psi}$$

$$P_{ni1} = \frac{\exp(\theta_n - \delta_{i1})}{\Psi}$$

$$P_{ni2} = \frac{\exp(2\theta_n - \delta_{i1} - \delta_{i2})}{\Psi} \tag{3}$$

Or in general the above equation will be stated as

$$P_{nik} = \frac{\exp(k\theta_n - \delta_{i1} - \delta_{i2} - \cdots - \delta_{ik})}{\Psi} \tag{4}$$

with Ψ as the numerator amount of the overall category.

In the analysis parameter esstimation using a certain software, for example CONQUEST, the δ parameter will be decomposed into the difficulty level parameter and the step parameter. In the 3-category scoring type toward a test item, there will be 2 step parameters and 1 item difficulty parameter. For example, $\delta_{ik} = b_i + \tau_k$ with *b* as the *i* item difficulty parameter and τ as the k step parameter. The probability of each step will be presented as follows.

$$P_{ni0} = \frac{1}{\Psi}$$

$$P_{ni1} = \frac{\exp(\theta_n - b_i + \tau_1 \ \ )}{\Psi}$$

$$P_{ni2} = \frac{\exp(2\theta_n - 2b_i + \tau_1 + \tau_2)}{\Psi}$$

$$\Psi = 1 + \exp(\theta_n - b_i + \tau_1) + \exp(2\theta_n - 2b_i + \tau_1 + \tau_2) \qquad (5)$$

The two groups that respond to the test item which has been identified as DIF will be regarded as the focal group and the reference group. The DIF index states the difference of signed area that displays the total probability of providing the correct response in each group. Camilli and Shepard (1994) named this method as Simple Area Indices. Within the test items that have uniform DIF, the DIF index might be identified by:

$$\text{SIGNED-AREA} = \int [P_R(\theta) - P_F(\theta)] d\theta \qquad (6)$$

and for the test items that have non-uniform DIF, the DIF index might be identified by:

$$\text{UNSIGNED-AREA} = \sqrt{\int [P_R(\theta) - P_F(\theta)]^2} \ d\theta \qquad (7)$$

By applying the concept of different probability in providing the correct response between the reference group and the focal group, this concept might be applied toward the function of the probability in providing the correct response in the polytomous data. This function is implemented in order to estimate the DIF index that has been developed by Retnawati (2014) by drawing the characteristic curve first. In the test items of polytomous-type test participants' responses that involve two categories, the characteristic curve might be seen in Figure 1.

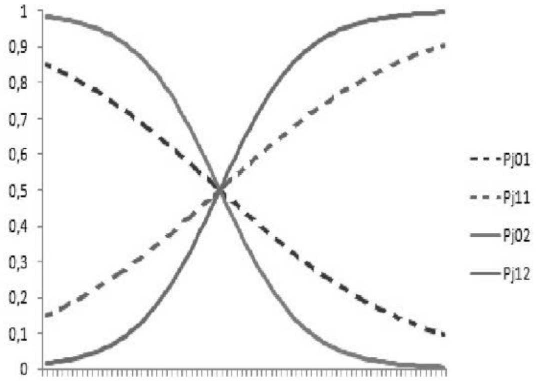

Figure 1.a. The item characteristic curve for the focal (1) a = 0.5 and b = -0.5 and the reference (2) a = 1.2 and b -.05 with 2 categories

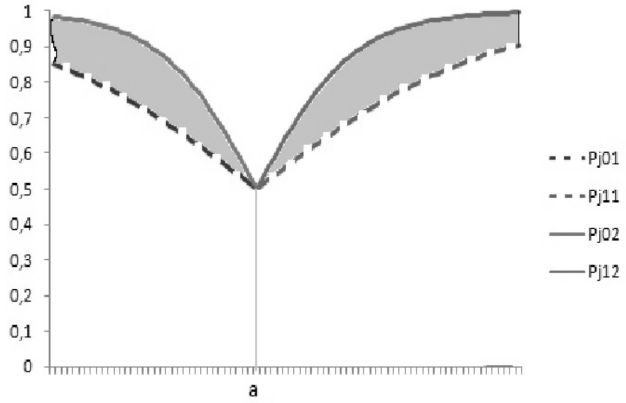

Figure 1.b. The item characteristic curve for calculating the uniform DIF index in PCM with 2 categories

The area between the two characteristic curves is named as the SIGNED AREA, which size might be calculated mathematically by means of integration method. The coverage of this area is the DIF index, which has been drawn in the Figure 1.b. Because in certain points, namely

θ =a, the curve P_{j02} and P_{j12} as well as P_{j01} and P_{j11} are intersecting to each other, the integral equation for the signed area will be:

$$\text{SIGNED-AREA} = \int_{-\sim}^{a}(P_{j02})d\theta + \int_{a}^{c}(P_{j12})d\theta + \int_{c}^{+\sim}(P_{j22})d\theta - \int_{-\sim}^{a}(P_{j01})d\theta - \int_{a}^{b}(P_{j11})d\theta - \int_{b}^{c}(P_{j21})d\theta \qquad (8)$$

Similar situation also applies in the 3-category polytomous data that are displayed in Figure 2.a and Figure 2.b. For example, the item parameters of the focal group a = 0.5 are and b_1 = -2.0 and b_2 = 1.0, while the item parameters of the reference group are a = 1.0 and b_1 = 2.0 and b_2 = 1.1. After the item characteristics have been described with the characteristic curve, it is apparent that these items contain the uniform DIF. The coverage of the signed area is formulated through the following equation:

$$\text{SIGNED-AREA} = \int_{-\sim}^{a}(P_{j02})d\theta + \int_{a}^{c}(P_{j12})d\theta + \int_{c}^{+\sim}(P_{j22})d\theta - \int_{-\sim}^{a}(P_{j01})d\theta - \int_{a}^{b}(P_{j11})d\theta - \int_{b}^{c}(P_{j21})d\theta \qquad (9)$$

