International Journal of Education & Literacy Studies
ISSN: 2202-9478
www.ijels.aiac.org.au

Sensitivity and Accuracy of the Mantel-Haenszel Method and Standardization Method: Detection of Item Functioning Differential

Ahmad Rustam[1]*, Dali Santun Naga[2], Yetti Supriyati[3]

[1]*Posgraduate Program, Universitas Negeri Jakarta, Pulo Gadung, Kota Jakarta Timur, Indonesia 13220*

[2]*Information Technology, Universitas Tarumanagara, Jl. Letjen S. Parman No. 1 Jakarta Barat, Indonesia 11440*

[3]*Educational Research and Evaluation, Universitas Negeri Jakarta, Pulo Gadung, Kota Jakarta Timur, Indonesia 13220*

The research is financed by BUDI-DN. No. FR1182018150300

Corresponding author: Ahmad Rustam, E-mail: ahmadrustam_pep16s3@mahasiswa.unj.ac.id

ARTICLE INFO

Article history
Received: April 15, 2019
Accepted: July 14, 2019
Published: July 31, 2019
Volume: 7 Issue: 3

Conflicts of interest: None
Funding: LPDP and the BUDI-DN KEMRISTEKDIKTI

ABSTRACT

Detection of differential item functioning (DIF) is needed in the development of tests to obtain useful items. The Mantel-Haenszel method and standardization are tools for DIF detection based on classical theory assumptions. The study was conducted to highlight the sensitivity and accuracy between the Mantel-Haenszel method and the standardization method in DIF detection. Simulation design (a) test participants consisted of 1000 responses in both the reference and focus groups, (b) the size of the proportion of DIF (0.1; 0.25; 0.50; and 0.75), and (c) the length of the multiple choice test with 40 choices the answer. Research shows that the Mantel-Haenszel method has the same sensitivity as the standardization method in DIF proportions of 10% and 25%, however, when the ratio of DIF proportions above 25% the standardization method is less sensitive, and conversely the sensitivity of the Mantel-Haenszel method increases. The standardization method has higher accuracy than the Mantel-Haenszel method in the DIF proportion of 10%, however, when the size of the DIF proportion above 10% the accuracy of the standardization method decreases, the accuracy of the Mantel-Haenszel method is higher than the standardization method. Thus, if the ratio of DIF is detected by the standardization method of (≤0.10), then the results of the standardization method are preferred as a reference. Conversely, if the proportion of DIF detected by the standardization method is (≥0.10), then the result of the Mantel-Haenszel method is chosen as a reference.

Key words: Mantel-Haenszel, Standardization, DIF, Sensitivity, Accuracy

INTRODUCTION

The measurement process through tests aims to obtain information about the extent to which students' achievements or abilities are after the learning process. The test package used consists of several items as a test tool to measure students' abilities. It is known that tests can be developed for different purposes in education. The tests developed can be formative and summative tests, class-based tests and others. Test items developed, then presented both in the form of dichotomy and polytomy.

Tests in the form of dichotomies are like multiple choice questions, where the test only consists of two models of answers which are correct with a score of one "1" and wrong with a score of "0" Whereas the test is in the form of a polytomy, which is like a description question, where students' answers are assessed based on the assessment criteria with the specified score range.

The form of the dichotomy test is very widely used as a test tool. This type of analysis is implemented by the Institute for Educational Assessment Centers to measure the ability of students in Indonesia.

A good test will provide information or results that are accurate in describing the actual abilities of students. Of course, a good test has a set of quality items. The development of a good test is inseparable from the item analysis process that truly fulfills and applies the criteria and procedures for developing test items.

Test items are well developed and analyzed, to produce quality test items, so that they can be used as test kits and provide exact information and in accordance with the purpose of the test, for example, being able to provide information about students who have or do not master the material provided.

The main objective of item analysis is to identify the weaknesses of the test as a measuring instrument used. The results of the item test analysis are used to examine and analyze various aspects related to feedback on student learning difficulties. Based on these objectives, the activity of analyz-

Published by Australian International Academic Centre PTY.LTD.

http://dx.doi.org/10.7575/aiac.ijels.v.7n.3p.28

ing test items provides many benefits, namely: 1) being able to help the users of test items in evaluating the tests used; 2) very relevant for the preparation of tests nationally and locally, for example, the test items provided by the teacher for students in the class; 3) help in writing effective test items; 4) content able to improve tests in class; and 5) able to improve grain validity and reliability (Anastasi, 1976; Urbina, 2004).

To obtain accurate information or test results, of course, quality items are needed that have good validity. Some experts argue that the validity of the test is the extent to which the measuring instrument is able to measure what should be measured (Allen & Yen, 1979; Azwar, 2000; Kerlinger, 1986; Nunnally, 1978).

Validity can be grouped into three types namely criterion validity, content validity, and construct validity (Allen & Yen, 1979; Kerlinger, 1986; Nunnally, 1978). From the three types of validity, it can be seen the fact that the validity is grouped into five, namely test content, response process, internal structure, relationships with other variables, and consequences of the test (AERA, APA, & NCME, 1999; Cizek, Rosenberg, & Koons, 2008).

