The Equating of Battery Test Packages of Mathematics National Examination 2013-2016

Badrun Kartowagiran[1] , *Sudji* Munadi[1], *Heri* Retnawati[2,*], and *Ezi* Apino[2]

[1]Educational Research and Evaluation, Graduate School, Universitas Negeri Yogyakarta, Indonesia
[2]Mathematics and Natural Science Faculty, Universitas Negeri Yogyakarta, Indonesia

Abstract. When a test implements several test instruments, there is an assurance that the test instruments that will be implemented are equal and this equality is an urgent matter. Therefore, this study aimed at confirming the equality of the test instruments that had been implemented in Mathematics National Examinations (MNE) 2013-2016. This study was conducted using quantitative approach. The data were gathered using students' response documentation technique and MNE test instruments; the test instruments were drawn from the packages that had been administered to the national examinations for junior high school students in the Province of Yogyakarta Special Region. The data were analyzed by using the stages of item parameter estimation, designing the equating equation, equating through concordance model, drawing the test characteristics curve, and interpretation. The results of the analysis showed that the instruments that have been administered are almost equal, both from one package to another and from one year to another, with the standards of test instrument 2013.

1 Introduction

In order to improve the National Examination (NE) quality and to decrease the fraud, the Republic of Indonesia Government implements a policy that the administration of NE should make use of several test packages. This policy has been implemented since 2010; in that year, the national examination only made use of 2 test packages. One year later, the number increased into 5 test packages. In 2013 and 2014, the NE was administered using 20 test packages. Then, in 2015 and 2016 the NE was administered using 5 packages again. The test packages that had been administered were developed based on the same guidelines so that these packages were expected to measure the same skills. Thereby, the test packages that had been administered in these NE are assumed as the parallel ones.

The use of several parallel test packages in the NE can decrease the potentials of fraud during the test administration [1]. In the same time, the use of those packages has another advantage namely that the secrecy of the test items are more assured [2]. With the abundant number of test packages that will be administered in the NE, the potentials of leakage can be minimized [3, 4]. However, the main challenge in administering several test packages during the NE is related to the assurance that these packages are equal and measure the

* Corresponding author: heri_retnawati@uny.ac.id

same indicator. In other words, there should be a guarantee that the test packages that will be implemented in the NE are the equal ones and this is an urgent matter.

The guarantee that the test packages are equal can be confirmed both theoretically and empirically. This confirmation is related to the concept equating or concordance of test score [5]. The equating can be performed using the classical approach (classical test theory approach) and the modern approach (item response theory approach) [6, 7]. The equating using the modern approach basically calculates the students' ablities and level of difficulty into certain scores with a linear equation [2, 8]. In order to perform this calculation, parameter of index discriminant (*a*), level of difficulty (*b*), and pseudo-guessing (*c*) should be estimated first. The estimation that involves these parameters in the item response theory is known as the 3PL Model. On the other hand, several methods that can be implemented in order to perform equating are namely mean and mean method, mean and sigma method, and item characteristics curve method which includes the Stocking and Lord method [9, 10].

In the mean and mean method, the equating constant α and β only involves the mean of the item difficulty index (*b*) and the mean of item discriminant index (*a*) after the parameters have been estimated first [2]. For example, if the score of test 1 is equated to that of test 2 using the 3PL Model using the mean and mean method according to [9], then the linear equation model will be as follows:

$$b_2 = \alpha b_1 + \beta \tag{1}$$

So that the researchers attain the following equations:

$$\overline{b_2} = \alpha \overline{b_1} + \beta \Leftrightarrow \beta = \overline{b_2} - \alpha \overline{b_1} \tag{2}$$

and

$$a_2 = \frac{a_1}{\alpha} \tag{3}$$

then

$$\overline{a_2} = \frac{\overline{a_1}}{\alpha} \Leftrightarrow \alpha = \frac{\overline{a_1}}{\overline{a_2}} \tag{4}$$

$\overline{b_1}$ and $\overline{b_2}$ are mean of common item difficulty index from test 1 and test 2. $\overline{a_1}$ and $\overline{a_2}$ are mean of common item discriminant index from test 1 and test 2. While, α, β is equating constants.

