THE APPLICATION OF THE GENERALIZED LORD'S CHI-SQUARE METHOD IN IDENTIFYING BIASED ITEMS

Herwin
Faculty of Education, Universitas Negeri Yogyakarta

Sophak Phonn
Ministry of Education, Youth, and Sport of Cambodia

Abstract
This study aims to find out the results of the analysis of item bias by using Generalized Lord's Chi-square test method on the test instrument of elementary school examination in a sub-district of Gowa Regency, Indonesia. This research is explorative research using quantitative approach. This research was conducted in the second semester of the academic year of 2017/2018 at the Bontomarannu District Elementary School in Gowa Regency. Data collection technique used is documentation. The data in this study were analyzed using the DIF method of Generalized Lord Chi-Square test. The results show that in using the Generalized Lord Chi-Square method, from 20 items of mathematics test of school exam in Bontomarannu District Elementary School in Gowa Regency of the academic year 2017/2018, there are two items which are detected containing the bias (DIF), i.e., item 5 and item 11, while the rest are not.
Keywords: *DIF, generalized Lord's chi-square, school examination*

Contact Herwin

herwin89@uny.ac.id

Department of Elementary-School-Teacher Training Education, Faculty of Education, Universitas Negeri Yogyakarta
Jl. Colombo No. 1, Karangmalang, Depok, Sleman, Yogyakarta 55281, Indonesia

Introduction

Assessment is an important component of education. One effort that can be done to improve the quality of education is to improve the quality of the assessment system. A good assessment system that will encourage teachers to determine good teaching strategies in motivating students to learn better. Therefore, to improve the quality of education, it is necessary to improve the assessment system that is applied. From this assessment, the teacher will obtain the ability of the portrait or profile of students to achieve the basic competencies set out in the curriculum (Herwin, 2016, p. 1).

A good rating system must be supported by various components. One component that supports the implementation of a good assessment is a quality assessment instrument or measuring instrument. In creating a measurement, a valid and consistent (reliable) measurement tool is needed. By using a measuring instrument that meets both of these criteria, the results of repeated measurements will be obtained in accordance with what is to be measured without being affected by other factors. A condition on a test instrument that is influenced by other factors, other than what is being measured, is called a bias on a test (Retnawati, 2014, p. 125).

One indication of a good instrument is an instrument that is free from the threat of bias. Bias is defined as a systematic error in the measurement process (Osterlind, 1983, p. 10). Bias is a negative condition that can threaten the quality of the test. Item bias can harm a group of students based on gender, religion, ethnicity, social, and economic status of different groups, when, in fact, the two groups have the same knowledge and also abilities.

The Differential Item Function (DIF) is a procedure used to identify bias test items. DIF is a psychometric method which is generally used to overcome justice in achieving standards, intelligence, certification, and testing of licenses (Jodoin & Gierl, 2001, p. 329). DIF occurs when participants from various groups show different probability of success in the item after matching the ability underlying the item intended to be measured (Zumbo, 1999, p. 12). The DIF analysis aims to detect differences in item responses in questionnaires, rating scales, or tests in different subgroups (e.g., sex) while controlling ability levels (Zhang, 2015, p. 1). Bias is a threat to the test. Bias can appear as an invalid problem, for example, in terms of construction problems or content that is not suitable for a particular group of students. The bias of items in an instrument can affect test results that are not good. That is, the decisions obtained from the test will be less objective if the instrument contains an item bias. The bias presence of items must be detected so that the test can be of higher quality.

Since the function and purpose of the test are very important, the measurement instrument used must be of high quality (free of bias). Thus, the detection of bias items in a measuring instrument is very important. The bias item procedure is used to determine whether individual items in the examination function in the same way for two groups of examinees are usually defined by racial and ethnic background, gender, age, and experience, or disability conditions (Scheuneman & Bleistein, 1999, p. 220).

Statistically, bias items occur when there is a difference in support (alignments) of an item against a particular group. To understand the source of the bias that occurs in an item, it is necessary to check whether the response to the item is systematically related to certain characteristics of the respondents such as gender, age, ethnicity and other differences (Bares, Andrade, Delva, Grogan-Kaylor, & Kamata, 2012, p. 387).