One source of the fact of the validity of the test is the consequence of the test. A test item that is unable to measure properly, gives inaccurate results. Test items that tend to benefit certain groups incorrectly answering an item are items that are not good and DIF identified. Thus, the item provides information that is inaccurate and not in accordance with the purpose of its manufacture, so that the consequences of the item are not in line with expectations. For this reason, the item has a different function and benefits, certain groups. By that, it is necessary to detect the difference in grain function (DIF) which is part of the validity of the test and the consequences of the test.

In 1959 Mantel and Haenszel presented a model for a group matching study. Based on the results of the study, Holland and Thayer (1988) used Mantel-Haenszel 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. In addition, the level of ability of the test participants can be classified into consecutive group M and for each test participant can be grouped into two groups, namely, focus and reference groups.

The Mantel-Haenszel model has an important role in determining the presence or absence of DIF in a population, for example, ethnic groups, gender (male and female) or population determined based on socio-demographic or geographical location. In addition, the Mantel-Haenszel method offered another method for DIF identification called the standardization method using the same information as was done in the Mantel-Haenszel procedure (Dorans & Kulick, 1986).

The aim of this study is to highlight the sensitivity and accuracy between the Mantel-Haenszel method and the standardization method in DIF detection for four DIF proportions (10%, 25%, 50%, and 75%) specified in the hypothesis.

LITERATURE REVIEW

Differential Item Functioning

Good items will provide accurate information, according to the function of the item. Poor items cause measurement bias. Common grain bias is known to be known in the analysis process with differential item functioning (DIF). DIF is a statistical term used to describe situations in which people from one group answer an item correctly more often than people who are equally knowledgeable from another group (Zumbo, 1999).

Item are declared not DIF, if the item can provide accurate information, and the item does not benefit one particular group in answering correctly. Hambleton, Swaminathan, and Rogers (1991) state that DIF items occur when several individuals from different groups have the same ability, however, do not have the same possibility of answering the items correctly. Also, DIF points out if the examinees from different groups also have different possibilities in answering a test item, after all the 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).

The causes of the occurrence of biased items in the implementation of the test are differences in race, gender, region, culture and ethnicity (Berk, 1982; Hulin, Drasgow, & Parsons, 1983). Also, biases or DIF appear on the grain and occur due to race and sex factors (Jensen, 1980).

Scores obtained by participants from the test results will provide information about the magnitude or dimensions measured by the test. However, sometimes the scores from the test results do not provide accurate information about the ability of the test takers, this occurs due to DIF, so it needs a tool or method to analyze carefully by carrying out DIF detection. This analysis is done so that injustice or loss in certain groups can be avoided to obtain information on students' abilities that are measured objectively.

Many methods have been developed. However, test developers need an easy and practical method to use. Two classic methods have advantages, and their use is not so difficult to do, namely the Mantel-Haenszel method and the standardization method. Both the Mantel-Haenszel method and standardization are solutions in the future because this technique requires low cost, practical, statistically good, and this technique is very good in DIF detection (Dorans & Holland, 1992).

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 results of the study (Dorans, 1989) explain the differences between the two methods that the standardization method uses focus group relative frequencies as weights,

while the Mantel-Haenszel method provides significance test results. Also, differences in the two methods, first, for each dimension interval matching the standardized method variable considers the difference in proportion values (P) between focus groups and reference groups. Second, standardization methods weigh differences related to specifically identified standardization groups, and such identification is typically a focus group (Masters & Keeves, 1999).

Both DIF detection methods, the Mantel-Haenszel method, and standardization methods, each have advantages and disadvantages. One of the factors that influence the strength or sensitivity of the two methods is the sample size of the respondents. For this reason, it is necessary to pay attention to the sample size of respondents in identifying the sensitivity of the two DIF detection methods. The results of simulation studies with the proportion of DIF are 20%, 40%, and 60% (Gierl, Gotzmann, & Boughton, 2004; Huggins, 2012). In addition, Hidalgo, Galindo-garre, and Gómez-benito (2015) conducted a study by increasing the size of the DIF proportion of 0%, 10%, 20%, 30% and 40% with a test length of 20 items, so that the number of successive DIF items 0, 2, 4, 6, and 8.

Whitmore and Schumacker (1999) conducted a simulation with DIF size variations of 5% and 15%. The latest research by increasing the proportion of smaller DIF sizes is 0%, 15%, and 30% (Oliveri, Ercikan, & Zumbo, 2014). Various results of previous studies that have been carried out with varying DIF proportions, after analysis the researchers determined four measures of DIF proportions, namely 10%, 25%, 50%, and 75%.

Mantel-Haenszel Method

The application of the Mantel-Haenszel method that is the test participant in each focus group and the reference group is made into M categories, based on the level of ability of the test participants. In the Mantel-Haenszel Method, the ability of the test participants is used, namely the total score data, which is then designed in a 2x2 contingency table of M pieces. Where M is the number of classifications based on the level of ability of the test participants. The form of the 2x2 contingency table is shown below:

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 different scores that are the same as m are referred to like groups with equal scores.

The parameter α is called the common odds ratio for every 2x2 contingency table, the value of α is the common odds for each m, i.e.:

$$\alpha_m = \frac{\frac{R_{rm}}{W_{rm}}}{\frac{R_{fm}}{W_{fm}}} = \frac{R_{rm}W_{fm}}{R_{fm}W_{rm}} \quad (1)$$

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

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

With

$$E\left(R_{rm}\right) = E\left(R_{rm}\,|\alpha = 1\right) = \frac{N_{rm}R_{tm}}{N_{tm}} \quad (3)$$

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

Test statistics $MH_{\chi^2_{obs}}$ in equation (2) is distributed according to chi square distribution with 1 degree of freedom, if H_0 right. The decision criteria are as follows. If $MH_{\chi^2_{obs}} > \chi^2_{\alpha;1}$, then the items examined were statistically significantly detected by DIF.

