MaPan : Jurnal Matematika dan Pembelajaran
p-ISSN: 2354-6883 ; e-ISSN: 2581-172X
Volume 7 No 1, June 2019 (16-31)
DOI: https://doi.org/10.24252/mapan.2019v7n1a2

A COMPARISON OF MANTEL-HAENSZEL AND STANDARDIZATION METHODS: DETECTING DIFFERENTIAL ITEM FUNCTIONING

Ahmad Rustam[1)], Dali Santun Naga[2)], Yetti Supriyati[3)]
[1, 2]Universitas Negeri Jakarta, [3]Universitas Tarumanagara
[1, 2]Jl. Rawamangun Muka RT 11/RW 14, Jakarta Timur
[3]Jl. Letjen S. Parman No. 1, Jakarta Barat
E-mail: ahmadrustam_pep16s3@mahasiswa.unj.ac.id[1)], dalinaga@gmail.com[2)], yetti.supriyati@unj.ac.id[3)]

Submitted: 22-04-2019, Revised: 14-05-2019, Accepted: 18-06-2019

Abstract:
The purpose of this study was to review the sensitivity of the two methods, the Mantel-Haenszel (MH and the Standardization methods to detect differences in function items (DIF). Sensitivity was based on the number of DIF grains. The data used in this study were generation data using the Wingen3 program in the form of a response dichotomy of 3054. Sample size was (200 and 1000) responses for the reference group and (200 and 1000) responses for the focus group. Samples were taken randomly as many as 35 replications. The distribution of the ability of the two groups was normal with average and variance, 0 and 1 respectively. The results of the study indicated that MH method were more sensitive than standardization method in DIF detection for samples of 400 and 2000. The finding also assumed there were possibility that standardization method was supreme when using a small sample or the number of population members of the focus group and reference was not balanced, while the focus group was less than the reference group.

Keywords: *Mantel-Haenszel, Standardization, DIF, Sensitivity*

PERBANDINGAN METODE MANTEL-HAENSZEL DAN METODE STANDARISASI: MENDETEKSI PERBEDAAN FUNGSI BUTIR

Abstrak:
Tujuan penelitian ini untuk meninjau sensitivitas dua metode yaitu metode *Mantel-Haenszel* (MH) dan metode Standarisasi dalam deteksi perbedaan fungsi butir atau *Differential item functioning* (DIF). Sensitivitas ditinjau dari banyaknya butir DIF. Data yang digunakan dalam penelitian ini adalah data generasi dengan menggunakan program Wingen3 yang berbentuk respons dikotomi sebanyak 3054. Ukuran sampel (200 and 1000) respons untuk kelompok referensi dan (200 and 1000) respons untuk kelompok fokus. Sampel diambil secara acak sebanyak 35 replikasi. Distribusi kemampuan kedua kelompok adalah distribusi normal dengan rata-rata dan varians yaitu 0 dan 1. Hasil penelitian menunjukkan bahwa metode MH lebih sensitif dari pada metode standarisasi dalam deteksi DIF untuk sampel 400 maupun sampel 2000. Dari hasil penelitian ini ditemukan bahwa ada kemungkinan metode standarisasi lebih unggul ketika menggunakan sampel yang kecil atau jumlah anggota populasi kelompok fokus dan referensi tidak seimbang, dimana kelompok fokus lebih sedikit dibandingkan kelompok referensi.

Kata Kunci: Mantel-Haenszel, Standardization, DIF, Sensitivitas

How to cite: Rustam, A., Naga, D. S, & Supriyati, Y. (2019). A comparison of Mantel-Haenszel and standardization methods: Detecting differential item functioning. *MaPan: Jurnal Matematika dan Pembelajaran*, *7*(1), 16-31.

INTRODUCTION

Good test items will provide accurate information or test results. When the item is not functioning properly, the results to be described are certainly not good. One of the factors that are not good is the imbalance in the distribution of correct answers among different groups of test takers. So, the results obtained are not accurate in describing the true abilities of students. Good test items will provide accurate information or test results. When the item is not functioning properly, the results are certainly not good. One of the factors that are not good is the imbalance of different answers to different groups of test takers. So, the results are not describing the true abilities of students. Detection of DIF, is an attempt to find out whether a test item acts fairly or unfairly against several different groups.