Many different ways can be used in detecting bias items, but in this study, the method used is the Generalized Lord's chi-square test method. The Generalized Lord's chi-square test method (Kim, Cohen, & Park, 1995), is a DIF identification method which is often referred to as Qj statistics, making it possible to detect uniform or non-uniform differential function among many groups by setting the appropriate item re-

sponse model. Therefore, based on some of these descriptions, this research was conducted. If the DIF indication level is practically significant, it can be tested by using a certain statistical test or just by looking at the index, then the item in question is said to be exposed to DIF, load DIF, or detected as DIF item (Budiyono, 2009, p. 3).

Based on the description of the background of the problem that has been explained, the research question is how the results of the bias item analysis using the Generalized Lord's Chi-square test method on the elementary school exam test instrument in Bontomarannu District, Gowa Regency in 2018? Based on the research question, the implementation of this study is aimed at determining the results of the bias item analysis using the Generalized Lord's Chi-square test method on the elementary school exam test instrument in the district of Bontomarannu, Gowa Regency, in the academic year of 2017/2018.

Research Method

This research is exploratory research using a quantitative approach, aimed at detecting the bias items or differential item functioning (DIF) on the class II elementary school exam test instrument in the district of Bontomarannu, Gowa Regency, in 2018. This research was conducted in the elementary school of Bontomarannu District, Gowa Regency, in the even semester of the academic year of 2017/2018.

Data collection techniques are carried out by documentation. The data in this study are class II elementary mathematics questions on school exams in Bontomarannu District, Gowa Regency in 2018, consisting of 20 items (15 multiple-choice items and five filling items). In addition to the question set, this research data are the results of the response or answers of the students as many as 400 answer sheets to be continued on the analysis of item bias detection. Data from the responses or answers of participants in the form of 1-0 dichotomous data are derived from the multiple choice objective test answers and objective filling tests.

In this study, the grouping is divided into four groups: rural women students, urban women groups, rural male groups, and urban male groups. In addition, in this study, the reference groups were labeled in rural women's groups. Meanwhile, focal groups were labeled in urban women's groups, rural men's groups, and urban male groups.

The data of this study were analyzed using the Generalized Lord Chi-Square test DIF detection method. This method is used to detect bias items in cases of more than one focal group (Magis, Béland, Tuerlinckx, & De Boeck, 2010, p. 852). Kim et al. (1995) expand the Lord Chi-square test method to be more than one focal group in a procedure called the Generalized Lord Chi-square test. Lord Chi-square statistics are then generalized to Equation 1:

$$Q_j = (Cv_j)'(C \Sigma_j C')^{-1}(Cv_j) \qquad (1)$$

Referring to Equation 1, vj is obtained by combining the vector parameter items estimated in the reference group and focal group. Σj is a diagonal block matrix where each diagonal block is a matrix of item variance-covariance parameters in each group. C matrix is a design matrix that shows the parameters of the items you want to compare between groups (for more details, see Kim et al., 1995). The threshold (or cut-score) for classifying items as DIF is calculated as the quantile of the chi-square distribution with a lower-tail probability of one minus alpha and p degrees of freedom.

Findings and Discussion

Findings

This research is focused on the application of the Generalized Lord Chi-square method to detect biased items in the school exams test instrument in the Elementary School of Bontomarannu District, Gowa Regency, in the academic year of 2017/2018. This Generalized Lord Chi-square method is applied because in this case, there is more than one focal group. This research is based on problem analysis using data from the test

participant's response results or the results of student answers to mathematical questions that have been done.

In the DIF study, there were at least two groups, namely, focus and reference groups. Focus groups are basically a minority group, for example, are potentially disadvantaged groups. Groups that are considered potentially benefited by this test are called reference groups. However, it should be stressed that naming groups is not always clear. Naming groups in such cases is often done randomly (Karami, 2012, p. 60).

There are two types of DIF, namely DIF that is uniform and non-uniform. Uniform DIF occurs when groups perform better than other groups at all levels of ability. That is, almost all group members outperform almost all other group members who are at the same level of ability. In the case of non-uniform DIF, members of one group are favored to the extent of the ability scale and from that point on the inverse relationship. Thus, there is an interaction between grouping and level of ability (Karami, 2012, p. 60).

