# Comparison of the Five-package National Examination in Math in 2011 Based on DIF (A Case Study In Jakarta)

Wardani Rahayu
*State University of Jakarta*
Corresponding author : wardani9164@yahoo.com

**Abstract**

This study was aimed to compare the Five-package National Examination in Math for Social Studies in 2011 in terms of the DIF items and non-DIF items in Jakarta. Data collection technique was simple random sampling. The independent variables were the type of school and the Five-package UN. The schools were divided into two types. They were Senior High School (SMA) and Madrasah Aliyah (MA). The package numbers of UN in Math for Social Studies were 12, 25, 39, 46 and 54. The National Examination data were analyzed by two-factor ANAVA in terms of DIF and non-DIF items. The detection of DIF items with Mantel-Haenszel's method was based on gender. The Five-package National Examination in 2011 consisted of 11 DIF items and 29 non-DIF items. Items used in this study are the items that fit model with the three-parameter logistic model. The results was on DIF items and DIF items for the two types of schools (SMA and MA), there was no difference between the scores of the National Examination in Math for Social Studies on packages 12, 25, 39, 46 and 54.

*Keywords*: National Examination in Math, five packages, DIF, SMA, MA

## INTRODUCTION

The Indonesian government has been implementing national exam in junior high schools and senior high schools since 1969. The national exam has several times changed its name with different rules of engagement. In the years of 1982 - 2002 this exam was held as the National Final Learning Evaluation (EBTANAS), in 2003-2004 as the National Final Exam (UNAS) and in the years of 2005-2013 as the National Examination (UN). National examination aims to measure the students competency in certain subjects performed simultaneously at the end of level of education in junior high schools and senior high schools nationwide in order to assess the achievement of national education standards. The implementation of National Examination is expected to improve the quality of education in Indonesia.

The quality improvement of National Examination is being conducted by the Ministry of National Education and Culture. One of which is the policy of using five-package National Examination test device in 2011. These are packages of exam questions that increasingly reduce the potential for cheating, because it is very difficult for schools to apply cheating.[6] The five-packages of test device using the same exam questions with different item sequence numbers. These test items were developed based on the Graduate Competency Standards (SKL) compiled by the National Education Standards Agency (BSNP).

Each package of test device were numbered, for the exam question packages of Social Science Mathematics of SMA/MTs in Jakarta had number 12, 25, 39, 46 and 54. The distribution of the test device on a class was performed in random by the exam committee at the related school. One class consisted of 20 people and as many as four people working on the same package, so that the people who cheat did not know which packages were being worked on by the students.[7] The policy of five-package National Examination questions and its implementation was an attempt of the Ministry of National Education and Culture to improve the quality of implementation of the National Examination.

Problems in the implementation of the Indonesian National Examination were, first of all, Indonesia is an archipelago with diversity of ethnic, religion, culture which is a challenge in the implementation of national examination. Between one region and another have different conditions. Schools in the city had met the eight education standards which were better than schools in other regions. The schools in Jakarta motivated students with the motto: 'let's achievers!', while there were some schools in other regions, particularly in inland motivated their students with the motto: 'let's go to school!' This diversity was what needs to be considered when preparing the test device which had a good quality for the National Examination. Items in National Examinations test device should be able to measure what will be measured. Therefore, the items validation process must be carried out before the implementation National Examination. In line with the cooperation between APA, AERA, & NCME in 1999 in Standards, ETS (2002) stated that the validation is the most important aspect in determining the quality of the measurements in the fields of psychology and education.[10] The test device items which provide different measurement results for test participants who have the same abilities are called bias. Item bias is always linked to the test participants in a group. The term of item bias is also called the differential item functional (DIF). Naga also

stated that the term DIF reflects the purpose of the method of item bias detection in identifying item bias which have different functions for different groups of test participants.[9] Holland & Thayer called the item bias as differential item functioning.[2] Hidalgo and Lopez added after matching on an ability of certain test participants.[8] One method for detecting DIF in an item by using the non-parametric statistical approach is the Mantel-Haenszel method. The Mantel-Haenszel procedure compares the odds for success between groups after conditioning on ability.[1]

The second problem in the implementation of the National Examination in 2011 was whether the five-packages distributed in one class would give the same National Examination participants scores on the test participants who had the same capabilities. The items on the five-package were the same, but the item numbers were different. Randomization of item numbers on each package would result in a change of item order based on the item difficulty levels. For example the 1st item on one test device would get an easy item, while the other packages would get difficult item. This would affect the psychology of test participants when solving the National Examination questions and resulted in the National Examination score. This problem became the anxiety of schools and parents.

