Differential item function of gender in the mathematics elementary school tryout test

M Ardiyaningrum¹, L Badriah¹, Trisniawati², Suhartini³, S A Widodo²

¹Pendidikan Guru MI, Universitas Alma Ata, Jl. Brawijaya no 99 Tamantirto Bantul, Yogyakarta, Indonesia
²Universitas Sarjanawiyata Tamansiswa, Jl. Batikan UH III/1043, Yogyakarta, Indonesia
³Universitas Alma Ata, Jl. Brawijaya no 99 Tamantirto Bantul, Yogyakarta, Indonesia

Email: martaliayudhono87@gmail.com

Abstract. The purpose of this paper is to describe: 1) number of items in the elementary school mathematics national examination tryout test containing DIF gender and pattern; 2) functionality of differential tests on elementary school mathematics national examination tryout test tool organized by the Pythagoras tutoring institute. This study is a quantitative explorative descriptive study with a type of case research, because it describes the characteristics of the elementary school mathematics national examination tryout tests device and is not generalizable. The amount of data used comes from the answer patterns of 919 students consisting of 431 female students and 488 male students. As for the results of this study are: 1) the elementary school mathematics national examination tryout test kit has a different power ranging from 0.282 to 1.129 and the level of grain difficulty ranges from -3.365 to 2.838; 2) Items in the elementary school mathematics national examination try out test containing 34 items or as much as 87.2%; 3) the elementary school mathematics national examination tryout test kit benefited the female group in theta > -3.06 which was seen from the DTF (differential test function) graph in the male and female groups.

1. Introduction
The Primary School National Examination Result is one of the requirements to enroll in Junior High School (SMP). This is a concern for parents so that their sons / daughters can get the highest score and can be accepted in expected junior high school. One of the efforts are conducted is to send their children in tutoring institution in the community. One of the tutoring institutions that routinely conducts elementary school national exam tryouts namely Pythagoras.

Analysis of the test instrument is needed to ensure that the instrument meets the established criteria and is one of the methods taken to ensure the quality of the test instrument [1]. Explains that a condition on a test device that is influenced by other factors, other than what is being measured is called a bias on a test [2]. Bias in a test is an unfair test condition, inconsistent, contaminated by factors apart from the aspects to be tested, and misuse of the test.

Unjust test conditions can be explained that the test contains partiality to a particular group. Then this condition states that the items in the test contain differential item functioning (DIF). An item shows bias if the item response function between subgroups is not identical, whereas if the item response function between subgroups is identical, it means that the item does not contain DIF [3]. Items containing DIF are items that provide more benefits to certain groups, such as regions (cities and villages), culture...
(Batak and Javanese), gender (men and women), school categories (state and private), and so on. The DIF analysis on the test kit is very necessary to avoid bias from each item making up the test device. The Pythagoras guidance agency has not carried out an analysis of the DIF specifically based on gender. Some studies also explain the differences in the results of large-scale academic tests in the field of mathematics caused by gender differences [4]–[10].

2. Method

This research is a quantitative study as a research approach explaining a phenomena by collecting numerical data that are analyzed using statistical approaches[11]. This research have a type of case research because it describes the characteristics of the primary school’s Mathematics National Exam try out. The main data used in the research was in the form of answers / responses to elementary school students who took part in the Mathematics try out organized by the Pythagoras Tutoring Institute. The type of data used was quantitative data. The sample of the research was all students who attended the try out in Pythagoras's tutoring. The study sample was 919 students consisting of 431 male students and 488 female students.

Before carrying out the DIF analysis, an analysis of the characteristics of the items of the Mathematics National Exam tryout test was carried out first with the item response theory (IRT) approach. The tryout test instrument had fulfilled the IRT assumptions, namely dimensions, local independence, and parameter invariance. The next step was to determine the suitability of the model parameter estimation of test items [12]. In determining the item parameters and chi-square score, BILOG program version 3.85 was used. After the chi-square score of each parameter model was compared, the results obtained that the parameter type 2 PL is best used for item analysis. So that the next analysis would use parameter test item estimation with 2 Logistics Parameters. The following is the logistic model of 2 parameters (2 PL) with the item parameters, namely the item difficulty index (bi) and the item difference power index (ai), which meet:

\[ P_i(\theta) = \frac{e^{a_i(\theta - b_i)}}{1 + e^{a_i(\theta - b_i)}} , \ i = 1,2,3, \ldots, n \]  

Whereas:

\[ P_i(\theta) = \text{the probability of test takers who have the ability } \theta \text{ randomly selected can answer item i correctly} \]

\[ \theta = \text{level of subject ability (as independent variables)} \]

\[ b_i = \text{i item difficulty index} \]

\[ e = \text{natural number which value is close to 2.718} \]

\[ n = \text{the number of items in the test}. \]

Furthermore, for the detection of the presence of DIF will be carried out by the Chi-Square method of Lord’d (\( \chi^2 \)) and supported by the ICC method to detect the type of DIF (Uniform or Non-Uniform) in each item of test equipment and also with the TCC method to detect types DTF (Differential Test Functioning) on the Mathematics National Exam tryout.