Based on the results of the research obtained from the question instrument, documentation, and responses, it is found that in the 2018 elementary school examination in Bontomarannu Subdistrict, specifically for Class II and Mathematics subject, the results showed that the instruments used were a multiple-choice objective test consisting of 15 items and five filling items. In addition, the results of the students' answers were also obtained in the form of response answers on each question that was done.

Analysis of item bias or detection of DIF in this study was viewed from two aspects, namely gender aspects (male and female) and location aspects (urban and rural). Data distribution from the subject of this study is described in Table 1.

Table 1. Distribution of Research Data Sources

Location / Gender	Rural	Urban
Female	127	130
Male	73	70

Table 1 shows that the overall data of respondents amounted to 400 response answers or 400 answer sheets. The data in the form of answers to the test participants are included in the application, namely the R Program, which is then estimated using the Generalized Lord Chi-Square test method. The results of the bias item analysis obtained from the R Program are presented in Table 2.

Table 2. Results of Item Bias Analysis using the GLC Method

Item	GLC	Cut-score	p	α	Bias
1	0.31	7.8147	0.94	0.05	no
2	0.63	7.8147	0.88	0.05	no
3	1.09	7.8147	0.77	0.05	no
4	0.39	7.8147	0.94	0.05	no
5	12.2	7.8147	0.00	0.05	yes
6	5.68	7.8147	0.12	0.05	no
7	0.94	7.8147	0.81	0.05	no
8	1.25	7.8147	0.74	0.05	no
9	1.11	7.8147	0.77	0.05	no
10	0.73	7.8147	0.86	0.05	no
11	32.8	7.8147	0.00	0.05	yes
12	2.58	7.8147	0.45	0.05	no
13	0.79	7.8147	0.85	0.05	no
14	1.92	7.8147	0.58	0.05	no
15	2.24	7.8147	0.52	0.05	no
16	0.43	7.8147	0.93	0.05	no
17	0.53	7.8147	0.91	0.05	no
18	7.03	7.8147	0.07	0.05	no
19	1.17	7.8147	0.75	0.05	no
20	2.77	7.8147	0.42	0.05	no

Table 2 shows information regarding the Generalized Lord's Chi-square statistical coefficient which is the analysis coefficient of DIF, the threshold/cut-score coefficient which is the comparative criterion, p-value which is an opportunity to reject the null hypothesis, α which is the significance level used, and the conclusion item analysis. Based on the results of the analysis presented in Table 2, it can be concluded that of the 20 items of mathematics questions in the school examination in the state elementary school

of Bontomarannu District, Gowa Regency, in the academic year of 2017/2018, there are 18 items that are categorized well (free from gender and location bias), while two items others were detected significantly containing bias items, namely: Item 5 and Item 11. Both items were in the form of multiple choices. In addition to the generalized Lord's Chi-square statistics and p-value, the results of the analysis are also presented in the form of item plots as shown in Figure 1.
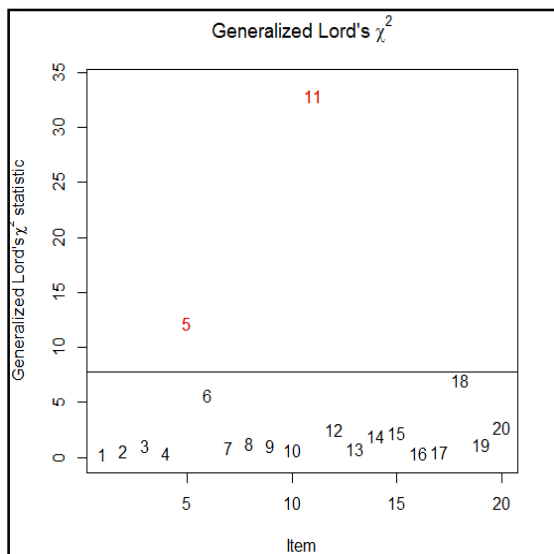

Figure 1. Generalized Lord's Chi Square Item Plot

The plot in Figure 1 provides information regarding item distribution based on bias analysis. Basically, the information in the picture is the same as the results in Table 2 in which two items are significantly detected containing DIF, namely Item 5 and Item 11. The straight line that intersects the plot is the cut-score area of 7.8147. This area serves to limit item bias. It is used to group items that contain significant bias or DIF and items that are free of DIF.