The study related to National Examination and DIF had been carried out, for example Badrun (2008), Effendi (2011) and Sudaryono (2012) compared several methods of DIF detection, Rahayu (2010) examined the accuracy of linking methods on DIF detection based on the number of false positive items.[11] The difference with this study was not to see the accuracy of the linking method or the DIF detection method, but the DIF method was only to see the DIF and non-DIF items, whether there was a difference in the National Examination scores by type of school and type of package, especially in Jakarta.

## METHOD

The method used was the method of ex post facto. The dependent variable was the National Examination scores in 2011, while the independent variable was the number of National Exam package, DIF and non-DIF items, and type of school. National exam package numbers were divided into five, namely 12, 25, 39, 46 and 54. Schools were divided into two types, namely Senior High School (SMA) and Madrasah Aliyah (MA).

The study design used was a 5x2x2 factorial design. The data used was the score of students' work from student respondents to the National Examination in mathematics IPS in Jakarta in 2011. Sampling of this study was determined by random sampling technique. Data analysis techniques using two-way ANOVA.

**Table 1** Number of Test Participant Responses as Samples

| Type of Schools | National Examination Package Number (A) | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | 12 | | 25 | | 39 | | 46 | | 54 | |
| | DIF | Non DIF | DIF | Non DIF | DIF | Non DIF | DIF | Non DIF | DIF | Non DIF |
| SMA | 300 | | 300 | | 300 | | 300 | | 300 | |
| MA | 300 | | 300 | | 300 | | 125 | | 125 | |

This study was limited to the National Examination scores derived from the items which were model-fit to the three-parameter logistic model and DIF detection with the Mantel-Haenszel method. The study procedures performed were (1) a data retrieval in the form of scores of the National Examination in Social Science Mathematics for students in Jakarta in the Center of Assessment and Testing of the Ministry of National Education. Data was in the form of zero-and-one-shaped score with a length of test device as many as 40 items. Test device consisting of 5 types of packages containing the same exam questions; the difference was that the item sequence numbers were randomized between one package and another. (2) The test device consisted of 40 items; those items were model-fit to the three-parameter logistic model (L3P). (3) The DIF of items which were model-fit would be detected by gender with the Mantel-Haenszel method. It aimed to find whether the items containing DIF or not. The number of responses of test participants of the package number 12 for reference (male) and focal (female) groups used for DIF detection was 3100 respectively. (3) The number of responses of test participant of each package would be drawn in random for SMA and MA groups. (4) The responses of test participants would then be summed in the form the National Examination scores compared from different groups of students transformed with T-Score formula: $T = 10 z + 50$.

**RESULTS AND DISCUSSION**

The National Examination test devices items for Social Science Mathematics in 2011 pursuant to items number in the package 12 which were model-fit to the three-parameter logistic model were 12 items not model-fit and 28 items model-fit. The items which were model-fit were items number 1, 8, 10, 12, 13, 14,15, 16, 17, 18, 19, 20, 21, 22, 23, 25, 27, 29, 31, 32, 33, 34, 35, 37, 38, 39, 40. DIF detection results by using the Mantel Haenszel method.

Descriptively, the average score of the National Examination in Social Science Mathematics in 2011 at SMA and MA groups for the five-packages was almost the same. It was shown from the average score between 49.95 till 50.02. Similarly, the National Examination score distribution in 2011 had a nearly the same homogeneity in the group of SMA and MA for the five-packages.