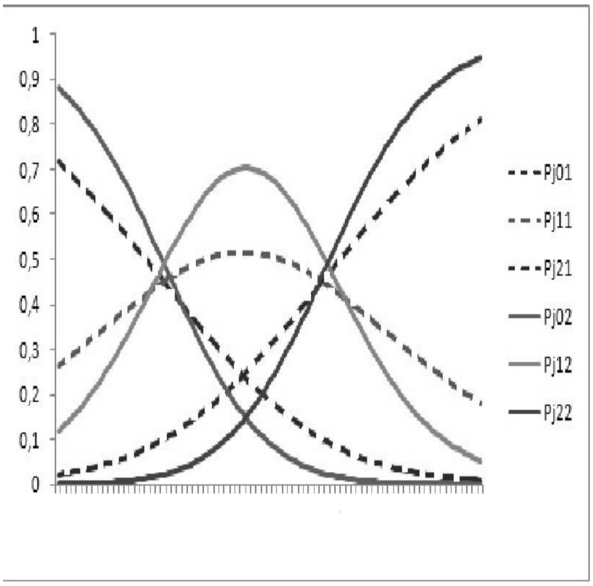

Figure 2.a. The characteristic curve for the focal group (1) and the focal group (2) with 3 categories

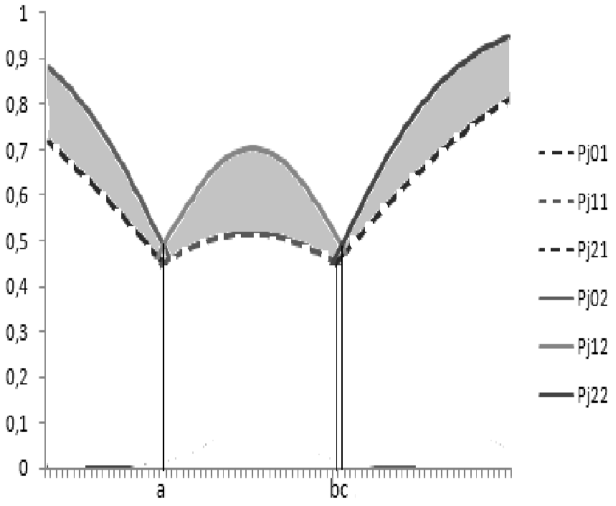

Figure 2.b. The item characteristic curve for calculating the uniform DIF index with 3 categories

In the test items that have non-uniform DIF loads, the DIF index might be identified by paying attention first to the characteristic curve in order to see the integral area. Then, the integral area should be used in calculating the probability coverage. An example of this situation will be provided in Figures 3 and 4.

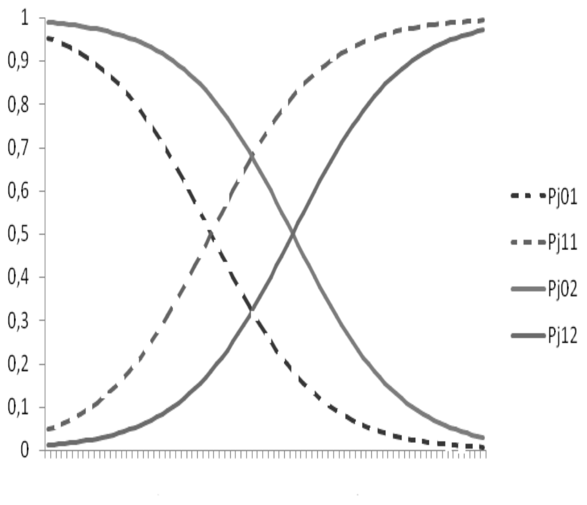

Figure 3.a. The CRF with 2 categories (containing non-uniform DIF loads)

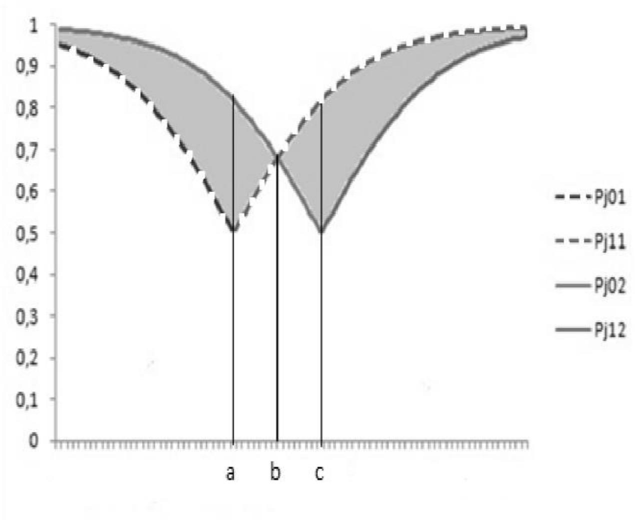

Figure 3.b. Part of the CRF that might be used in calculating the integral of non-uniform DIF loads index in 2 categories

$$\text{UNSIGNED-AREA} = \int_{-\sim}^{a}(P_{j01} - P_{j02})d\theta + \int_{a}^{b}(P_{j01} - P_{j12})d\theta + \int_{b}^{c}(P_{j11} - P_{j02})d\theta + \int_{c}^{+\sim}(P_{j11} - P_{j12})d\theta \qquad (10)$$