The Generalized Lord's Chi-square statistical method is an item bias detection method or DIF-based item response theory. For this reason, we can describe the item response theory curve to see the difference between items containing DIF and items that are free of DIF. The curve referred to is presented in Figure 2 and Figure 3.
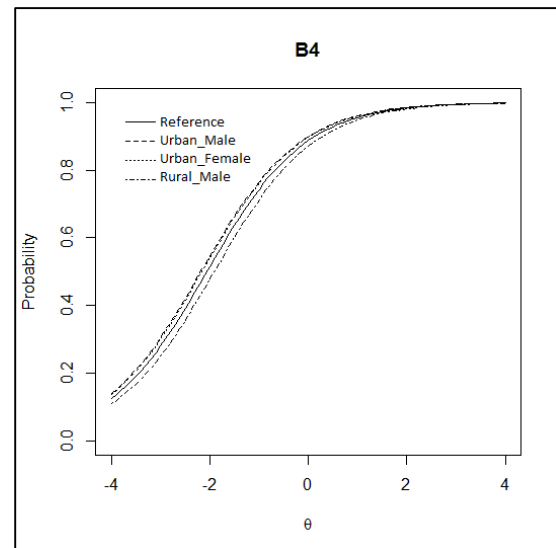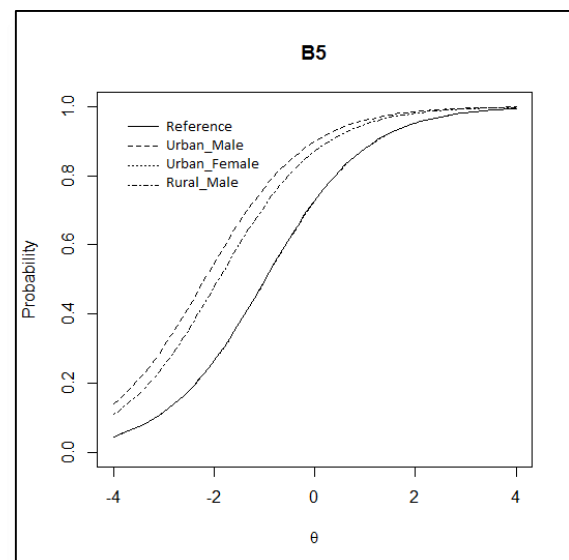

Figure 2. Plot Item 4

Figure 3. Plot Item 5

Figure 2 and Figure 3 show two different item characteristics: Item 4 and Item 5. Based on the previous analysis on the Generalized Lord's Chi-square coefficient and p-value, Item 4 includes items free of bias or temporary DIF, while Item 5 includes items that significantly load DIF. Figure 2 shows that in Item 4, the curve between the four groups looks like each other. It shows that the item does not contain DIF because the chances of test-takers with the same ability to answer Item 4 are also relatively the same.

Another thing is found in Item 5. From Figure 3, it can be seen that the curve

between groups is separate. It shows that Item 5 contains DIF because the probability of examinee with the same ability to answer Item 5 correctly seems different/not the same. It becomes the basis for concluding that Item 5 on the math questions of the school examination in the State Elementary School of Bontomarannu District, Gowa Regency, in the academic year of 2017/2018, if viewed from the aspect of the item, may be in unfavorable category, and need further revision and evaluation for reuse in future.

In addition to Item 5, the results of this study also found that one more item detected significantly contained DIF, namely Item 11. Based on the previous analysis on Generalized Lord's Chi-square, statistical coefficient of 32.89 with a threshold/cut-score of 7.8147 and p-value of 0.00 with α 0.05 indicates that Item 11 includes items that significantly contain bias or DIF. For more details, Item 11 is presented in the form of characteristic items, as shown in Figure 4.
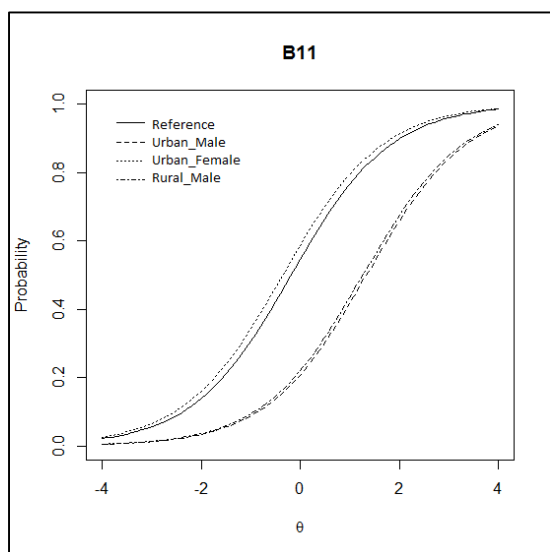