3. Result and Discussion

3.1. Estimated parameters of test items

The item analysis model that is suitable to be used on the Mathematics Elementary School try out test kit is 2 PL. Then the data were analyzed using the 2 PL model, while the results are presented in the table 1.

| Item number | a   | b   | Item number | a   | b   |
|-------------|-----|-----|-------------|-----|-----|
| 1           | 0.462| -2.584| 21          | 0.505| -0.003|
| 2           | 0.622| -2.06 | 22          | 0.433| 0.716 |
| 3           | 0.622| -0.214| 23          | 0.310| -0.724|
Table 1 shows that the difficulty level of all items is between -3,365 and 2,838. Item 34 has the lowest difficulty level and item 40 has the highest level of difficulty. Item 34 is an item that has the opportunity to be answered correctly by all students, on the contrary for item 40 which requires higher thinking skills to get the right answer. While the grain difference value is between 0.282 to 1.129. This shows that from the different aspects of power, all of these items are good items because they have a range between 0 and 2 \[13\].

### 3.2. Differential Item Function (DIF) Analysis

To test the presence or absence of DIF on the item test kits, student responses were grouped by gender of students (male and female). After that, an analysis of the presence of DIF was carried out on the items in the Mathematics tryout test with the Lord-squared method based on student gender.

In the statistical test using the Lord-Squared method of Lord, the steps taken to detect DIF are as follows [2]:

1) estimate the item parameters and variances-covariances of the focal groups and reference groups separately. In this research, the focal group was a male group and the reference group was a group of female.

2) determine the constant equalization of parameters between the focal group and the reference group.

3) counting \( \chi^2 \) value and compare it with a critical value table \( \chi^2 \) with degree of freedom 2 and could use the desired level of significance. Below is the equation \( \chi^2 \) which would be used to determine the presence or absence of DIF\[14\], \[15\]:

\[
(\chi^2) = v_i'\Sigma^{-1}v_i
\]  

\[\text{whereas:}\]

\( v_i \) = vectors of differences in estimating the fourth item parameters between male and female groups

\( \Sigma^{-1} \) = is a variance-covariance matrix for differences in estimating item parameters.

The following will be described in each step in detecting the presence of DIF on each item using the Lord’s Chi-Square method:

1) Estimating item parameters and variances-covariances of the male and female group separately
Kim and Cohen state that in the DIF study, item parameter estimates must first be placed in the same metric before a comparison is made to determine whether there are items that function differently[16]. In accordance with the results of the selection of the most appropriate model for the Mathematics tryout test kit, the model used was analysis with 2 PL. Each gender group (male and female) will first be estimated the item parameters with the help of the BILOGMG 3.0 program. After processing the data with the BILOGMG 3.0 program, results were obtained in the form of item parameters and the value of male and female groups. Based on these results, it was obtained the fact that different power item for the male group was from 0.320 to 1.091, while for the female group it was from 0.299 to 1.162. The parameter of item difficulty level for the male group has ranged from -3,415 to 2,445, while for the female group it was from -2,808 to 2,959. Based on these results it appears that the level of difficulty for both groups is still in the range of -4 to 4, still in the good category. This is in line with the initial steps that have been taken, namely testing the most suitable logistic parameter model that can be used for parameter estimation on item response theory. With these assumptions, it is expected that the analysis using the item response theory can be done with the right parameter model approach.

In this study an analysis of 40 items will be conducted, but after going through the data processing using BILOGMG, the fact that item 1 cannot be estimated because the parameter estimation results for the male group does not come out, so it is assumed that item 1 has been automatically annulled by the program BILOGMG. Furthermore, the DIF analysis will be carried out for items number 2 to point 40.