Standardization Method

The concept of the Standardization method on items that indicate DIF is the ability of the same test participant or the same score, but different in responding to an item.

Detection of DIF with standardization method is similar to other methods, namely population divided into two groups namely reference groups and focus groups. In the same score, the proportion of correct reference groups and focus groups was calculated. Illustrations of determining the same score from both groups for one item are shown in Table 2.

In Table 2, the A_i score shows the same total score. Furthermore, the number of respondents in the reference group at the same score stated the mR and the number of respondents in the focus group on the same score was stated as mF. Calculation of the proportions of the correct answers of the two groups on the same score as follows,

$$P_R = mR/MR, \text{ and } P_F = mF/MF \quad (5)$$

Table 1. 2×2 Contingency tables for specific grains at the mth capability level

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

Table 2. Illustration table determining the same score for the standardization method

Score	Reference Group	Focus Group
A_1	Reference Member	Focus member
A_2	Reference Member	Focus member
A_i	Reference Member	Focus member

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

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

$$D = P_F - P_R \quad (6)$$

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. Differential Item Functioning determination criteria for grains, if the PD value is more than 0.1 or less than -0.1 (Neil J Dorans, 1989; Nell J. Dorans, Schmitt, & Bleistein, 1988; Muniz, Hambleton, & Xing, 2001). The greater the D, the greater the difference between the two groups, so the greater the PD, the more DIF the item (Naga, 1992).

Sensitivity

Sensitivity was first introduced by Yerushalmy (1947) on health measurements that sensitivity is the ability to correctly diagnose a person who is sick, meaning that the test results are positive and hurt. This is associated with the measurement that sensitivity is the proportion of DIF positive items in the population and after being identified by the detection method it turns out that the item is DIF. In other words, the sensitivity is the possibility of DIF grains being detected correctly or the probability of each DIF item being identified correctly with the DIF detection method. A correct item is DIF, and after being detected by a certain method the result is positive DIF called True positive. In addition, known type I error rates and type II error rates. Type I error rate is a method of not detecting grains as DIF, but in reality, the item is DIF. Type II error rate is a method of detecting grain as DIF, however, in reality, the item is not DIF.

Loong (2003) explains the sensitivity that sensitivity = TP/TP + FN, where true positive (TP) is the real positive number and false negative (FN) is the number of false negatives. Based on the Loong formula (2003), it can be written the sensitivity formulation referred to in this study, namely,

$$Sensitivity = \frac{\sum true\ positive}{\sum true\ positive + \sum false\ negative} \quad (8)$$

Based on the formula, the DIF detection method, the fewer false negatives detected, the more sensitive the method is. Conversely, a DIF detection method, if more false negatives are detected, the less sensitive the method.

Low sensitivity is caused by the DIF detection method that passes many grains containing DIF. This can be said that a DIF detection method with low sensitivity will increase some false negative (FN) numbers.

Accuracy

Accuracy is the level of accuracy and accuracy in measurement. According to Ercikan, Roth, Simon, Sandilands, and Lyons-Thomas (2014) accuracy is the ability to correctly identify the correct DIF items, while not identifying items that are not DIF.

Simply stated, the purpose of the accuracy analysis is that the DIF detection method is able to detect items that are truly DIF, and able to detect items that are truly non-DIF, so that no detection errors occur.

Formulation of the accuracy value of a DIF detection method (Zhu, Zeng, & Wang, 2010) as follows,

$$Accuracy = \frac{\sum true\ positive + \sum true\ negative}{\sum true\ positive + \sum true\ negative + \sum false\ positive + \sum false\ negative} \quad (9)$$

There are two methods used to detect differences in grain function (DIF), and then tested the sensitivity and accuracy of the method. The two methods used in this research are the Mantel-Haenzel method of standardization method.

METHOD

The research method used is experimental design with treatment design. The research variables consist of independent variables and dependent variables. The dependent variable in this study is the sensitivity value and accuracy value, while the independent variable is the DIF detection method consisting of the Mantel-Haenszel method and the standardization method. DIF detection using the Mantel-Haenszel method uses the help of the SPSS program (Azen & Walker, 2011), while the standardization method uses Microsoft Excel-based AR-DIF programs developed by researchers.

Data

The main data used is the student work result data in the form of a score, in which the form of student score results in the form of responses "0" and "1" with a test length of 40 items. Source of data from the Education Assessment and Education Center of the Ministry of National Education in the form of student national exam results (UN) response data in 2015.

Population and Sample

The population in this study was the response of the 2015 national exam (UN) test participants in Bone and Luwuk Timur districts, Bunggai District with 3,245 test participants.

Based on the response of the test participant, a population estimated with the help of the BILOG program to obtain the distribution value of the test participants' abilities. The ability distribution data is used to determine the value of ability distribution in the data generated with the Wingen3.0 program rock.