Figure 4. Plot Item 11

Figure 4 shows that in Item 11, there is a curve between separate groups. It is the basis for concluding that Item 11 on the math questions of the school exam in the elementary school of Bontomarannu District, Gowa, in the academic year of 2017/2018 if viewed from the aspect of the item, maybe in unfavorable category and need further revision and evaluation for reuse in future.

The results of this study found that from 20 math questions in the school examination in the elementary school of the District of Bontomarannu, Gowa Regency, in 2017/2018, two multiple-choice items were detected containing bias (DIF) (item 5 and item 11), while the remainder were not detected as bias items. Thus, the two biased items need further evaluation while the other 18 items are good and can be maintained to be used in the test in the future.

Discussion

The results of the study obtained empirical findings; namely, two questions contain bias or contain the DIF (Item 5 and Item 11). It is undoubtedly a material for future evaluations for developers of math questions in school exams in Elementary Schools of Bontomarannu District, Gowa. According to Retnawati (2013, p. 276), ideally, the implementation of the test is based on objectivity, transparency, accountability, and non-discrimination. If a test contains items that are in favor of a particular group, then the test is said to contain bias or contain DIF.

Item 5 is one of the items that is significantly detected to contain bias. Figure 4 shows that the item contains gender bias. The Plot Item 5 shows that the item favored the group of male students. It shows that the chances of male students answering Item 5 correctly are greater than female students, even though their abilities are the same. Further analysis can be done by observing the context of the Question Item 5. For more details, Item 5 (in Indonesian) is presented in Figure 5.
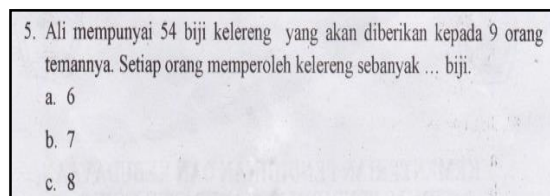

Figure 5. Question Item 5

Note: Translation for Question Item 5
Ali has 54 marbles which will be given to 9 persons. Thus, each person will have … marbles.

Item 5 contains the distribution operation material. Item 5 questions are made in the form of story questions with an alternative choice of 3 answers. Based on the story contest discussed in Item 5, it can be seen that the developer questions raised the story of the marbles to be given. Based on the results of this study, the question turned out to favor male students (see Figure 3) in the Bontomarannu District Elementary School in Gowa, in the academic year of 2017/2018. It shows that the context of the story about the game of marbles seems to be closer to male students. This is attributed to the charge bias in Item 5. According to Kim et al. (1995, p. 1), item bias is basically a condition in which there are differences in the chances of answering correctly on a question item in several groups of respondents or test participants even though the ability between several groups is the same.

In addition, Lautamo, Laakso, Aro, Ahonen, and Törmäkangas (2011, p. 223) explain that bias or often known as DIF occurs when people from different groups have different probabilities of getting a specific score on a test item. DIF analysis provides an indication of unexpected behavior based on items on the test. DIF is also basically a process of validating people's responses (test-takers). Bias can be examined by DIF analysis using a sample consisting of subgroups that differ in essential characteristics that will affect the measured phenomenon. Thus, when one or more item parameters differ between groups, in that condition the item contains DIF (bias).

Further, Karami (2012, p. 59) asserts that Differential Item Functioning (DIF) has been increasingly applied in the study of justice in psychometric circles. DIF occurs when two groups with the same level of ability, not equally capable (different opportunities) answer an item correctly. In other words, one group does not have the same opportunity to get the right item even though its members have a level of ability compared to other groups. If the factor leading to DIF is not part of the construct that is being tested, then the test is biased.