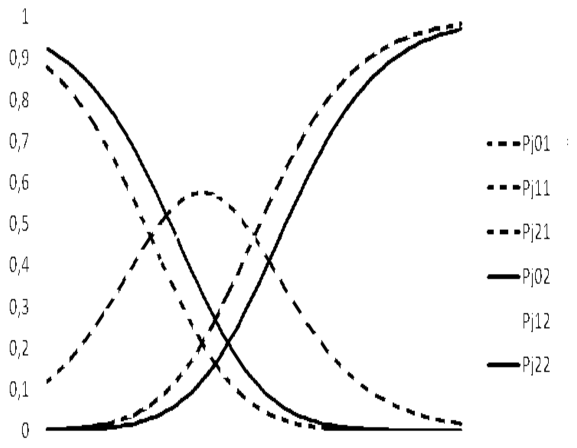

Figure 4.a. The CRF with 3 categories (containing non-uniform DIF loads)

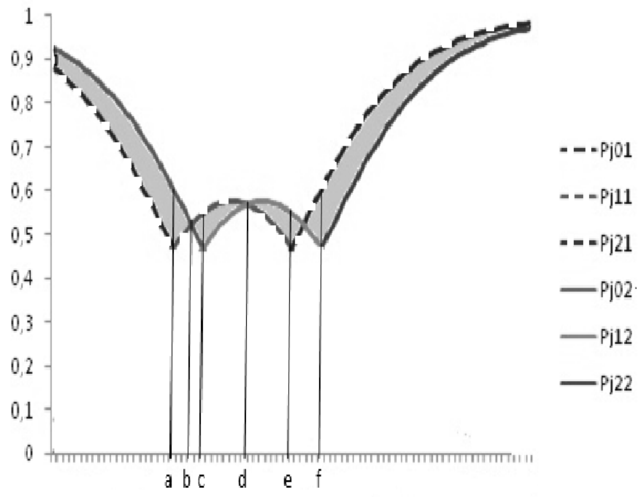

Figure 4.b. Part of the CRF that might be used in calculating the integral of non-uniform DIF loads index in 3 categories

If the function is considered too complicated, the calculation of this integral might be conducted through the Rieman sum calculation assistance by turning the integral area into small area (Varberg & Purchell, 2001) and then calculating these areas by means of numeric approach.

2. METHOD

The study was a descriptive explorative research that identified the DIF loads in the polytomous scoring-type PISA-like test items. The approach in the study was the quantitative one. The study not only identify the load of DIF, but also identify the type and the significance of differential item functioning (DIF) in the partial credit model (PCM) polytomous data.

2.1. Data Collection Method

The data collection of the study utilized test. The test was the PISA-like test instrument that had been developed by Wulandari (Jailani, et al, 2015). The test instrument were developed by adopting the PISA released items from 3 periods (2003, 2007 and 2011); the number of the items was 21 units. The 4 test items had been the constructed responsewith dichotomous scoring (0-1) and 17 test items had been the constructed response with 3 category polytomous scoring (0-1-2).The test contained domain of context (that included the personal context, the societal context, the occupational context and the scientific context) and the domain of process (that included formulate, employ and interpret). The PISA-like test were in *bahasa Indonesia* and utilizing Indonesian contexts.

2.2. The Participants

The test participants of the study are 386 ninth grade students (third grade students of junior high school) and 460 tenth grade students (first grade students of senior high school) whose age were about 15-16 years old. The completion of these items involved the students from 4 regencies and 1 municipality in the Province of Yogyakarta Special Region in Indonesia and these students came from both the state schools and the private schools; the category of these schools are high, moderate and low based on the results of their achievement in the National Examination. The ninth grade students belonged to the focal group, while the tenth grade students belonged to the reference group.

2.3. Data Analysis

The item characteristic analysis utilizing classroom-based student categorization was conducted through the PCM by applying the CONQUEST software (Wu, Adam, and Wilson, 1997). Then, by applying the item characteristics, the researcher draw the category response function (CRF) graphic in order to compare the discrepancy between the item difficulty level and the item error.

The detailed steps in performing the analysis would be given as follows:

1) Estimating the item parameter by means of Rasch model both for the dichotomous data and the polytomous data with the CONQUEST assistance
2) Selecting the fit items by implementing the Rasch model
3) Estimating the item parameters for the ninth grade students' responses and the tenth grade students' responses in the polytomous and the dichotomous data with the CONQUEST assistance
4) Drawing the CRF with the assistance of EXCEL software in order to identify whether the items had been neutral, containing uniform DIF loads or containing non-uniform DIF loads
5) Calculating the DIF index using Rieman sum technique.
6) Determining the DIF significance by comparing the different estimation of item difficulty level parameters and the two group-estimation error with the assistance of CONQUEST program, using criterion an item contains DIF significantly if the discrepancy of the difficulty index is more than twice of its standard error (Adams & Wu, 2010).

7) Interpreting the results of the analysis, including identifying the reasons why the items had been difficult for the students, comparing the substance of the test items and comparing the position of these materials in the curriculum contain within the schools.

3. FINDINGS

The characteristics of the test item instruments were in the form of difficulty level, step parameter and model fitness. The results of the analysis would be displayed in the Table 1.