The response data sample was determined through data generation with the help of the Wingen3.0 program. The sample size was determined based on the study design, which amounted to 2000 responses consisting of 1000 responses as a reference group (R) or male group and 1000 responses as focus groups (F) or groups of women.

The number of UN test items is 40 items, and there are two groups of test participants, namely male and female groups. The items were detected using two methods, the Mantel Haenszel method and the standardization method.

Research Procedure

There are several procedures carried out in this study after the generation of research data in the form of zero responses "0" and one "1", namely as follows:

Data design and simulation

The research design was carried out by involving two DIF detection methods, a measure of the proportion of DIF. The responses used were 2000 consisting of 1000 reference groups and 1000 focus groups and used dichotomous response data of 40 items with 5 answer choices.

Data generation

This study uses response data in the form of a value of 1 (true) and a value of 0 (wrong) for each item. The number of respondents for the reference group was 1000 responses and the focus group was 1000 responses.

The generation of data is done using the Wingen3.0 program. Wingen3.0 is a response data generation program that can be conditioned based on the needs of research analysis. The generation of data using the Wingen3.0 program was also carried out by (Han, 2007; Oliveri et al., 2014). The response data generation settings are explained in the next procedure.

Estimation of empirical data ability parameters

For each test participant in the reference group or focus group, the latent ability distribution or theta (eta) was determined based on empirical data, the ability data were normally distributed with an average of -0.0123 and the variance was 0.9393, can be written in the form of N (-0.0123, 0.9393). The distribution of capabilities from empirical data is obtained through estimation using the BILOG program, so that the ability data distribution is obtained, namely N (-0.0123, 0.9393).

Item parameters

For each test item, both the reference group and the focus group are determined by each item parameter value, namely the different power parameters (a), difficulty level (b), and guesses (c) based on the item response theory.

Determine DIF and non-DIF items

For each test item the probability of P (θ) is calculated according to the three-parameter logistic model (3PL). Based on data distribution in step (c) and predetermined parameters, the response data in the form of values 0 and 1 are obtained from the results of data generation using Wingen3.0.

In this data generation, the value of the item parameters such as a, b, and c are determined in such a way that there is a DIF charge with the size of the desired DIF proportion. Furthermore, the values of parameters a, b, and c in this data generation is assumed to be the actual parameter values that exist in the population and are used to determine the true DIF.

The generation of data for each item is determined by parameter values a, b, and c. research results of Oliveri et al. (2014) using different power parameter values (b) of 0.6. In addition, the research design of Budiyono (2009) determines the items affected by DIF, the values of the reference group parameters are a = 1.5, b = -0.5, and c = 0.1 and the values of the focus group parameters are a = 1.5, b = 0.5, and c = 0.1. For items not affected by DIF, the reference group parameter values are a = 1.5, b = 0.0, and c = 0.1 and the focus group parameter values are a = 1.5, b = 0.0, and c = 0.1. Based on the values of the item parameters, all true DIFs are designed as DIF. The number of items used is by the empirical data on the SMP UN questions as many as 40 items.

Replication

For the purpose of analyzing the hypothesis of data generation, 30 replications were performed for each treatment.

DIF detection using the Mantel-Haenszel method

Detection of differential item functioning (DIF) using the Mantel-Haenszel method can be done with the help of the SPSS program. The steps to detect DIF with the Mantel-Haenszel method, first is to prepare response data from each reference group and focus groups that have been in the generation. Next, input data in the SPSS worksheet and make groupings in each reference group and focus. After that, it was analyzed in SPSS to obtain DIF detection results using the Mantel-Haenszel method.

DIF detection with the standardization method

DIF detection by standardization method uses a program that was designed by researchers in Microsoft Excel based applications. The same thing with the Mantel-Haenszel method is to prepare each response data from reference groups and focus groups that have been in the generation. Then, input in the AR-DIF program worksheet, after that run the program to obtain DIF and non-DIF items.

Data Analysis Technique

Before conducting parametric inferential statistical analysis, first examine the results of several prerequisite tests, namely the data normality test and homogeneity of variance using the Levene test.

Examination of group differences in terms of data from the analysis of the value of sensitivity and the value of accuracy after detection of two methods consisted of 30 replications. The average difference test used is different test two independent samples or two independent samples t test with a significant level $\alpha = 0.05$, the calculation using SPSS program assistance.

RESULTS AND DISCUSSION

This study identified items that were DIF and not DIF from 40 items with 2000 responses consisting of 1000 reference group responses and 1000 focus group responses. Then, the conditions for measuring the proportion of DIF are determined, namely (10%, 25%, 50%, and 75%). Where 10% consists of 4 DIF and 36 non-DIF grains, 25% consists of 10 DIF grains and 30 non-DIF grains, 50% consists of 20 DIF grains and 20 non-DIF grains, 75% consisting of 30 DIF grains and ten non-DIF items. Furthermore, the DIF detection method uses the Mantel-Haenszel method and standardization method. Following this, the results of the research on DIF grain detection are presented using two classic methods, the Mantel-Haenszel method, and standardization.

Descriptive Data on DIF Item Detection Results

DIF detection analysis uses two methods, namely the Mantel-Haenszel method and the Standardization method with a response sample size of 2000 (NR = 1000; NF = 1000). The results of the sensitivity analysis based on data generation between the Mantel-Haenszel method and standardization obtained the results in Table 3.