Based on this view, Item 5, which has been detected by DIF, has a different function in the sub-group analyzed, namely gender. Based on the results of empirical analysis of male groups and groups of women have different opportunities to answer Item 5 correctly, even though compared to the same ability parameters.

Beside Item 5, Item 11 is also one of the items that are significantly detected to contain bias. Item 11 in Figure 4 shows that the item contains gender bias. The Item 11 plot shows that the item favored the female student group. It shows that the chances of female students answering Item 11 correctly are higher than male students, even though their abilities are the same. Further analysis can be done by observing the context of the Item 11 question. For more details, Item 11 (in Indonesian) questions are presented in Figure 6.
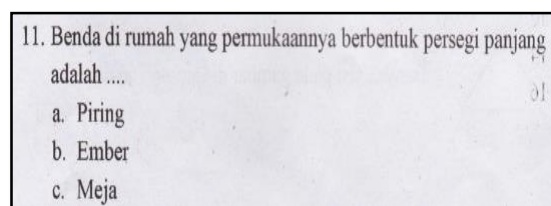

Figure 6. Question Item 11

Note: Translation for Question Item 11
An object in the house which has a rectangle-shaped surface is ….
    A. Plate
    B. Bucket
    C. Table

Item 11 basically contains flat material. Based on the question theme discussed in Item 11, it can be seen that the developer of the question raised about examples of objects in the house, such as plates, buckets, and tables. Empirically, based on the results of this stud-y, the question turned out to favor female students in elementary school of Bontomarannu District, Gowa, Academic Year 2017/2018. It shows that the context of objects in the house, such as plates, buckets, and tables seems to be closer to female students. It is attributed to the content of bias on Item 11. Bias items are important be-

cause the purpose of the DIF analysis is to detect differences in item responses in questionnaires, rating scales, or tests in different subgroups (e.g., sex) while controlling level of ability (Zhang, 2015, p. 1).

De Leo, Van Dam, Hobkirk, and Earleywine (2011, p. 570) state that the DIF charge on an item is often caused by inaccurate questions. These inaccurate questions present one group higher than the other group on certain traits because group membership is not the difference in the characteristics of the real abilities. The results of this study relate to this study that the bias found in the two items is caused by gender differences alone, not differences in the characteristics of the actual ability of the test takers. It is what needs attention in the future related to the development of better question items.

Based on the IRT model, an item displays DIF if the probability of responding in different categories varies across the groups studied, given an equivalent level of the underlying attributes. If the DIF is contained in an instrument, there are several options available. One extreme option is to remove the items from bank items, but the risk is that if the number of items omitted is large, then the condition has a disadvantage, namely, the measurement of precision and flexibility of administration of items which is blocked. Another alternative is available when the number of items with DIF is relatively small, namely making repairs or revisions to items detected loading DIF. In this case, in-depth analysis is needed regarding the causes of the existence of the items in a particular group. If it can be shown that an instrument is free from bias items, then all items in the instrument can be used to correctly estimate the value of the ability parameter (Weisscher, Glas, Vermeulen, & De Haan, 2010, p. 545).

Item 5 and Item 11 are the items that experience DIF. Basically, the two items are unwanted items on the school exam in the Elementary School of Bontomarannu District of Gowa in the academic year of 2017/2018 because they will tend to show differences in the number of subject attributes that actually do not exist. Items like this are items that can discriminate against one group compared to another, so it needs to be repaired or excluded from the test on the next examination.

The findings of this study indicate that there is a bias in gender aspects only, while in the location (urban and rural) aspects, no significant items were found to contain DIF. It means that the location of the school (both rural and urban) does not affect the alignment of the test items. It means that each item on the exam of the Elementary School of Bontomarannu District of Gowa Regency in the academic year of 2017/2018 does not favor the sub-groups of student locations (rural and urban).

Another thing happens in the aspect of gender. The existence of items that have different treatment in terms of gender aspects causes the item detected to contain bias. It is in line with the results of research conducted by Chiesi, Ciancaleoni, Galli, Morsanyi, and Primi (2012, p. 391) that gender is one of the potential aspects that might lead to bias in measuring ability in general. Although some studies (Colom & García-López, 2002, p. 445) also state that there is no significant difference in terms of gender, but empirically this study found a difference in the probability of successfully answering items correctly even in the case of similar abilities.