Table 1. The Overall Item Characteristics and Model Fitness

Item	Category	Difficulty Level	Step 1 Parameter	Step 2 Parameter	MNSQ	Model Fitness
CR113	2	-2.093			1.02	Fit
CR117	2	-1.676			0.91	Fit
CR119	2	-2.275			0.94	Fit
CR127	2	2.092			1.04	Fit
CR203	3	-1.099	0.369	-0.369	1.06	Fit
CR204	3	-0.694	-0.083	0.083	1.02	Fit
CR207	3	1.084	2.702	-2.702	0.55	Fit
CR212	3	-0.074	1.705	-1.705	0.93	Fit
CR214	3	-2.891	1.097	-1.097	0.96	Fit
CR215	3	0.224	0.680	-0.680	1.16	Fit
CR216	3	0.105	0.861	-0.861	0.92	Fit
CR220	3	0.762	-0.675	0.675	0.94	Fit
CR221	3	-0.867	0.799	-0.799	1.19	Fit
CR222	3	-0.948	0.297	-0.297	0.95	Fit
CR223	3	-0.091	0.523	-0.523	1.00	Fit
CR224	3	0.822	1.576	-1.576	0.61	Fit
CR225	3	0.096	-0.901	0.901	0.95	Fit
CR226	3	0.523	-1.205	1.205	1.16	Fit
CR228	3	3.513			0.56	Fit
CR229	3	2.064			0.86	Fit
CR230	3	1.421			0.90	Fit

Based on the results that had been displayed in the Table 1, all items were compatible to the Rasch model. There was a tendency that the items that had 2 scoring categories or more would be easier to compare than those that had polytomous scoring categories. In the last 3 items that are CR228, CR229, CR230 the category parameters did not appear in the analysis results; instead, the difficulty level parameters appeared in the analysis results. The reason was that these items had been responded only by some of the test participants. For the item CR228, only 7.41% of testees got 1 score and none got 2 score. For the item CR230, only 25.53% of testee got 1 score and only 4.26% got 2 score. Then, the three items were excluded from the analysis results.

Furthermore, the researcher estimated the parameters of each item both for the ninth grade students and the tenth grade students. The complete results of the estimation would be displayed in the Table 2. Based on the results that had been displayed in the Table 2, the researcher found that there had been different parameters between the ninth grade students and the tenth grade students. Although the difference was not prominent, both groups seemed to have different characteristics.

Table 2. The Test Item Parameters that had been Estimated Separately Based on the Data of the Ninth Grade Students and the Tenth Grade Students

Item	Category	Ninth Grade			Tenth Grade		
		Level Difficulty	Step 1 Parameter	Step 2 Parameter	Level Difficulty	Step 1 Parameter	Step 2 Parameter
CR113	2	0.023			-0.023		
CR117	2	0.230			-0.230		
CR119	2	0.847			-0.847		
CR127	2	-0.364			0.364		
CR203	3	0.194	0.606	-0.606	-0.194	0.364	-0.364
CR204	3	0.155	0.186	-0.186	-0.155	-0.087	0.087
CR207	3	0.275	1.017	-1.017	-0.275	2.698	-2.698
CR212	3	-0.124	1.374	-1.374	0.124	1.701	-1.701
CR214	3	0.670	1.310	-1.310	-0.670	1.089	-1.089
CR215	3	0.068	0.817	-0.817	-0.068	0.676	-0.676
CR216	3	0.009	0.798	-0.798	-0.009	0.857	-0.857
CR220	3	-0.040	-0.951	0.951	0.040	-0.679	0.679
CR221	3	0.531	0.644	-0.644	-0.531	0.796	-0.796
CR222	3	0.018	0.040	-0.040	-0.018	0.294	-0.294
CR223	3	0.115	0.269	-0.269	-0.115	0.520	-0.520
CR224	3	0.308	-0.339	0.339	-0.308	1.574	-1.574
CR225	3	-0.084	-0.881	0.881	0.084	-0.905	0.905
CR226	3	0.143	-0.522	0.522	-0.143	-1.208	1.208

Utilizing the item parameters in the Table 2, the researcher might describe the category response function for each item and the researcher might identify whether the DIF loads of an item had been identified or not. Based on the CRF description, the researcher might identify as well whether an item had been beneficial for the ninth grade students or for the tenth grade students. An example of CRF description for the DIF analysis toward several items would be displayed in the Figure 1 until Figure 4.

Also by using the item parameters, the researcher might identify the DIF index by means of integral that had been approached by Rieman sum calculation. The significance of DIF loads might be identified from the comparison between the item parameters discrepancy and the twice of its standard errors that had been calculated by means of CONQUEST. The results of CRF description and the table of DIF identification toward the overall items would be displayed in the Table 3.

Table 3. The Results of DIF Significance Test

Item	Category	Identification of DIF Load Based on the CRF	Type of DIF	DIF Index	Discrepancy on the Difficulty Index	Two-Folded Standard Errors	Significance of DIF Load
CR113	2	Not Loading	-	-	0.046	0.236	-
CR117	2	Loading	Non-Uniform	0.444	0.460	0.246	Significant
CR119	2	Loading	Non-Uniform	1.673	1.694	0.262	Significant
CR127	2	Loading	Non-Uniform	0.703	-0.728	0.964	Not Significant
CR203	3	Loading	Non-Uniform	0.172	0.388	0.172	Significant
CR204	3	Loading	Non-Uniform	0.093	0.310	0.180	Significant
CR207	3	Loading	Non-Uniform	3.081	0.550	0.272	Significant
CR212	3	Loading	Non-Uniform	0.342	-0.248	0.174	Not Significant
CR214	3	Loading	Non-Uniform	0.911	1.340	0.218	Significant
CR215	3	Not Loading	-	-	0.136	0.186	-
CR216	3	Not Loading	-	-	0.018	0.182	-
CR220	3	Loading	Non-Uniform	0.161	-0.080	0.274	Not Significant
CR221	3	Loading	Non-Uniform	0.875	1.062	0.242	Significant
CR222	3	Loading	Non-Uniform	0.250	0.036	0.206	Not Significant
CR223	3	Loading	Non-Uniform	0.554	0.230	0.224	Significant
CR224	3	Loading	Non-Uniform	2.722	0.616	0.460	Significant
CR225	3	Not Loading	-	-	-0.168	0.226	-
CR226	3	Loading	Non-Uniform	0.216	0.286	0.330	Not Significant

From 21 items that had been analyzed, 3 items were excluded from the DIF analysis; as a result, there were 18 items which had been tested. From the overall items and based on the characteristic curve, the researcher attained information that all items had been identified to have the non-uniform DIF loads. From the 18 items, there were 4 items which had not been identified as DIF, there were 5 items that had been identified containing DIF but not statistically significant and there were 9 items that had been identified containing DIF significantly.