A good item is being able to provide accurate information, in other words, that the item does not benefit one particular group, so that the item is accurate in retrieving data on the ability of the respondent. Hambleton, Swaminathan, & Rogers (1991), an item is said to be DIF, when several individuals from different groups have the same ability but do not have the same possibility of answering the items correctly. In addition, DIF can also be interpreted as a difference that is not expected among several groups of exams which should have comparable test results based on the attributes measured by the items in the tests performed (Wiberg, 2007). It is said that DIF where examinees from different groups also have different possibilities in answering a test item after all, abilities are controlled (Gierl, Khalid, & Boughton, 1999). Furthermore, DIF is defined as different probabilities of examinees from different groups but with the same ability to respond correctly to items (Ong, 2010).

There are two forms of DIF, namely DIF Uniform and Non-uniform DIF. Uniform DIF occurs when the probability of answering an item correctly is consistently higher for one group than for another group at all ability levels. This is indicated by the presence of two parallel Characteristics Curve (ICC). In this case, there is no interaction between ability level and group membership

(Mellenbergh, 1982). Non-uniform DIF occurs when the probability difference for answering items correctly differs in different directions for different levels of ability for different groups. Two intersecting ICCs indicate this. In this case, there is an interaction between the level of ability and group membership.

The causes of the occurrence of biased items in the implementation of the test are differences in race, gender, region, culture, and ethnicity. This is also confirmed by Jensen (1980) that appearing grain bias occurs due to race and sex factors (Berk, 1982; Hulin, Drasgow, & Parsons, 1983).

The values obtained from the test results are presented to provide information about the magnitude or dimensions measured by the test. Sometimes scores on test results do not provide accurate information about test takers. The test package can not function properly due to the different functions of the items.

Based on this description, the ability of the detection method to check whether there is a DIF on each item test, it is desirable to conduct a measurement process to be carried out so that injustice or loss in certain groups can be avoided and the students' abilities are measured objectively.

In 1959 Mantel and Haenszel presented a model for a group matching study. Based on the results of this study, Holland & Thayer (1988) used it for DIF detection and subsequently, known as the Mantel-Haenszel (MH) method.

The use of the Mantel-Haenszel method is based on the assumptions that the ability of the test participant is expressed in the total score obtained by the test taker from all the test items assuming that each test item has the same weight. Also, the level of ability of test participants can be classified into following group M, and for each test, the participant can be grouped into two groups, namely focus and reference groups (Budiyono, 2005). The Mantel-Haenszel method is used to determine whether there is a DIF present in a population. Jensen (1980) emphasized that emerging DIF happened due to race and sex.

Dorans & Kulick (1986) offer another method for identification of DIF called the Standardization method or standardization method using the same information as done in the Mantel-Haenszel procedure.

It should be noted that all methods available for identification of DIF are designed to match groups either directly or indirectly, for abilities measured by items so that that group differences can be observed. Also, all the methods and techniques that have been developed to identify DIF have the same assumptions.

The main difference from the two methods, first is that in each matching variable dimension interval, standardization or standardization method considers the difference in the proportion value (P) for the focus group and the reference group. Second, standardization methods weigh differences about specifically identified standardization groups; such identification is typically a focus group (Masters & Keeves, 1999).

Based on the description above, the detection of DIF in the classical method can be done in two ways, the coat-Haenszel method and the standardization method. Thus, the purpose of this study is to review the sensitivity of two methods, namely the Mantel-Haenszel method and the Standardization method in DIF detection.

The Mantel-Haenszel method and standardization method are two methods which are new in the classical test theory used for DIF detection. The two methods are based on the same assumption that is using raw scores or row scores without estimating responses, so it needs an assessment to explore the differences from the two methods. The results of this study provide information for test developers, to determine the most appropriate method for detecting DIF test items.

Mantel-Haenszel method

At the end of 1979, Scheuneman published methods that were (similar but not identical) Chi-square to evaluate DIF, but this method was subsequently criticized by (Baker, 1981) because it produced values that were influenced by sample size and distribution of unknown sampling. The Chi-square procedure requires groups to be divided into total score intervals and form a 2x2 table for each interval that shows a pass-fail on one axis and two groups on another (Holland & Wainer, 1993). Nine years after the development by Scheuneman, Holland & Thayer (1988) outlined the procedure for investigating DIF using a technique designed by Mantel-Haenszel in a retrospective study they performed on the disease.