In Table 3 it is shown that the sensitivity value of DIF detection of both methods is Mantel-Haenszel and the standardization method is in the proportion of 10% and the proportion of 25% is descriptively the same value. The sensitivity value of DIF detection in proportions above 25% of the two methods is different, descriptively the sensitivity value of DIF detection in the Mantel-Haenszel method is superior to the standardization method.

To see the difference in the results of the calculation of the sensitivity of the two methods simply, it can be seen in the graph in Figure 1.

From the results of the analysis in Figure 1, it can be seen that the sensitivity value of the standardization method is decreasing, when the proportion of DIF is getting bigger.

In addition, the results of the accuracy analysis based on data generation between the Mantel-Haenszel method and Standardization results are obtained in Table 4.

In Table 4, it is shown that the accuracy value of DIF detection of the two methods, namely the Standardization method is more accurate than Mantel-Haenszel at the proportion of 10%. In contrast, the Mantel-Haenszel method is more accurate than Standardization in proportions above 10%.

To see the difference in the results of the calculation of the accuracy of the two methods simply, it can be seen in the graph in Figure 2.

From the analysis in Figure 2, it can be seen that when the size of the DIF proportion is smaller, the accuracy of the P-difference method is more accurate, on the contrary, when the size of the DIF proportion increases, the Mantel-Haenszel method's accuracy is more accurate.

The results of the analysis of the normality of DIF detection data distribution from both methods, the Mantel-Haenszel and Standardization methods, obtained that there is abnormal data distribution. In connection with these results, the central limit theorem assumes that it is not a problem of any population distribution, estimation of the sample will remain equally distributed (normally distributed) that applies to n> 30 (Agresri & Finlay, 2009; Berenson, Levine, & Krehbiel, 2012; Lind, Marchal, & Wather, 2012).

Based on the results of hypothesis testing on the generation of data based on empirical data, it was found that (1.a) there was no difference in sensitivity of the Mantel-Haenszel method with the Standardization method in the normal data distribution and the proportion of DIF 10%, (1.b) Haenszel with the standardization method for normal data distribution and the proportion of DIF 25%, (1.c) the Mantel-Haenszel method is more sensitive than the Standardized method for normal data distribution and proportion of DIF 50%, (1.d) Mantel-Haenszel method is more sensitive than with the Standardization method on normal data distribution

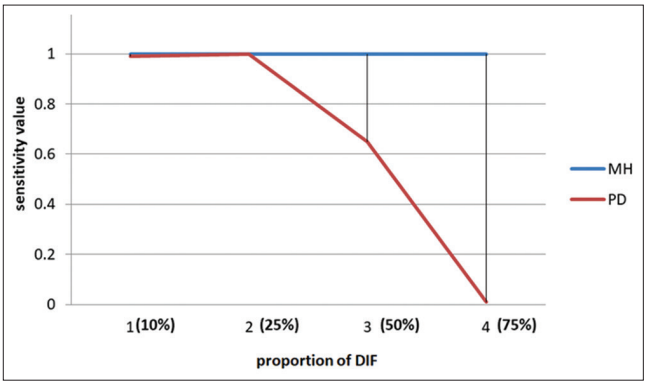

Figure 1. Sensitivity comparison chart between Mantel-Haenszel method and standardization

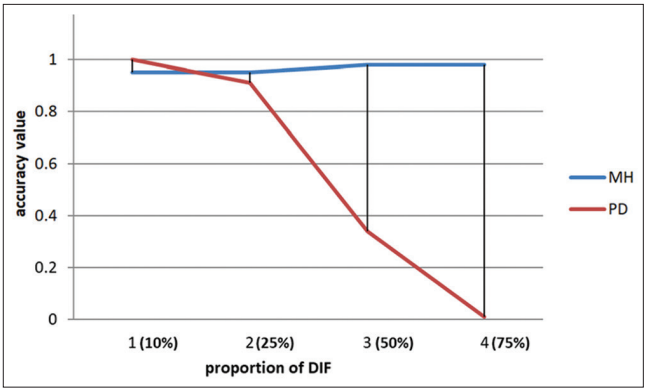

Figure 2. Graph of the accuracy comparison between the Mantel-Haenszel method and standardization