The bias that occurs in the case of this study is internal bias, which is also commonly referred to as item bias. This item's bias is an aspect of the bias in the test relating to the psychometric properties of a test item and the overall test. This internal bias is focused on investigations about whether each item has similar behavior, namely the similarity in measuring psychometric properties (Adams, 1992, p. 178). If this view is related to the results of this study, it turns out that two items have behavioral inequality or tend to favor certain groups. It is what underlies that there is an interaction between group members on the performance of the items in the examination of Elementary School in the District of Bontomarannu, Gowa Regency, in the academic year of 2017/2018.

Refractive analysis of items is done to check whether each item is fair for each student (examinee) without being caused by inherent differences from the student, such as differences in sex, language, ethnicity, parental education, and others. Detection of bias items in a measuring instrument is very important, considering that the community has time to be more critical or very concerned with the results of measurements such as school exams, especially if these measurements can have a direct impact on students, such as whether a student can be categorized as failed or successful on a test.

The general approach to managing item bias is to issue any item that shows or loads the DIF of an instrument. However, maintaining all items is much better because the development of expensive and time-consuming instruments and testing is quite long. Therefore, one way to deal with items that show DIF is to correct bias by maintaining matching items with the opposite bias on an item. To match the appropriate items and address DIF items at the scale level, the direction, and type of DIF must be recognized correctly (Cho, Martin, Conger, & Widaman, 2010, p. 176). This view implies that, basically, detected items loaded with DIF are generally acted upon by removing or removing the item from the instrument. On the other hand, removing items in the instrument will interfere with the validity of the instrument's contents so that the item is fixed, and is a good alternative for energy efficiency and time for test developers. However, to correct and maintain the item, the developer must recognize the direction and source of bias correctly.

The bias content of the appraisal item is a bad thing and decreases the level of credibility of the test (assessment). If an assessment instrument is dominated by many bias items, then, the results of decisions or conclusions issued from the implementation of the assessment will also be biased, because one of the indicators of the quality of the assessment is the quality of the instrument or appraisal itself, although there is no doubt other factors besides instrument factors.

In the case under study, namely the elementary school exam instrument in the Bontomarannu District, Gowa Regency, in the academic year of 2017/2018, it is found that in general, the instruments used are basically good. Only a small number of items detected contain bias, while most others have good quality in terms of bias. It is the basis of evaluation for items that are still biased and retain items that have been assessed as good.

Conclusion and Suggestions

Based on the results of the research conducted, it can be concluded that by using the Generalized Lord Chi-square method showing 20 items of math questions in the school examination in the elementary school of Bontomarannu District, Gowa Regency, in the academic year of 2017/2018, two items contain detected bias (DIF), which are item 5 and item 11. Meanwhile, 18 other items such as Items 1, 2, 3, 4, 6, 7, 8, 9, 10, 12, 13, 14, 15, 16, 17, 18, 19, 20 are not detected as bias items.

Based on the conclusions obtained in this study, some suggestions are proposed. (1) To the developer of the School Examination in the State Elementary School of Bontomarannu District, Gowa Regency: to always evaluate the test instruments used every year by biased detection of items so that the quality of the instruments used annually can be guaranteed. It is considered crucial because basically, the right measurement results will produce objective decisions. (2) To analyze bias items, it is recommended to use the Generalized Lord Chi-square method, because by using this method, we can detect DIF to more focal groups so that the sub-groups analyzed can be more complex. (3) The application of the Generalized Lord Chi-square method in this study only reaches the 1-parameter model so that it is suggested to be further analyzed in the next parameter (2-parameters and 3-parameters). (4) In conducting the next research, it is recommended to use more than one method to compare accuracy.