In the use of the Mantel-Haenszel method, the test participants in each group (focus group and reference group) were classified into M categories based on the level of ability of the test participants. The ability of these test participants is called the matching variable, which is the variable used as the basis for matching (Holland & Thayer, 1988). In the Mantel-Haenszel method, the ability of the test participants is represented by the total score of the test takers. The data used in the Mantel-Haenszel method is data in 2x2 contingency tables as

many as M pieces of data in a sizeable 2x2xM contingency table, with M being the number of classifications based on the level of ability of the test participants. The 2x2 contingency table shape is shown in table 1.

Tabel 1. 2x2 Contingency Tables for Specific Grains at the Mth Capability Level

Group	Number of Participants Test Right	Number of Wrong Test Participants	Overall Number of Test Participants
Focus (f)	R_{fm}	W_{fm}	N_{fm}
Reference (r)	R_{rm}	W_{Rm}	N_{rm}
Total group (t)	R_{tm}	W_{tm}	N_{tm}

In table 1, the row contains the number of parties examined for reference groups and focus groups, while the column includes the number of parties examined for the correct and wrong responses to the items. Meanwhile, for each table refers to the specific value of matching variable m. The groups of parties examined for various scores that are the same as m are referred to like groups with equal scores.

Mantel and Haenszel provide estimates for the common odds ratio as follows (Neil J Dorans & Holland, 1992; Paul W Holland & Thayer, 1988),

$$\hat{\alpha}_{MH} = \frac{\sum_m \frac{R_{rm}W_{fm}}{N_{tm}}}{\sum_m \frac{R_{fm}W_{rm}}{N_{tm}}} \tag{1}$$

If $\hat{\alpha}_{MH} > 1$, then the investigated item is affected by DIF which benefits the reference group. If $\hat{\alpha}_{MH} < 1$, then the items investigated are affected by DIF which benefits the focus group.

Test the significance of the null hypothesis H_0: $\alpha_m = 1$, for each m, use the chi-square test statistics as follows (Dorans & Holland, 1992; Holland & Thayer, 1988)

$$MH_{\chi^2} = \frac{\left[\left|\sum_m R_{rm} - \sum_m E(R_{rm})\right| - 0{,}5\right]^2}{\sum_m Var(R_{rm})} \tag{2}$$

With

$$E(R_{rm}) = E(R_{rm}|\alpha = 1) = \frac{N_{rm}R_{tm}}{N_{tm}} \tag{3}$$

$$Var(R_{rm}) = Var(R_{rm}|\alpha = 1) = \frac{N_{rm}R_{tm}N_{fm}W_{tm}}{N_{tm}^2(N_{tm}-1)} \quad (4)$$

Test statistics $MH_{\chi^2_{obs}}$ The test statistics in equation (5) are distributed according to the square distribution with the degree of freedom 1, if H0 is true. The decision criteria are as follows. If, then the items examined were statistically significantly detected by DIF.

Standardization Method

In the early eighties (80s) Dorans & Kulick (1983) developed a standardized approach after consulting Holland. The formula developed is standardization (Dorans & Holland, 1992; Dorans & Kulick, 2006). The concept of standardization methods on items that indicate DIF is the ability of the same test participant or the same score, but different in responding to an item.

DIF detection by standardization method is similar to other methods. Namely, the population is divided into two subpopulations, reference sub-population, and focus sub-population. In the same score, the proportions of the correct reference sub-population and sub-population focus are calculated. Illustrations of determining the same score from both groups for one item are shown in table 2.

Table 2. Illustration Tables Determine the Same Score for Standardization Methods

Score	Sub-populasi Referensi	Sub-populasi Fokus
A_1 Score	Reference Member	Focus member
A_2 Score	Reference Member	Focus member
A_i Score	Reference Member	Focus member

In table 2, the Ai score shows the same total score. Furthermore, the number of respondents in the reference sub-population at the same score stated the mR, and the number of respondents in the sub-population focused on the same score stated mF. Calculation of the proportion of the correct answers of the two subpopulations on the same score as follows,

$$PR = mR / MR, \text{ dan } PF = mF / MF \quad (5)$$

Where MR and MF each represent the number of respondents in the reference sub-group and focus sub-group, for each of the same scores (A_i).