Utilizing items paramaters from Table 2, item characteristic curve can be drawn. From its ICC, researcher got information about nature of items, in every category. The categories gave information, wether the step item favored a group of testees. In Figure 5, 6 and 7 explaine the three items with different cases.

The item with the code CR117 had been a test item with a food context that the students commonly read, namely *martabak*. This item had two stimuli namely two types of *martabak*; in the test item, there were two *martabak* with different circular shape and different price but they had the same thickness. These *martabak* would be smeared with the combination of two jam layers and the students, then, were asked to define the amount of the combination.

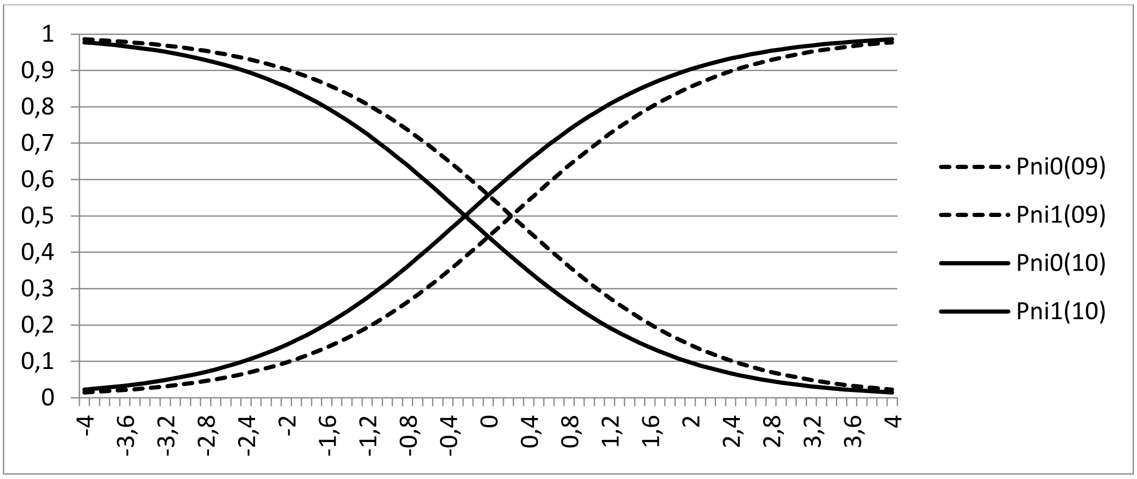

Figure 5. The Graphic of Category Response Function for the Item CR117

Although probability had been studied in the eighth grade, this item demanded specific understanding through the provision of narrative test item. In the item CR117, the tenth grade students had greater chance to score 1 in comparison to the ninth grade students. The reason was that such test items had usually been exercised when the students would attend the national examination; therefore, the tenth grade students, since they used to attend the national examination, would have higher probability in scoring than the ninth grade students.

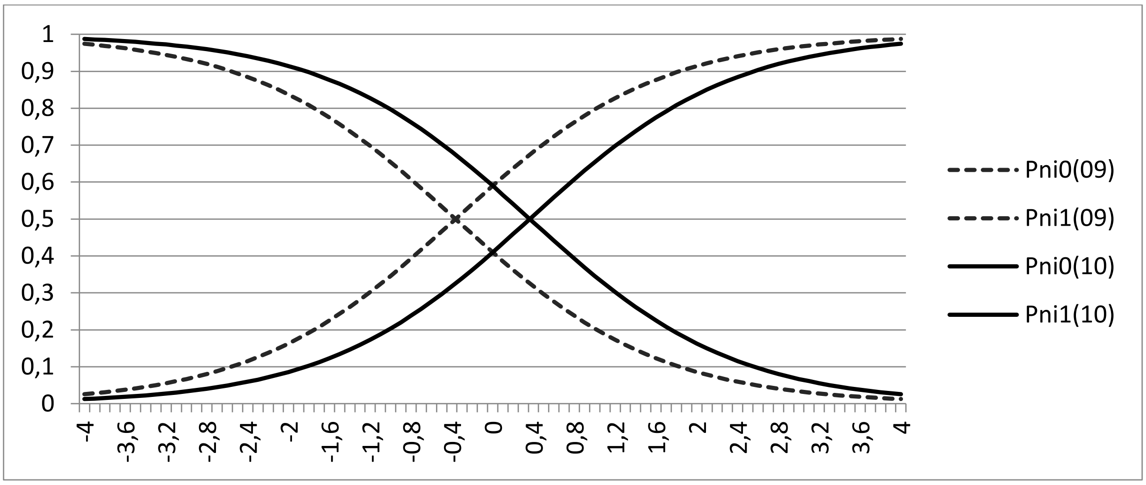

Figure 6. The Graphic of Category Response Function for the item CR127

The item CR127 contained a context where a telecommunication company would like to build a transmitter tower. In this test item, the students were provided with a stimulus of tower construction and of government advice with regards to the construction. Through the concept of distance, the students were asked to provide a reason why the government advice had not been compatible to the regulations of tower construction. The CRF graphic was displayed in the Figure 6. In this item as well, the probability to score 1 among the ninth grade students was higher than that among the tenth grade students. The reason was that the concept of distance had been an easy concept and had been studied much when these students are in the seventh grade. As a result, the ninth grade students had greater probability to memorize this concept than the tenth grade students. It caused the DIF index of the items is equate big, but it is not significantly contain DIF.