Results of testing with a two-way ANOVA was that there was no difference in the National Examination scores on SMA and MA groups, no difference in the National Examination scores on the five-packages, no difference in the National Examination scores either on DIF items, non-DIF items or the whole items. Similarly, there was no interaction between the type of the National Examination package and the type of school to the National Examination scores, no interaction between the type of National Examination package with DIF items, non-DIF items and the whole items to the National Examination scores, and no interaction between the type of the National Examination package, type of school, as well as DIF items, non-DIF items and the whole items to the National Examination scores.

**Table 2** Mean and Standard Deviation Based on Type of UN Package and Type of School

| Item | Scholl | Package 12 | | Package 25 | | Package 39 | | Package 46 | | Package 54 | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Mean | St. Dev | Mean | St. Dev | Mean | St. Dev | Mean | St. Dev | Mean | St. Dev |
| Non DIF | SMA | 50.02 | 9.99 | 50.04 | 10.00 | 49.96 | 10.02 | 50.02 | 10.02 | 49.99 | 9.99 |
| | MA | 50.01 | 9.95 | 49.96 | 9.99 | 49.99 | 10.01 | 49.87 | 9.96 | 50.02 | 10.01 |
| DIF | SMA | 50.00 | 10.02 | 49.99 | 9.99 | 50.01 | 10.01 | 50,00 | 10,00 | 50.00 | 10.00 |
| | MA | 50.00 | 10.02 | 50.00 | 10.00 | 50.01 | 10.00 | 50.00 | 10.00 | 50.00 | 10.00 |
| Total | SMA | 49.96 | 10.02 | 49.95 | 10.01 | 50.00 | 10.00 | 49.99 | 9.98 | 49.99 | 9.99 |
| | MA | 50.00 | 10.02 | 50.00 | 9.99 | 50.01 | 10.00 | 50.00 | 10.00 | 50.01 | 9.98 |

**Table 3** Analysis Results of the Two-Way ANOVA

| Source | Df | Mean Square | F | Sig. |
|---|---|---|---|---|
| Corrected Model | 29 | .175 | .002 | 1.000 |
| Intercept | 1 | 1.984E7 | 1.984E5 | .000 |
| Package | 4 | .024 | .000 | 1.000 |
| School | 1 | .015 | .000 | .990 |
| Bias | 2 | .126 | .001 | .999 |
| Package * School | 4 | .153 | .002 | 1.000 |
| Package * Item Bias | 8 | .116 | .001 | 1.000 |
| School * Item Bias | 2 | .573 | .006 | .994 |
| Package * School *Item Bias | 8 | .253 | .003 | 1.000 |

Inferential test results supported by the average value of the National Examination scores in SMA and MA groups for the five-packages was almost the same, which was close to the score of 50 and the score distribution which the homogeneity was almost the same. The not-different National Examination scores on SMA and MA in 2011 by type of school and type of package was due to the tests used had an item difficult level value between -1.194 and 1.199. According to Hambleton and Swaminathan (1990: 36), the item difficult level value of $b_j$ approaching –2, it can be said that the test item is easy while if the value of $b_j$ approaching 2, it can be said that the test item is difficult. The difficult level value of 26 items of test device which was model-fit approaching -2 and 2

items approaching 2. Therefore, 28 items of the National Examination test device of SMA/MA in 2011 which were model-fit to the three-parameter logistic model (L3P) consisted of 26 easy test items and 2 difficult test items. The 35th and 40th items were the difficult test item and non-DIF item. The DIF item which had the most easy difficulty level was the 1st item and the non-DIF item was the 20th item.

The 35th item measures the student's ability to determine the expected frequency appearing at least two images of the three coins thrown together as much as 600 times. The 40th item measures the student's ability to determine the number of ways to take 20 rosebuds randomly taken as many as 15 rosebuds. These items with difficult category were probability. Probability in the majoring in Social Studies of SMA/MTs was the course matter which was considered difficult. Therefore, additional time was required to learning outside the classroom; it can be in the form of group work with assignment assessment or project assessment, thus enabling students to have competence relating to probability.