The difference in the proportion of the reference sub-population and focus sub-population is used as a benchmark for determining the DIF whether or not the item is

$$D = PF - PR \quad (6)$$

Furthermore, the value of standardization (P_D) is formulated as follows,

$$P_D = \frac{\sum_{i=1}^{A} Dm_{iF}}{\sum_{i=1}^{A} m_{iF}} \quad (7)$$

Where miF is the number of sub-group members focused on the Ai score. DIF determination criteria for grains, if the P_D value is more than 0.1 or less than -0.1 (Dorans, 1989; Dorans, Schmitt, & Bleistein, 1988; Muniz, Hambleton, & Xing, 2001). The greater the D, the higher the difference between the two sub-populations, so the more significant the P_D, the more DIF the item (Naga, 1992).

Sample Size

The sample size is one of the essential considerations in item analysis. Various research results have been carried out using a variety of sample sizes.

The results of Spray's (1989) study of simulation studies using the Mantel-Haenszel method and standardization methods suggest that good sample size for both methods is more than 250 sample sizes. Also, (Fidalgo, Ferreres, & Muñiz, 2004) used the Mantel-Haenszel method with sample sizes varying for each reference group (R) and focus group (F), including 4000 responses (3000 references; 1000 focus) and 750 (500 references; 250 focus). Likewise, Swaminathan & Rogers (1990) used a large sample in DIF detection, where there were two variations in a sample size of 250 for each group and 500 for each group.

Fidalgo et al. (2004) compared the Mantel-Haenszel and SIBTEST method using two sample variations namely 4000 (NR = 3000; NF = 1000) and 750 (NR = 500; NF = 250). Dorans & Kulick (2006) DIF detection applied the standardization method and the Mantel-Haenszel Method using two sample sizes of 1546 responses (NR = 891; NF = 655).

Narayanan & Swaminathan (1996) simulations with reference group sample sizes (500 and 1000) and sample size of focus groups (200 and 500) results showed that when the DIF items were removed and matched the total scores of the two groups, it was found that Type I errors were maintained in two conditions: (a) the first condition, the sample size of the two balanced groups

500 and (b) the second condition, the reference group 1000 and the focus group 200.

In addition to the results of these studies, the results of the Guler & Penfield (2009) study also show that sample sizes 300 and 1,000 can be used for DIF detection using the Mantel-Haenszel method.

Based on various research results presented, the tendency of researchers to use sample sizes above 200, then in this study used two different sample size variations from the results of previous studies designed with a balanced sample size of 400 (NR = 200; NF = 200) and 2000 (NR = 1000; NF = 1000).

RESEARCH METHOD

The research method used was experimental approach with treatment design. The research variables consisted of independent variables and dependent variables. The dependent variable in this study was the number of DIF items. Whereas, the independent variable was the DIF detection method consisting of the Mantel-Haenszel method and the standardization method. DIF detection using the Mantel-Haenszel method was supported by SPSS program, while the standardization method used Microsoft Excel-based AR-DIF program developed by researchers.

Data

The data were generated using the Wingen program 3. Total number of responses were 3054 which consisted of 1527 male students as reference group and 1527 female as focus group responses. The research data was the responses in the form of scores where it was "0" and "1" with the length of the test or the number of items in the test was 40 items. Determined ability (θ) of the two groups was normally distributed with the average is 0, and the variance was 1.

Population and Samples

The population in this study are responses in the form of "0" and "1". The response population of this study was 3,054 test participants. From this population, random response samples were taken based on reference groups and focus groups. For sample size 400 consisted (NR = 200; NF = 200) response and sample size 2000 (NR = 1000; NF = 1000) response. The number of randomized replications or sampling carried out in this study was 35 replications for each group and the number of responses.

Research procedure

a. There are several procedures carried out in this study after generating research data in the form of zero responses "0" and one "1", namely as follows:
b. From the response population, 35 random response samples (replication) were drawn using the SPSS program for each reference group and focus group. The samples taken were as many as (200 and 1000) responses of the reference group and (200 and 1000) responses for the focus group.
c. Unidimensional analysis as a condition, to find out items measuring one dimension. The analysis uses factor analysis with the help of SPSS.
d. Analysis of DIF grain detection using the Mantel-Haenszel method and standardization method, first preparing data from each reference group and focusing. The Mantel-Haenszel method was calculated with the help of the SPSS program and interpreted. Then, the standardization method is calculated using the AR-DIF program, which was designed by Microsoft Excel-based researchers.

Data analysis technique

Before conducting parametric inferential statistical analysis to test the difference in the two averages using the t-test, first examine the results of several prerequisite tests, namely the data normality test and homogeneity of variance using the Levene test.