The item C212 was beneficial for the tenth grade students both for scoring 1 and scoring 2. This item was related to the materials of probability that had been used in selecting the soccer players who would take on the penalty shootout and who would have a great probability to be the top scorer. Paying attention to the curriculum that had been applied in the schools, this material was studied by the ninth grade students in their final period. It was the reason why the tenth grade students had higher probability to provide the correct response in order to score 1 or 2. The complete CRF graphic for this item would be displayed in the Figure 7.

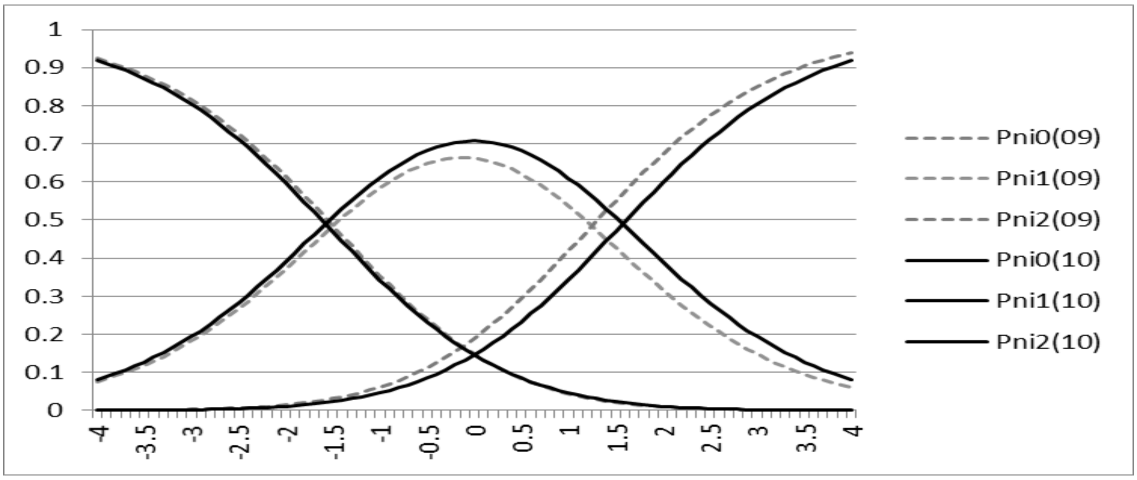

Figure 7. The Graphic of Category Response Function for the Item CR212

A quite different matter was found in the item CR212, which also occurred in the item CR221. The item CR221 had the score 1 category and the tenth grade students had higher probability to score 1 than the ninth grade students. However, in the score 2 category both the ninth grade students and the tenth grade students had the same probability. The reason was that the material in this item had been related to the context of changing the mean values when the test data changed; this material was studied by the ninth grade students in their final period. The CRF graphic for this item would be displayed in the Figure 8.

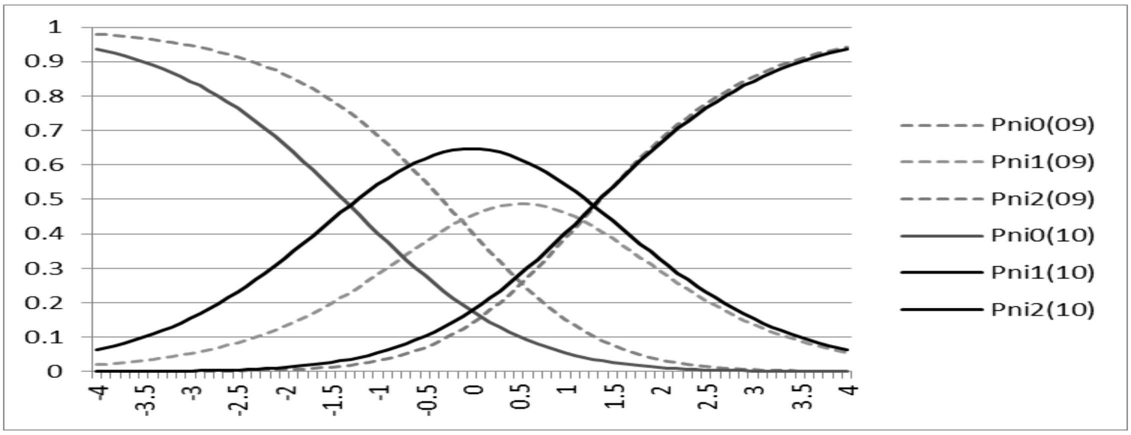

Figure 8. The Graphic of Category Response Function for the Item CR221

The item CR225 had been one of the items that did not have DIF loads. This item had been an item that contained the context of constructing fence in such a way that its circumference would be equal to the length of the wood that the owner had. In order to complete this test item, the students should use their knowledge regarding the concept of determining the circumference of all planes. This material was studied in the elementary school and was deepened in the seventh grade. Such situation was the reason why the item CR225 did not have any DIF loads.

The item CR225 was also a quite unique item. In the score 1 category curve, the score of maximum probability was lower than the probability score in the intersection of 0 score category and 2 score category. This situation indicated that in this item there had been few students who scored 1 and, as a result, this item might be simplified from 3 answer categories into 2 answer categories. The CRF graphic would be displayed in the Figure 9.