2) Determine the constant equalization of parameters between male and female groups
As explained in the description above, table 1, that the power range of different groups of male and female groups is almost the same, then the two groups have equal item parameters. Likewise, the parameters of item difficulty level have almost the same range between group of male and group of female

3) Counting $\chi^2$ value and compare it with the critical value table for female groups $\chi^2$
In determining $\chi^2$, The researcher used the help of Microsoft Excel, by first determining the vector of the difference in estimating the item-i parameter between the male group and the female group, then determining the inverse of the matrix variance-covariance difference in estimating the item parameters. $\chi^2$ formula can be seen in the equation (2). The following are the results of data processing using the method of $\chi^2$. Other than items 2, 10, 14, 20, and 37 contain DIF.

| Item Number | $\chi^2_{\text{count}}$ | $\chi^2_{\text{table}}$ | Description |
|-------------|--------------------------|--------------------------|-------------|
| 2           | 0.293829                 | 5.99                     | Does not contain DIF |
| 10          | 3.545124                 | 5.99                     | Does not contain DIF |
| 14          | 1.544126                 | 5.99                     | Does not contain DIF |
| 20          | 3.965017                 | 5.99                     | Does not contain DIF |
| 37          | 0.821541                 | 5.99                     | Does not contain DIF |

DIF Identification Table with the results comparing $\chi^2_{\text{count}}$ with $\chi^2_{\text{table}}$ presents the results of the presence or absence of DIF in each item on the Mathematics National Exam tryout. In determining the presence or absence of DIF, the provision is used if $\chi^2_{\text{count}} < \chi^2_{\text{table}}$, then it can be concluded that the item does not contain DIF. As many as 5 items, namely 2, 10, 14, 20, and 37 are items that do not contain DIF, meaning that the items are not partial / beneficial to one group. Items that do not contain DIF in this gender group can be given to both male and female students. While the items containing DIF were 34 items or around 87.2%. Items containing this DIF if given to students of different gender will produce a response that is more favorable to one gender group.

To find out which group is more profitable can be done using the Item Characteristic Curve (ICC) method. In using the ICC method, the first thing to do is estimate the parameters of each item for the male and female groups. The item analysis model that is suitable for this case is 2 PL, so only
the difficulty level parameters are used \( b \) and different power \( a \). Each item parameter is substituted in equation 1.

The probability value of each item for the male and female groups is compared through the ICC curve as below. The figure below only presents 5 items from the 39 items identified in the DIF analysis. The ICC of the 5 items is presented in the following figure.

Figure 1. Item 2 ICC

Figure 1 above represents item ICC 2 for male and female groups. Graphically, it appears that there is intersection of male group graphs with female groups, this indicates the possibility of a non-uniform DIF. Non-uniform DIF can be identified through the intersection between the curves of the male group and the curves of the female group. Through the curve, it is obtained the fact that at theta \(<-1.60\) point 2 shows that the male group has a higher chance to answer correctly than the female group. However, for theta \(>1.6\) this happens otherwise, female groups have a higher chance to answer correctly. Statistically through Lord’s Chi Square method shows that item 2 does not have a significant DIF. This second item has a form of a contextual problem about multiplication and division operations. For a higher level of ability, this question is more beneficial for female students, this is in line with the opinion of Willingham & Cole which states that female students tend to be better at arithmetic and algebra operations [17]. But overall, this item does not contain DIF.

Figure 2. Item 10 ICC

Figure 2 above represents item ICC 10 for male and female groups.
The ICC figure 2 for Item 10 is not much different from item 2, which has a curve intersection at theta -1.4 so it is assumed that this item contains a non-uniform type DIF. However, after seeing the chi-square value, it turns out that this item is significantly an item that does not contain DIF.

Item 37 has conditions that is slightly different from the previous curve, this result is shown in figure 3. The 37th ICC item between male and female groups tends to coincide. In the ICC method this condition can indicate that item 37 is DIF. Likewise after being tested statistically this item does not contain DIF. This condition can be interpreted that item 37 does not benefit one gender group.

ICC item 3 shows a fairly wide area between the curves of the male group and the female group. with the ICC method, this shows the existence of DIF between male and female groups, as well as statistically, this item contains DIF. The gender group that benefited through item 3 is the male group. DIF in point 3 is included in the type of DIF UNIFORM. This 3rd item is an item that measures students' abilities in the material of rank. Some previous studies revealed that in algebraic material, male have a tendency to outperform female [5], [17]. The results in item 3 also show that items measure algebraic abilities are more beneficial for male students.
Different from item 3, point 16 contains non DIF uniform. The ICC point 16 on figure 5 shows the broad area formed by the male curve and the female curve. This condition indicates the presence of DIF in point 16. Female students with $\theta > -2.2$ have more profits at point 16. Item number 6 is an item that measures students' abilities in solving geometric problems. In terms of non-routine problem solving, female students benefit more from male [5], [18]. Several other studies also mention that gender allows for different results on the routine processing of items learned in the classroom [18]–[20].