Table 3. Description of data sensitivity results of dif detection method

Replication	DIF Proportion							
	10%		25%		50%		75%	
	Sensitivity		Sensitivity		Sensitivity		Sensitivity	
	MH	PD	MH	PD	MH	PD	MH	PD
1	1.00	1.00	1.00	1.00	1.00	0.60	1.00	0.03
2	1.00	1.00	1.00	1.00	1.00	0.85	1.00	0.00
3	1.00	1.00	1.00	1.00	1.00	0.55	1.00	0.00
4	1.00	1.00	1.00	1.00	1.00	0.60	1.00	0.00
5	1.00	1.00	1.00	1.00	1.00	0.75	1.00	0.00
6	1.00	1.00	1.00	1.00	1.00	0.75	1.00	0.00
7	1.00	1.00	1.00	1.00	1.00	0.55	1.00	0.00
8	1.00	1.00	1.00	1.00	1.00	0.65	1.00	0.00
9	1.00	1.00	1.00	1.00	1.00	0.70	1.00	0.03
10	1.00	1.00	1.00	1.00	1.00	0.55	1.00	0.03
11	1.00	1.00	1.00	1.00	1.00	0.60	1.00	0.00
12	1.00	1.00	1.00	1.00	1.00	0.65	1.00	0.07
13	1.00	1.00	1.00	1.00	1.00	0.60	1.00	0.00
14	1.00	1.00	1.00	1.00	1.00	0.70	1.00	0.00
15	1.00	1.00	1.00	1.00	1.00	0.60	1.00	0.03
16	1.00	1.00	1.00	1.00	1.00	0.60	1.00	0.00
17	1.00	1.00	1.00	1.00	1.00	0.80	1.00	0.03
18	1.00	1.00	1.00	1.00	1.00	0.75	1.00	0.00
19	1.00	1.00	1.00	1.00	1.00	0.65	1.00	0.00
20	1.00	1.00	1.00	1.00	1.00	0.70	1.00	0.00
21	1.00	0.80	1.00	1.00	1.00	0.60	1.00	0.00
22	1.00	1.00	1.00	1.00	1.00	0.70	1.00	0.00
23	1.00	1.00	1.00	1.00	1.00	0.80	1.00	0.00
24	1.00	1.00	1.00	1.00	1.00	0.50	1.00	0.00
25	1.00	1.00	1.00	0.85	1.00	0.60	1.00	0.00
26	1.00	1.00	1.00	1.00	1.00	0.55	1.00	0.00
27	1.00	1.00	1.00	1.00	1.00	0.55	1.00	0.00
28	1.00	1.00	1.00	1.00	1.00	0.70	1.00	0.00
29	1.00	1.00	1.00	1.00	1.00	0.65	1.00	0.00
30	1.00	1.00	1.00	1.00	1.00	0.65	1.00	0.03

and 75% DIF proportion, (2.a) Standardization method is more accurate than the Mantel-Haenszel method in normal data distribution and the proportion of DIF 10%, (2.b) Mantel-Haenszel method is more accurate than with the Standardization method on normal data distribution and the proportion of DIF 25%, (2.c) the Mantel-Haenszel method is more accurate than the Standardized method on normal data distribution and DIF 50% proportion, (2.d) Mantel-Haenszel method is more accurate than Standardization method in normal data distribution and 75% DIF proportion.

DISCUSSION

This study identified the sensitivity and accuracy of two DIF detection methods. The results of this study examine two hypotheses relating to two DIF detection methods with a sample size0 of 2000 respondents and consisting of 4 measures of DIF proportions, namely 10%, 25%, 50%, and 75%.

The Sensitivity of the Mantel-Haenszel method and the Standardization method in DIF Detection for Four Proportions of DIF

From the analysis results obtained that the Mantel-Haenszel method has the same sensitivity as the Standardization method in the DIF proportion of 10% and 25%, but when the size of the DIF proportion increases causing the standardization method to be less sensitive, the Mantel-Haenszel method is increasingly sensitive when the proportion value of DIF above 25%. The Mantel-Haenszel method has a better sensitivity than the Standardization method when the size of the DIF proportion is above 25%. This happens when the pro-

Table 4. Description of data accuracy results of dif detection method

Replication	DIF Proportion							
	10%		25%		50%		75%	
	Accuracy		Accuracy		Accuracy		Accuracy	
	MH	PD	MH	PD	MH	PD	MH	PD
1	0.95	1.00	0.93	0.93	1.00	0.30	0.98	0.03
2	0.93	0.98	0.95	0.85	0.95	0.43	0.98	0.00
3	0.98	1.00	0.95	0.90	1.00	0.28	0.98	0.00
4	0.98	1.00	1.00	0.90	1.00	0.30	1.00	0.00
5	0.90	1.00	1.00	0.85	1.00	0.40	1.00	0.00
6	0.95	0.98	0.95	0.88	1.00	0.38	1.00	0.00
7	0.98	1.00	1.00	0.98	0.98	0.28	0.95	0.00
8	0.88	1.00	0.98	0.95	1.00	0.33	0.98	0.00
9	0.95	1.00	0.95	0.88	0.98	0.38	0.98	0.03
10	0.95	1.00	0.98	0.95	1.00	0.30	0.95	0.03
11	0.98	1.00	0.98	0.88	1.00	0.30	0.98	0.00
12	0.98	1.00	0.98	0.93	0.98	0.33	1.00	0.05
13	0.95	1.00	0.98	0.85	1.00	0.30	0.98	0.00
14	0.88	1.00	0.98	0.95	0.98	0.35	1.00	0.00
15	1.00	1.00	0.98	0.90	1.00	0.35	1.00	0.03
16	0.93	1.00	0.85	0.85	0.98	0.38	1.00	0.00
17	0.93	1.00	0.93	0.88	0.95	0.40	1.00	0.03
18	0.98	1.00	0.98	0.85	1.00	0.38	0.95	0.00
19	0.95	1.00	0.98	0.93	0.98	0.35	0.93	0.00
20	0.95	1.00	0.95	0.93	1.00	0.38	0.93	0.00
21	0.93	1.00	0.93	0.93	1.00	0.33	1.00	0.00
22	0.95	1.00	0.95	0.95	0.98	0.35	0.98	0.00
23	0.98	1.00	0.98	0.90	0.98	0.40	0.98	0.00
24	0.95	1.00	1.00	0.90	0.98	0.28	0.98	0.00
25	0.90	0.98	0.98	0.90	0.98	0.33	1.00	0.00
26	0.90	0.98	0.98	0.90	0.98	0.35	1.00	0.00
27	0.93	1.00	0.98	0.90	0.98	0.30	0.98	0.00
28	1.00	1.00	0.95	0.95	0.95	0.40	1.00	0.00
29	0.93	1.00	0.95	0.95	0.95	0.33	1.00	0.00
30	0.98	1.00	1.00	0.93	0.98	0.35	0.95	0.03

portion of DIF is large, so the disadvantaged group is that the focus group mostly gets low scores, on the other hand, the reference group mostly gets high scores. As a result, the score between focus and reference becomes unbalanced, both from low scores and high scores. The imbalance of scores between the two groups has an impact on the small value of the difference in proportion (D = Pf - Pr), so the items that were originally DIF after detection of standardized methods are not DIF.