The prerequisite test aimed to find out the DIF detection data from a normally distributed population. The data used are DIF detection results using the Mantel-Haenszel method and standardization method. The sample normality test uses the Kolmogorov-Smirnov analysis, at a significance level of $\alpha = 0.05$.

RESEARCH RESULTS AND DISCUSSION

Descriptive Data on DIF Detection Results

DIF detection analysis using two methods, the Mantel-Haenszel method and the Standardization method with 400 response sample sizes (NR = 200; NF = 200) and 2000 (NR = 1000; NF = 1000). DIF detection results for a sample of 400 responses, descriptively shown in table 3,

Table 3. Description of DIF Item Data MH Method and Standardization Method for Sample 400

No.	Statistics	Method	
		MH	**Standardization**
1	Mean	20.26	18.80
2	Median	21.00	19.00
3	Mode	21.00	18.00
4	Standard Deviation	2.17	2.03
5	Variance	4.73	4.11
6	Minimum	16.00	14.00
7	Maximum	24.00	22.00
8	Total	709.00	658.00

The results of the analysis in table 3 for the Mantel-Haenszel method show that the highest number of grains detected by DIF for all replications is 24 DIF, while the lowest is 16 DIF points. If taken the average of DIF detection for all replications is 20.26, and for all replications, the most DIF detection is 21 items obtained from the mode value. Also, the variance is 4.73. This illustrates that the DIF variation in data items is not much different while the standardization method for the number of grains detected by DIF for all replications was 22 DIF, while the lowest was 14 DIF. If taken from the average of DIF grain detection for all replications of 18.80, and for all replications, the most DIF detection was 19 items obtained from the mode value. The variance of DIF detection data of 4.11, indicates that the variation in data items that DIF is not so much different.

The results of DIF detection using the Mantel-Haenszel method on sample size 2000, are descriptively shown in table 4,

Table 4. Description of DIF Item Data MH Sample 2000 Method

No.	Statistics	Method	
		MH	**Standardization**
1	Mean	30.43	19.26
2	Median	30.00	19.00
3	Mode	30.00	19.00
4	Standard Deviation	0.85	1.09
5	Variance	0.72	1.20
6	Minimum	29.00	17.00
7	Maximum	32.00	22.00
8	Total	1065.00	674.00

Description of DIF grain data from MH sample 2000 method in table 4, found that for the Mantel-Haenszel method the number of grains detected by DIF for all replications was 32 DIF, and the lowest was 29 DIF. If taken the average of DIF detection for all replications is 30.43, and for all replications, the most DIF detection is 30 items obtained from the mode value. The variance value of DIF detection data is 0.72. This illustrates that there are not many variations in the number of DIF grains. Whereas, the standardized method is that the highest number of grains detected by DIF for all replications is 22 DIF, while the lowest is 17 DIF points. If taken the average of DIF detection for all replications is 19.26, and for all replications, the most DIF detection is 19 items which are seen in the mode value. The variance value is 1.20, which illustrates not much data variation.

The results of hypothesis testing based on inferential statistical analysis using a test of the difference of two independent samples for each hypothesis is shown in the following SPSS output.

Table 5. SPPSS Output for Inferential Statistical Analysis

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
		F	Sig.	t	Df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
									Lower	Upper
Sampel_400	Equal variances assumed	.487	.488	2.901	68	.005	1.45714	.50234	.45475	2.45954
	Equal variances not assumed			2.901	67.666	.005	1.45714	.50234	.45466	2.45963
Sampel_2000	Equal variances assumed	.980	.326	47.705	68	.000	11.17143	.23418	10.7041	11.6387

	Equal variances not assumed	47.705	64.092	.000	11.17143	.23418	10.7036	11.6392

a. The Mantel-Haenszel method is more sensitive than the Standardization method in DIF detection for a sample of 400

The results of the test obtained from the value of 2.901 and the probability value (sig.) Are smaller than α = 0.05 which is 0.005, meaning that the hypothesis H0 is rejected. Thus, the Mantel-Haenszel Method is more sensitive than the standardization method in DIF detection for a sample of 400.

b. The Mantel-Haenszel method is more sensitive than the Standardization method in DIF detection for a sample of 2000

The results of the test analysis obtained obtained t value of 47,705 and the probability value (sig.) Is smaller than α = 0.05 which is 0,000, meaning the hypothesis H0 is rejected. Thus, the Mantel-Haenszel method is more sensitive than the standardization method in DIF detection for the 2000 sample.