The test device information function of National Examination in SMA / MA at 7.3092 on the test participants' ability $\theta$ = -1. The item information function was between 0.0142 and 0.6012. The 20th item had the biggest item information function which was 0.6012. The 4th item contributed most to the test information function compared to other items. Item information function described the power of an item on the test device. The information function was critical to be used in the selection of items; the higher the information provided by the item, the better the item. Item analysis results empirically using IRT showed that 28 items which were model-fit to the three-parameter logistic model (L3P) in the National Final Examination test of SMA/MA in 2011 in mathematics were categorized as good items. This indicated that the item validation process of the National Examination test device in 2011 had been well-implemented so that it produced items that had high information function. However, in the implementation of item validation of the National Examination test device, whether or not the item is bias is need to be seen with respect to gender, region, ethnicity in all regions of Indonesia.

This study needs to be followed up with further study by comparing the National Examination scores in other regions such as Central Part of Indonesia and Eastern Part of Indonesia, especially for rural areas in order to obtain a picture of weakness of students competence in mathematics. Other regions in Indonesia have the characteristics that are different from Jakarta. Where are the regions that have the same score with Jakarta. Future studies can be conducted to detect items with non-parametric approach without being restricted by items which are model-fit in IRT models.

## CONCLUSION

National Examination Results for Social Science Mathematics in 2011 in SMA and MA groups for the five-package, and for the group of DIF and non-DIF items did not differ in Jakarta area. Most of the National Examination items which were model-fit to the three-parameter logistic model had the easy difficulty level so that randomization of item numbers on the five-package did not affect the National Examination Results. The National Examination validity should be kept to a minimum bias because Indonesia is an archipelago country with a diversity of ethnic, religion, culture.

## REFERENCES

Clauser, Brian, Kathy Mazor, Ronald K Hambleton. (2013) The Effects of Purification of Matching Criterion on The dentification of DIF Using The Mantel-Haenszel Procedure. *Applied Measurement in Education*, Volume 6, Issue 4, 1993, 269-279. http://www.tandfonline.com/ (acces 17 April 2013)

Camili, Gregory., A Shepard Lorrie. (1994). *Methods for Identifying Biased Test Items*. London: Sage Publications.

Effendi. (2012) "Detection of Crossing Dif: A Comparison of Raju's Area Measure, Lord's Chi-Square, And Likelihood Ratio Test." *Jurnal Evaluasi Pendidikan*. Vol. 2 No 2, 2012

Hambleton, Ronald K dan H Swaminathan. (1990). *Item Response Theory: Principles and Aplications*. Boston : Kluwer.Nijhoff Publishing.

Kartowagiran, Badrun. (2013) Perbandingan Berbagai Metode Untuk Mendeteksi Bias Butir. http://etd.ugm.ac.id/ (acces 17 April 2013).

Lima Paket Soal UN Akan Tekan Kecurangan. http://www.republika.co.id/berita/pendidikan/berita/ (acces 17 April 2013).

Lima Paket Soal UN untuk Cegah Kecurangan. http://suaramerdeka.com/v1/index.php/read/news/ (acces 17 April 2013).

Montesinos, Maria Dolores Hidalgo, Jose Antonio Lopez-Pina. (2002) "Two-Stage Equating In Differential Item Functioning Detection Under The Graded Response Model With The Raju Area Measures And The Lord Statistic." *Educational and Psychological Measurement*, Vol. 62 No. 1, February 2002, 32-44

Naga, Dali S. (1992). *Pengantar Teori Sekor Pada Pengukuran* . Jakarta: Besbtas.

Rahayu, Wardani and Ali Ridho. (2008). Analisis Data *Indonesian National Assessment Program* (INAP) Tahun 2007. Jakarta: Pusat Penilaian Pendidikan Badan Penelitian dan Pengembangan DIKNAS.

Rahayu, Wardani. (2010) "Linking Method and Flase Positive Item on DIF Detection Based on Item Respons Theory." *Jurnal Penelitian dan Evaluasi Pendidikan*, 14 (1), 2010.

Sudaryono. (2013) Perbandingan Sensitivitas Metode *Chi-Square* Scheuneman, Mantel-Haenszel, Dan Teori Responsi Butir Model Rasch Pada Pendeteksian *Differential Item Functioning* (*Dif*). *Jurnal Evaluasi Pendidikan*. Vol. 3 No. 1, 2013.