Standardization method, when most values are small, it has an impact on the value of standardization (PD), so that the item has a high chance of being detected as DIF. Other than that, on certain points the same score from group pairs, when more focus group members answer correctly, the number of focus group members will increase, so that when becoming a divider in the standardized results formula or the PD value will be smaller. Thus, the small and large number of items detected by DIF on standardization methods are very dependent 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 and Kulick (2006), the value of PD is very sensitive to the sample so that the increase or magnitude of the sample size will reduce the value of PD which means reducing the sensitivity of the Standardization method.

The sample size in both DIF detection methods also influences DIF detection. Based on the results of this study, it was shown that by increasing the number of focus members who answered correctly on the same score as the reference group, the chance of the Mantel-Haenszel method was greater and more sensitive in detecting DIF and non-DIF items.

Likewise, the standardization method that sample size is a problem when the sample used is small. Therefore, a sample of 2000 responses is not sufficient for analysis of standardization methods. Also, a large proportion of DIF size causes standardization methods to be less sensitive.

In the Standardization method, according to Masters and Keeves (1999) that this method considers differences in P values for focus groups and reference groups. Also, standardization methods weigh differences related to specifically identified standardization groups. Specific identification, typically a focus group, so that with special identification of focus groups makes this method less sensitive in DIF grain detection, because the increase in the focus group that answers correctly to the same score as the reference group will result in the standardization (PD) being low, consequently standardization methods become less sensitive.

DIF items if the same ability or score from different groups do not have the same opportunity to answer the item correctly so that the item benefits one group.

The sensitivity of the standardization method is strongly influenced by the response sample size. If more responses are analyzed, the greater the chances of focus groups and groups having the same score. For this reason, the size of the response sample plays a major role in DIF detection using the standardization method. This is in line with the results of the study (Dorans & Kulick, 1986: 366) found that the weaknesses of standardization methods require relatively large samples.

Accuracy of the Mantel-Haenszel method and Standardization method in DIF Detection for Four Proportions of DIF

The accuracy analysis of the DIF detection method is strongly influenced by the strength of the sensitivity of the method. Also, accuracy also takes into account the detection results of non-DIF items correctly detected.

From the analysis results obtained that the Standardization method has higher accuracy than the Mantel-Haenszel method in the DIF proportion of 10%, but when the size of the DIF proportion increases ie, above 10% the accuracy of the standardization method decreases, so the Mantel-Haenszel method accuracy is higher than the method standardization.

The accuracy of the standardization method is higher than the Mantel-Haenszel method when the size of the proportion of DIF is 10%. This is due to the small proportion of DIF, so the disadvantaged group is the focus group on the number of members of the two focus groups and references that have equal scores.

The accuracy of the standardization method is getting lower when the size of the DIF proportion is getting bigger. This is caused by an imbalance between members of the reference group and the focus on the same score. In other words, the chances of focus members having a higher score are getting smaller, on the other hand, members of the reference group have a high chance of getting a high score. The imbalance of this score allows error detection of both DIF and non-DIF items, so the actual item is non-DIF, however, it is detected as DIF. Conversely, the actual grain is DIF detected as non-DIF.

CONCLUSION

The Mantel-Haenszel method has the same sensitivity as the Standardization method in DIF proportions of 10% and 25%, but when the size of the DIF proportion increases causing the standardization method to be less sensitive, the Mantel-Haenszel method is more sensitive when the DIF proportion is above 25%.

The Standardization method has higher accuracy than the Mantel-Haenszel method in the DIF proportion of 10%, but when the size of the DIF proportion increases which is above 10% the accuracy of the standardization method decreases, so the accuracy of the Mantel-Haenszel method is higher than the Standardization method.

Detection of DIF by using two methods, namely Mantel-Haenszel and standardization requires carefulness in choosing a method that must be the benchmark of the results of an analysis to obtain accurate DIF information. Therefore, if the proportion of DIF is detected in the test with the standardization method of (≤ 0.10), the results of the standardization method are more a reference than the results of the Mantel-Haenszel method. Conversely, if the proportion of DIF detected in the test with the standardization method is (≥ 0.10), then the results of the Mantel-Haenszel method are more a reference than the results of the standardization method.

To improve the quality of the development of test items as an educational assessment tool, it is necessary to detect DIF before using test items to students. Specifically, the assessment process in the classroom, for teachers who still have limited ability to develop good tests and do not harm the participants, it is suitable to use the Mantel-Haenszel method. However, the standardization method also has advantages and is very suitable for professional test developers, as an initial detection for detecting DIF tests.