Discussion

This study identified the sensitivity of the DIF detection method. The sensitivity referred to here is that many items are detected as DIF by a method compared to other methods. The more grains DIF detects by a method, the more sensitive or sensitive the method is.

The results of the analysis show that the Mantel-Haenszel method is more sensitive than the standardization method in DIF detection for sample 400. The results of the Mantel-Haenszel method analysis for 400 sample sizes in several replications more detect DIF items compared to standardization methods. The result is 25 of the 35 replications that the Mantel-Haenszel method outperforms the standardization method.

The Mantel-Haenszel method is more sensitive than the standardization method in DIF detection. This, it can be explained that in the standardization method for the first replication, it shows that most values of the miF are small, so that it has an impact on the value of standardization so that the item has a high chance of being detected as DIF. On specific points the same score from group pairs, if more focus group members answer correctly, then the number of

focus group members will increase so that when becoming a divider in the standardization formula the results will be small. Thus, the small and large number of items detected by DIF on the Standardization method depends very much on the number of focus sub-population members and also depends on the number of focus sub-population members who answer correctly for the same score. As explained by Dorans & Kulick (2006) that the PD value is very sensitive to the sample, so the increase or size of the sample size will reduce the PD value which means reducing the sensitivity of the standardization method.

The standardized method on certain items, when the number of members focuses on the same score and turns out that few answers correctly, cause the grain to be detected by DIF, and benefit the reference group.

From the results of the analysis, the value of the difference in the large proportion of the two groups, caused the grain to be detected as DIF. The minus PD value indicates that the item benefits the focus group. The first replication mostly benefits focus groups as much as eight items and as many as four items that benefit the reference group.

Judging from the sensitivity averages between the Mantel-Haenszel method and the standardization method show that Mantel-Haenszel is superior to the standardization model for sample size 400, the Mantel-Haenszel method has an average DIF detection rate of 20.26 and the standardization model has an average number DIF detection of 18.80.

The results of the DIF detection analysis in the Mantel-Haenszel method are superior to the standardized method in the 2000 sample size, which shows that from 35 the Mantel-Haenszel replication method outperformed the standardization method in detecting DIF grains. After calculating the average, the Mantel-Haenszel method is higher than the standardization method. The Mantel-Haenszel method average value is 30.43, while the average standardization method is 19.26.

The sample size in the two DIF detection methods is very influential, based on the results of this study indicating that with increasing sample size it shows that the Mantel-Haenszel method is more sensitive or more detects DIF items. In line with the results of the research by Rogers, Swaminathan, & Jane (1993) that for larger sample sizes it would be better to maintain the accuracy of detection compared to small samples.

In the standardization method, according to Masters & Keeves (1999) that this method considers differences in P values for focus groups and reference groups. Also, standardization methods weigh the differences associated with

specifically identified standardization groups, where the specific identification is usually the focus group. Specific identification of the focus group will make this method less sensitive in detecting the number of DIF items, because the increase in the number of focus members who answer correctly, the meal minimizes the standardization value which results in less sensitive methods in DIF detection. Little or many points detected by DIF on standardization methods depend heavily on the number of focus sub-population members and also depend on the number of focus sub-population members who answer correctly for the same score.

CONCLUSION

This study concludes that the Mantel-Haenszel method is more sensitive than the standardization method in DIF detection for a sample of 400. Furthermore, the Mantel-Haenszel method is more sensitive than the standardization method in DIF detection for an example of 2000. The results of this study assume that there are possible methods standardization is superior when using a small sample or the number of members of the focus group and reference population is not balanced, where the focus group is less than the reference group.

ACKNOWLEDGEMENTS

Many thanks are presented to my sponsorship, Indonesia Endowment Fund for Education/LPDP and the BUDI-DN KEMRISTEKDIKTI which funds my research.

REFERENCES

Baker, F. B. (1981). A criticism of scheuneman's item bias technique. *Journal of Educational Measurement*, *18*(1), 59–62.

Berk, R. A. (1982). *Handbook of methods for detecting test bias*. Baltimore, Maryland: The Johns Hopkins University Press.

Budiyono. (2005). *Perbandingan metode Mantel-Haenszel, SIBTEST, regresi logistik dan perbedaan peluang dalam mendeteksi keberadaan DIF*. Universitas Negeri Yogyakarta.