The figure 6 shows that the area formed between the curves between the male and female groups is quite extensive. This indicates the presence of DTF on the Mathematics Try Out Test kit. This test is beneficial for the female group than the male group in $\theta > 3.00$. Female students with abilities of more than 3.00 have a greater chance to answer correctly compared to the male group. The DTF of this test kit is in the non-uniform category, because there is a cross between the male and female groups.

4. Conclusion
Based on the description of the results and discussion above, conclusions are obtained as follows. The test tool for the try out Mathematics National Exam of the Pythagoras Tutoring Institute consists of 40 items that have different power ranging from 0.282 to 1.129, with the lowest difference in power possessed by item 31, while the highest difference power is owned by the 17th parameter. Next parameter item is the level of difficulty of items which ranges from -3.365 to 2.838. Item in the
Elementary School Mathematics National Examination Test try out containing 34 items of DIF or 87.2%. Items that do not contain DIF only consist of 5 items, namely items 2, 10, 14, 20, and 37, the rest contain DIF. Items containing DIF have 2 types, namely; some are in the form of uniform and some are in the form of non-uniform. The Try Out of Mathematics National Exam for Elementary School benefits the female group in theta > -3.06 which is seen from the DTF (Differential Test Function) graph in the male and female groups. DTF for this test device falls into the Non-Uniform DTF category because it has the intersection of the female curve and the male curve.

Acknowledgments
Thank you to the Phytagoras Study Guidance Institute for providing research data.

References
[1] T. Martalia Ardiyaningrum, Cahya Kusuma, “ANALISIS BUTIR TRY OUT UJIAN NASIONAL MATEMATIKA SEKOLAH DASAR DI DAERAH ISTIMEWA YOGYAKARTA TAHUN 2017,” Taman Cendekia, vol. 02, no. 02, pp. 206–211, 2018.
[2] H. Retnawati, “Teori Respon Butir dan Penerapannya (Untuk Peneliti, Praktisi Pengukuran dan Pengujian, Mahasiswa Pascasarjana),” 2014.
[3] R. K. Hambleton, “Customized Tests and Customized Norms,” Appl. Meas. Educ., 1991.
[4] N. Abedalaziz, “Detecting DIF using item characteristic curve approaches,” Int. J. Educ. Psychol. Assess., 2011.
[5] N. Abedalaziz, “Detecting gender related DIF using logistic Regression and Mantel-Haenszel approaches,” in Procedia - Social and Behavioral Sciences, 2010.
[6] J. D. Scheuneman and A. Grima, “Characteristics of Quantitative Word Items Associated with Differential Performance for Female and Black Examinees,” Appl. Meas. Educ., 1997.
[7] M. M. Kimball, “A New Perspective on Women’s Math Achievement,” Psychological Bulletin. 1989.
[8] J. S. Hyde and M. C. Linn, “Gender similarities in mathematics and science,” Science. 2006.
[9] D. C. Geary, “Sexual selection and sex differences in mathematical abilities,” Behav. Brain Sci., 1996.
[10] D. C. Geary, “Sexual selection and sex differences in spatial cognition,” Learn. Individ. Differ., 1995.
[11] Aliaga and Gunderson, “Quantitative and Qualitative Research Methods | SkillsYouNeed,” skills you need, 2005.
[12] A. A. Rupp, “Item Response Modeling With BILOG-MG and MULTILOG for Windows,” Int. J. Test., 2004.
[13] R. K. Hambleton and H. Swaminathan, Item Response Theory Princible and Applications. 1985.
[14] S. Stark, O. S. Chernyshenko, and F. Drasgow, “Detecting differential item functioning with confirmatory factor analysis and item response theory: Toward a unified strategy,” J. Appl. Psychol., 2006.
[15] G. Camilli and L. A. Shepard, “A computer program to aid the detection of biased test items,” Educ. Psychol. Meas., 1985.
[16] S. H. Kim and A. S. Cohen, “A Comparison of Lord’s Chi-Square, Raju’s Area Measures, and the Likelihood Ratio Test on Detection of Differential Item Functioning,” Appl. Meas. Educ., 1995.
[17] N. S. C. Warren W. Willingham, “Gender and fair assessment,” Choice Rev. Online, 2013.
[18] G. Seegers and M. Boekaerts, “Gender-Related Differences in Self-Referenced Cognitions in Relation to Mathematics,” J. Res. Math. Educ., 2006.
[19] A. M. Gallagher, R. De Lisi, P. C. Holst, A. V. McGillicuddy-De Lisi, M. Morely, and C. Cahalan, “Gender Differences in Advanced Mathematical Problem Solving,” J. Exp. Child Psychol., 2000.
[20] J. S. Hyde, E. Fennema, and S. J. Lamon, “Gender Differences in Mathematics Performance: A Meta-Analysis,” Psychol. Bull., 1990.