References

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

Bares, C., Andrade, F., Delva, J., Grogan-Kaylor, A., & Kamata, A. (2012). Differential item functioning due to gender between depression and anxiety items among Chilean adolescents. *International Journal of Social Psychiatry*, *58*(4), 386–392. https://doi.org/10.1177/0020764011400999

Budiyono, B. (2009). Ketepatan metode Mantel-Haenszal, Sibtest, dan Regresi Logistik untuk mendeteksi differential item function. *Jurnal Penelitian Dan Evaluasi Pendidikan, 13*(1), 1–20. https://doi.org/10.21831/pep.v13i1.1398

Chiesi, F., Ciancaleoni, M., Galli, S., Morsanyi, K., & Primi, C. (2012). Item response theory analysis and differential item functioning across age, gender and country of a short form of the Advanced Progressive Matrices. *Learning and Individual Differences*, *22*(3), 390–396. https://doi.org/10.1016/j.lindif.2011.12.007

Cho, Y. I., Martin, M. J., Conger, R. D., & Widaman, K. F. (2010). Differential item functioning on antisocial behavior scale items for adolescents and young adults from single-parent and two-parent families. *Journal of Psychopathology and Behavioral Assessment*, *32*(2), 157–168. https://doi.org/10.1007/s10862-009-9145-1

Colom, R., & García-López, O. (2002). Sex differences in fluid intelligence among high school graduates. *Personality and Individual Differences*, *32*(3), 445–451. https://doi.org/10.1016/S0191-8869(01)00040-X

De Leo, J. A., Van Dam, N. T., Hobkirk, A. L., & Earleywine, M. (2011). Examining bias in the impulsive sensation seeking (ImpSS) Scale using Differential Item Functioning (DIF) – An item response analysis. *Personality and Individual Differences*, *50*(5), 570–576. https://doi.org/10.1016/j.paid.2010.11.030

Herwin, H. (2016). An application of the generalized logistic regression method in identifying DIF. In *Proceeding of the International Conference on Educational Research and Evaluation (ICERE).* Yogyakarta: Universitas Negeri Yogyakarta.

Jodoin, M. G., & Gierl, M. J. (2001). Evaluating type I error and power rates using an effect size measure with the logistic regression procedure for DIF detection. *Applied Measurement in Education*, *14*(4), 329–349. https://doi.org/10.1207/S15324818AME1404_2

Karami, H. (2012). An introduction to differential item functioning. *The International Journal of Educational and Psychological Assessment*, *11*(2), 59–76.

Kim, S.-H., Cohen, A. S., & Park, T.-H. (1995). Detection of differential item functioning in multiple groups. *Journal of Educational Measurement*, *32*(3), 261–276. https://doi.org/10.2307/1435297

Lautamo, T., Laakso, M.-L., Aro, T., Ahonen, T., & Törmäkangas, K. (2011). Validity of the play assessment for group settings: An evaluation of differential item functioning between children with specific language impairment and typically developing peers. *Australian Occupational Therapy Journal*, *58*(4), 222–230. https://doi.org/10.1111/j.1440-1630.2011.00941.x

Magis, D., Béland, S., Tuerlinckx, F., & De Boeck, P. (2010). A general framework and an R package for the detection of dichotomous differential item functioning. *Behavior Research Methods*,

42(3), 847–862. https://doi.org/10.3758/BRM.42.3.847

Osterlind, S. J. (1983). *Test item bias. Series: Quantitative applications in the social sciences.* Beverly Hills, CA: Sage.

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. https://doi.org/10.21831/pep.v17i2.1700

Retnawati, H. (2014). *Teori respons butir dan penerapannya: Untuk peneliti, praktisi pengukuran dan pengujian, mahasiswa pascasarjana.* Yogyakarta: Nuha Medika.

Scheuneman, J. D., & Bleistein, C. A. (1999). Item bias. In G. N. Masters & J. P. Keeves (Eds.), *Advances in Measurement in Educational Research and Assessment* (pp. 220–234). New York, NY: Elsevier.

Weisscher, N., Glas, C. A., Vermeulen, M., & De Haan, R. J. (2010). The use of an item response theory-based disability item bank across diseases: Accounting for differential item functioning. *Journal of Clinical Epidemiology*, *63*(5), 543–549. https://doi.org/10.1016/j.jclinepi.2009.07.016

Zhang, Y. (2015). Multiple ways to detect differential item functioning in SAS. In *Proceedings of SAS Global Forum 2015 Conference* (pp. 1–9). Dallas, TX: SAS Global Forum. Retrieved from https://pdfs.semanticscholar.org/01f8/0a01287893f8f3f1029aa817b9cce3983901.pdf

Zumbo, B. D. (1999). *A handbook on the theory and methods of differential item functioning (DIF).* Ottawa, Canada: Directorate of Human Resources Research and Evaluation.