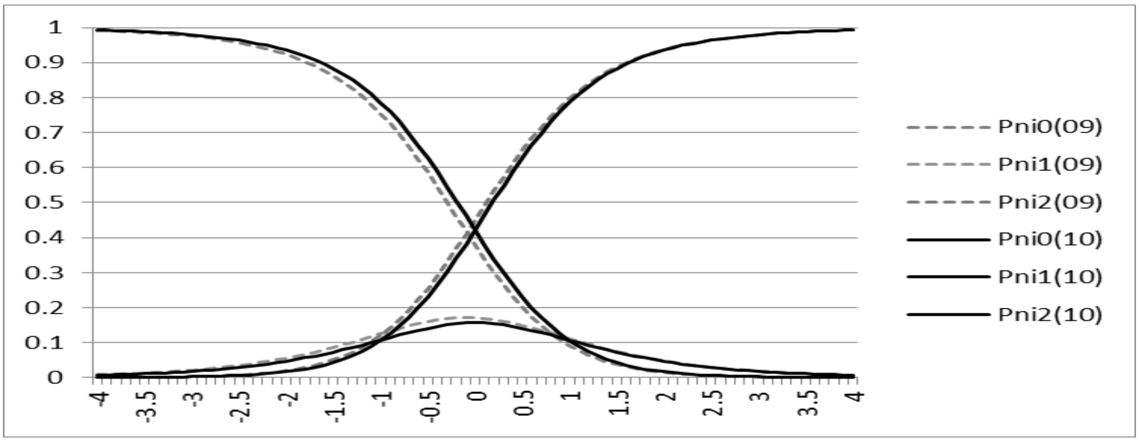

Figure 9. The Graphic of Item Response Category for the Item CR225

The results of DIF significance test in the Table 3 should be given attention as well. By benefitting the estimation resulted-item parameters and the Rieman sum calculation, the researcher attained the DIF index. After the index had been attained, the DIF load significance test was conducted by comparing the discrepancy between the item difficulty level and the parameter estimation errors of the two-group. It turned out that testing the significance through this manner had not been consistent. There were the items which DIF index had been huge but they did not significantly had the DIF loads. On the other hand, there were the items which DIF index had not been huge but they significantly had the DIF loads. In relation to this situation, there should be another study that should pay attention to the comparison in the methods of DIF load identification by using the polytomous data.

Observing each item containing DIF, the most of items contain DIF favoring students aged about 15 years who were in Grade 10, and not favoring students who are about 15 years old but was in grade 9. Based on these results, it can be described the reason why the same age but different classes have the different probability to answer items of PISA-like rightly. The recapitulation of the content and step of items load DIF significantly were showed in Table 4.

Table 4. Recapitulation of content and steps of items load DIF significantly

Item	Content	Step Favore testees from class	
		1	2
CR117	Uncertainty	10	-
CR119	Statistics and Data	10	-
CR203	Geometry	-	10
CR204	Geometry	-	10
CR207	Statistics and Data	10	9
CR214	Uncertainty	10	10
CR221	Statistics and Data	10	-
CR223	Arithmatica	10	-
CR224	Geometry	10	-

4. CONCLUSION AND SUGGESTIONS

The results of the analysis showed that from 18 items that had been analyzed there were 4 items which had not been identified as DIF, there were 5 items that had been identified containing DIF but not statistically significant and there were 9 items that had been identified containing DIF significantly. Many items favored students in grade 9, and another items favored students in grade 10. They were caused by the content of items and depended the posision of the content in the curriculum.

The students aged about 15 years who were in grade 10 had finished studying the subject more than students of the same age, but was in grade 9. It can be seen from the curriculum standards of education in Indonesia (Kementrian Pendidikan Nasional, 2006; 2016). The chapter about statistics and data, and also uncertainty has been learnt by student in the end of 9th, so that those items with this content benefit students in grade 10. Other factor was students of grade 10 has been pass the national exam. Before take this exam, students did a lot of exercises accompanied by deepening material (Sumarno, Sumardiningsih, Muhson, Retnawati, Basuki, 2011). The second thing is what affects the DIF load those polytomous items shaped mathematical literacy is more favor group of participants in grade 10, when compared with a group of students from grade 9. This gives a hint of the development of mathematical literacy skills from grade 9 to grade 10.

The reseach result about DIF load in items of literacy test is in line with many research. The reseach result of Akour, Sabah, and Hammouri (2015) shows that many science items of PISA test contain net and global DIF, and so do in the reading items (da Costa & Araujo, 2012). In mathematics items of PISA, many items in multiple choiche format load DIF favouring male and many items in constructed response load DIF favouring female (Lyons-Thomas, Sandilands, & Ercikan, 2014).

Some future research can be done related to the results of this study. The comparison difficulties of students grade 9 and grade 10 to solve the problems or questions of PISA released items or PISA-like can be done. The development of mathematical literacy skills in grades 9 and 10, or grade level more can be done, either by utilizing the approach of classical test theory and item response theory. Details of students' skills in mathematical literacy, such as domain content, context, and process can be further investigated. The studies result can then be utilized for the improvement of the learning of mathematics.

5. REFERENCES

Acara, T. (2011). Sample size in differential item functioning: An application of hierarchical linear modeling. Kuramve Uygulamada Eğitim Bilimleri (Educational Sciences: Theory & Practice), 11(1), 284-288.

Adams, R.J. (1992). Item Bias. In Keeves, J.P. (Ed), *The IEA technical handbook* (pp. 177-187). The Hague: The International Association for the Evaluation of Educational Achiement (IEA).

Adams, R., & Wu, M. (2010). *Differential Item Functioning*. Retrieved from https://www.acer.org/files/Conquest-Tutorial-6-DifferentialItemFunctioning.pdf

Akour, M., Sabah, S., & Hammouri, H. (2015). Net and global differential item fuctioning in PISA polytomously scored science items: application of the differential step functioning framework. Journal of Psychoeducational Assessment. 33(2), 166-176.

Budiono, B. (2004). Perbandingan metode Mantel-Haenszel, sibtest, regresi logistik, dan perbedaan peluang dalam mendeteksi keberbedaan fungsi butir. *Dissertasion*. Universitas Negeri Yogyakarta, Indonesia.

Bulut, O., & Suh, Y. (2017). Functioning with the multiple indicators multiple causes model, the item response theory likelihood ratio test, and logistic regression. Frontiers in Education, October 2017, 1-14.