Dorans, N. J. (1989). Applied measurement in education two new approaches to assessing differential item functioning : Standardization and the Mantel-Haenszel method. *Applied Measurement in Education*, *2*(3), 217–233.

Dorans, N. J., & Holland, P. W. (1992). *DIF detection and description: Mantel-Haenszel and Standardization*. New Jersey.

Dorans, N. J., & Kulick, E. (1983). *Assessing Unexpected differential item performance of female candidates on SAT and TSWE forms administered in December 1977: An Application of the Standardization Approach*. New Jersey.

Dorans, N. J., & Kulick, E. (1986). Demonstrating the utility of the Standardization approach to assessing unexpected differential item performance on the scholastic aptitude test. *Journal of Educational Measurement*, *23*(4), 355–368.

Dorans, N. J., & Kulick, E. (2006). Differential Item functioning on the mini-mental state examination state examination: an application of the Mantel-Haenszel and standardization procedures. *Medical Care*, *44*(11), 107–114.

Dorans, N. J., Schmitt, A. P., & Bleistein, C. A. (1988). The standardization approach to assessing differential speededness.

Fidalgo, Á. M., Ferreres, D., & Muñiz, J. (2004). Liberal and Conservative differential item functioning detection using Mantel-Haenszel and SIBTEST : Implications for type I and type II error rates. *Journal of Experimental Education*, *73*(1), 23–39.

Gierl, M., Khalid, S. N., & Boughton, K. (1999). Gender differential item functioning in mathematics and science : Prevalence and policy implications. In *Improving Large-Scale Assessment in Education* (pp. 1–25). Canada: Centre for Research in Applied Measurement and Evaluation University of Alberta Pap.

Guler, N., & Penfield, R. D. (2009). A comparison of the logistic regression and contingency table methods for simultaneous detection of uniform and nonuniform DIF. *Journal of Educational Measurement*, *46*(3), 314–329.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. California: SAGE Publications Inc.

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. Braun (Eds.). In *Test Validity* (pp. 129–145). Erlbaum: Hillsdale, NJ.

Holland, P. W., & Wainer, H. (1993). Differential item functioning. *Psicothema*, 453.

Hulin, C. L., Drasgow, F., & Parsons, C. K. (1983). *Item response theory, application to psychological measurement*. Illinois: Down Jones-Irwin Homewood.

Jensen, A. R. (1980). *Bias in mental testing*. New York: A Division of Macmillan Publishing Co., Inc.

Masters, G. N., & Keeves, J. P. (1999). *Advances in measurement in educational research and assessment*. United Kingdom: Elsevier Science Ltd.

Mellenbergh, G. J. (1982). Contingency table models for assessing item bias, *7*(2), 105–118.

Muniz, J., Hambleton, R. K., & Xing, D. (2001). Small sample studies to detect flaws in item translations. *International Journal of Testing*, *1*(2), 115–135.

Naga, D. S. (1992). *Pengantar teori skor pada pengukuran pendidikan*. Jakarta:

Besbats.

Narayanan, P., & Swaminathan, H. (1996). Identification of Items that show nonuniform DIF. *Applied Psychological Measurement*, *20*(3), 257–274.

Ong, Y. M. (2010). *Understanding differential functioning by gender in mathematics assessment*. University of Manchester for the degree of Doctor of Philosophy.

Rogers, H. J., Swaminathan, H., & Jane, H. (1993). A comparison of logistic regression and Mantel-Haenszel procedures for detecting differential item functioning. *Applied Psychological Measurement*, *17*(2), 105–116.

Spray, J. A. (1989). Performance of three conditional DIF statistics in detecting differential item functioning on simulated tests, (October).

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedure. *Journal of Educational Measurement*, *27*(4), 361–370.

Teresi, J. A., Ramirez, M., Lai, J.-S., & Silver, S. (2008). Occurrences and sources of differential item functioning (DIF) in patient-reported outcome measures: Description of DIF methods, and review of measures of depression, quality of life and general health. *Psychol Sci Q*, *50*(4), 538.

Wiberg, M. (2007). Measuring and detecting differential item functioning in criterion-referenced licensing test: A theoretic comparison of methods. *EM*, (60).

Zumbo, B. D. (1999). *A handbook on the theory and methods of differential item functioning (DIF): Logistic regression modeling as a unitary framework for binary and likert-type (ordinal) item scores*. Canada: Directorate of Human Resources Research and Evaluation, Department of National Defense.