ACKNOWLEDGEMENTS

Many thanks are presented to my sponsorship, Indonesia Endowment Fund for Education/LPDP and the BUDI-DN KEMRISTEKDIKTI which funded my study for my Doctoral program.

REFERENCES

AERA, APA, & NCME. (1999). Standards for Educational and Psychological Testing. American Psychological Assosiation: Washington, DC.

Agresri, A., & Finlay, B. (2009). Statistical Methods for the Social Sciences. USA: Pearson.

Allen, M. J., & Yen, W. M. (1979). *Introduction to Measurement Theory*. Monterey: Brooks/Cole Publishing Company.

Anastasi, A. (1976). *Psychological Testing*. New York: Macmillan Publishing Co., Inc.

Azen, R., & Walker, C. M. (2011). *Categorical Data Analysis for the Behavioral and Social Sciences*. New York: Routledge Taylor and Francis Group.

Azwar, S. (2000). *Reliabilitas dan Validitas (Edisi 4)*. Yogyakarta: Pustaka Pelajar.

Berenson, M. L., Levine, D. M., & Krehbiel, T. C. (2012). *Basic Business Statistics: Concepts and Applications*.

(Eric Svendsen, Ed.) (Twelfth Ed). New Jersey: Prentice Hall.

Berk, R. A. (1982). *Handbook of Methods for Detecting Test Bias*. Baltimore, Maryland: The Johns Hopkins University Press.

Budiyono. (2009). The Accuracy of Mantel-Haenszel, Sibstest, and Logistic regression Methods in Differential Item Functioning Detection. *Jurnal Penelitian Dan Evaluasi Pendidikan*, *12*(1), 1–20.

Cizek, G. J., Rosenberg, S. L., & Koons, H. H. (2008). Sources of Validity Evidence for Educational and Psychological Tests. *Educational and Psychological Measurement*, *68*(3), 397–412.

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. https://doi.org/10.1207/s15324818ame0203

Dorans, N. J., & Holland, P. W. (1992). *DIF Detection and Description: Mantel-Haenszel and Standardization*. 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: 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.

Ercikan, K., Roth, W., Simon, M., Sandilands, D., & Lyons-thomas, J. (2014). Inconsistencies in DIF Detection for Sub-Groups in Heterogeneous Language Groups. *Applied Measurement in Education*, *27*(4), 273–285. https://doi.org/10.1080/08957347.2014.944306

Gierl, M. J., Gotzmann, A., & Boughton, K. A. (2004). Performance of SIBTEST When the Percentage of DIF Items is Large. *Applied Measurement in Education*, *17*(3), 241–264. https://doi.org/10.1207/s15324818ame1703

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.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of Item Response Theory*. California: SAGE Publications Inc.

Han, K. T. (2007). WinGen : Windows Software That Generates Item Response Theory Parameters and Item Responses. *Applied Psychological Measurement*, *31*(5), 457–459. https://doi.org/10.1177/0146621607299271

Hidalgo, M. D., Galindo-garre, F., & Gómez-benito, J. (2015). Differential item functioning and cut-off scores : Implications for test score interpretation *. *Anuario de Psicología/The UB Journal of Psychology*, *45*(1), 55-69.

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.

Huggins, A. C. (2012). *The Effect of Differential Item Functioning on Population Invariance of Item Response Theory True Score Equating*. University of Miami.

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.

Kerlinger, F. N. (1986). *Asas-asas Penelitian Behavioral (terjemahan L.R. Simatupang)*. Yogyakarta: Gajahmada University Press.

Lind, D. A., Marchal, W. G., & Wather, S. A. (2012). *Statistical Tecnique in Business & Economics*. New York: McGraw-Hill Companies, Inc.

Loong, T. (2003). Understanding sensitivity and specificity with the right side of the brain. *BMJ*, *327*, 716–719.

Masters, G. N., & Keeves, J. P. (1999). *Advances in Measurement in Educational Research and Assessment*. United Kingdom: Elsevier Science Ltd.

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.

Nunnally, J. C. (1978). *Psychometric theory*. New York: McGraw Hill.

Oliveri, M. E., Ercikan, K., & Zumbo, B. D. (2014). Effects of Population Heterogeneity on Accuracy of DIF Detection. *Applied Measurement in Education*, *27*(4), 286–300. https://doi.org/10.1080/08957347.2014.944305

Ong, Y. M. (2010). *Understanding Differential Functioning By Gender in Mathematics Assessment*. University of Manchester for the degree of Doctor of Philosophy.

Urbina, S. (2004). *Essentials of Psychological Testing*. New Jersey: John Wiley & Sons,Inc.

Whitmore, M. L., & Schumacker, R. E. (1999). A comparison of logistic regression and analysis of variance differential item functioning detection methods. *Educational and Psychological Measurement*, *59*(6), 910–927.

Yerushalmy, J. (1947). Statistical Problems in Assessing Methods of Medical Diagnosis, with Special Reference to X-Ray Techniques. *Public Health Reports*, *62*(40), 1432–1449.

Zhu, W., Zeng, N., & Wang, N. (2010). Sensitivity, Specificity, Accuracy, Associated Confidence Interval and ROC Analysis with Practical SAS ® Implementations. In *Section of Health Care and Life Sciences* (pp. 1–9). Maryland: Northeast SAS User Group proceedings.

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.