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

Da Costa, P.D., & Araujo, L. (2012). Differential item functioning (DIF): What function differently for Immigrant students in PISA 2009 reading items? JRC Scientific and Policy Reports. Luxembourg: European Commission.

Elosua, P., & Wells, C. S. (2013). Detecting dif in polytomous items using MACS, IRT and ordinal logistic regression. Psicológica, 34(2), 327-34

Hambleton, R. K. & Swaminathan, H. (1985). *Item response theory*. Boston, MA: Kluwer Inc.

Hambleton, R.K., Swaminathan, H. & Rogers, H.J. (1991). *Fundamental of item response theory*. Newbury Park, CA: Sage Publication Inc.

Holland, P.W. & Thayer, D.T. (1988). Differential Item Performance and the Mantel-Haenszel Procedure. In Wainer, Howard; Braun, Henry I. (eds.) Test Validity (p p129-145). Hillsdale, NJ: Lawrence Erlbaum.

Jailani, J., Retnawati, H., Musfiqi, S., Arifin, Z., Riadi, A., Susanto, E., Wulandari, N. F. (2015). Pengembangan perangkat pembelajaran berbasis higher order thinking skills. *Research Report*. LPPM Universitas Negeri Yogyakarta.

Kartowagiran, B. & Retnawati, H. (2008). Pengembangan mengembangkan metode pendeteksian keberfungsian butir pembeda (differential item functioning, DIF) multidimensi. *Laporan Penelitian*. Lembaga Penelitian Universitas Negeri Yogyakarta.

Kementerian Pendidikan dan Kebudayaan Republik Indonesia. (2016). Peraturan Menteri Pendidikan dan Kebudayaan Republik Indonesia Nomor 21 tahun 2016 tentang Standar Isi. [Ministry of National Education of Republik Indonesia. (2016). *Regulation of Ministry of National Education of Republik Indonesia No 22 Year 2006 about Content Standard in Education*.]

Kementerian Pendidikan Nasional Republik Indonesia. (2006). *Peraturan Menteri Pendidikan Nasional Nomor 22 tahun 2006 tentang Standar Isi*. [Ministry of National Education of Republik Indonesia. (2006). *Regulation of Ministry of National Education of Republik Indonesia No. 22 Year 2006 about Content Standard in Education*.]

Khalid, M.N., & Glass, C.A.W. (2013). A step-wise method for evaluation of differential item functioning. Journal of Applied Quantitative Methods, 8(2), 25-47.

Lyons-Thomas, J., Sandilands, D., & Ercikan, K. (2014). Gender differential item functioning in mathematics in four international jurisdictions. Eğitim ve Bilim (Education and Science) 39(172), 20-32.

Masters, G.N. (2010). The partial credit model. In Nering, M.L., & Ostini, R. (Eds). *Handbook of ıtem response theory models*. New York: Routlegde.

Mazor, K. M., Kanjee, A., & Clauser, B. (1995) Using logistic regression and Maentel-Haenszel with multiple ability estimates to detect differential item functioning. *Journal of Educational Measurement*, *32* (2), 131-144.

Muraki, E., & Bock, R.D. (1997). *Parscale 3: IRT based test scoring and item analysis for graded items and rating scales*. Chicago: Scientific Software.

OECD. (2014). *PISA 2012 results: what students know and can do - student performance in mathematics, reading and science*. Paris: OECD Publishing.

Ogbebor, U., & Onuka, A. (2013). Differential item functioning method as an item bias indicator. *Educational Research. 4*(4), 367-373.

Osterlind, S.J. (1983). *Test item bias*. Beverly Hills, CA: Sage Publications Inc.

Plake, B.S., Patience, W.M., & Whitney, D. R. (1988). Differential item performance in mathematics achievement test items: Effect of item arrangement. Educational and Psychological Measurement, 48(4), 885-894.

Retnawati, H. (2003). Keberfungsian butir diferensial pada perangkat tes seleksi masuk smp mata pelalajaran matematika. Jurnal Penelitian dan Evaluasi Pendidikan, 5(6), 45-58.

Retnawati, H. (2013). Pendeteksian keberfungsian butir pembeda dengan indeks volume sederhana berdasarkan teori respons butir multidimensi. Jurnal Penelitian dan Evaluasi Pendidikan, 17(2), 275-286.

Retnawati, H. (2014). Teorı respons butir dan penerapannya. Yogyakarta: Parama.

Salehi, M. & Tayebi, A. (2012). Differential Item Functioning: Implications for Test Validation. Journal of Language Teaching and Research, 3(1), 84-92.

Sumarno, S., Sumardiningsih, S., Muhson, A., Retnawati, H., & Basuki, A. (2013). Faktor yang mempengaruhi menurunnya capaian siswa pada Ujian Nasional 2013. *Laporan Penelitian*. Direktorat PSMP Kementerian Pendidikan Republik Indonesia. [Sumarno, S; Sumardiningsih, S; Muhson, A.; Retnawati, H.; Basuki, A. (2013). Factors affecting students achievement in national examination 2013. *Research report*. Directorate of Secondary School of Ministry Education Office of Republik Indonesia.]

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. 67-113). Hillsdale NJ: Erlbaum.

Varberg, D. & Purchell, E.J. (2001). *Calculus* (Kalkulus, translated by Susia, I.N.). Bandung: Interaksara.

Wang, W.C., Yeh, Y.L., & Yi, C. (2003). Effect of anchor item methods on differential item functioning detection with the likelihood ratio test. Applied Psychological Measurement. 27(6), 479-498.

Wu, M.L., Adams, R.J., & Wilson, M.R. (1997). *ConQuest: Multi-aspect test software [computer program]*. Camberwell: Australian Council for Educational Research.