Then, in the mean and sigma method the determination of equating constant between α and β involves the mean and the standard deviation of the item parameters. For example, if the score of test 1 is equated to that of test 2 using the 3PL Model, according to [9], the equating model will be as follows :

$$b_2 = \alpha b_1 + \beta \tag{5}$$

As a result, the researchers attain the following equation :

$$\overline{b_2} = \alpha \overline{b_1} + \beta \Leftrightarrow \beta = \overline{b_2} - \alpha \overline{b_1} \tag{6}$$

and

$$S_2 = \alpha S_1 \Leftrightarrow \alpha = \frac{S_2}{S_1} \tag{7}$$

with S_1 and S_2 are standard deviation of item difficulty index from test 1 and test 2

Furthermore, in the item response theory, if the item response model is compatible to a data set then any linear transformation from the measurement scale will also be compatible for the data. This means that there has been relationship between the measurement scales from both tests. Thereby, if the scale of test 1 is equated to that of test 2 for the 3PL Model

then the relationship between the item parameter and the participants' abilities for both scales can be stated as follows [10].

$$q_{1i}^{*} = aq_{1i} + b \tag{8}$$

$$a_{1j}^{*} = \frac{a_{1j}}{\alpha} \tag{9}$$

$$b_{1j}^{*} = \alpha b_{1j} + \beta \tag{10}$$

$$c_{1j}^{*} = c_{1j} \tag{11}$$

a_{1j}, b_{1j}, c_{1j} are item parameter for item j on the test 1 scale. a_{1j}^{*}, b_{1j}^{*}, c_{1j}^{*} are item parameter for item j on the test 1 scale after having been equated to test 2. q_{1i} is the ability of i participant on the test 1 scale and q_{1i}^{*} is the abilty of i participant on the test 1 scale after having been equated to test 2. The c parameter is not transformed because its value does not depend on the metric q or, in other words, the parameter c is free from the scale transformation [10]. Moreover, the constant that links the score of test 1 and that of test 2 can be calculated using multiple methods known as test score linking method.

In relation to the use of several equating methods, the results of a study by [11] showed that the Stocking and Lord-type characteristics curve equating and the mean and sigma method have equally been good. On the other hand, [12] through the results of their study stated that there is not any best single method for equating the test scores. As a result, the researchers can select one of the equating methods in order to confirm the quality among the test packages according to the needs. Based on the above problems and the theoretical review, this study then aimed at confirming the equality of the test packages that had been administered in the MNE 2013-2016

2 Method

This study was a descriptive explorative which aimed at describing the equality of the test packages that had been administered in the MNE from 2013-2016. The data were gathered using the students' response documentation technique and the MNE test instruments from the test packages that had been administered in the NE for junior high school students in the Province of Yogyakarta Special Region. There were 9 test instruments that had been selected as the object of this study from the MNE 2013, 8 instruments from MNE 2014, and 5 instruments from MNE 2015 and 2016. Then, the test package 1 from MNE 2013 had been defined as the standard of reference for equating the inter-year test packages. The data analysis was performed through the stages of item parameter estimation, design of equating equation, equating using the concordance model, drawing of test curve characteristics (TCC), and interpretation.

3 Results and Discussions

After the item parameter estimation that included the index of difficulty (*b*), the index discriminant (*a*), and the pseudo-guessing (*c*) had been performed, the researcher equated the model without common items (concordance) using the mean and sigma method. The results of concordance analysis yielded the transformation equations for each item parameter. These parameters were the equal ones and then these parameters were manipulated in order to illustrate the characteristics curve from each test instruments. From the test characteristics curve (TCC), the researchers attained the information regarding the

test instruments equality both the inter-package one and the inter-year one. Then, the TCC from each instrument was presented in Figure 1, Figure 2, Figure 3, and Figure 4.

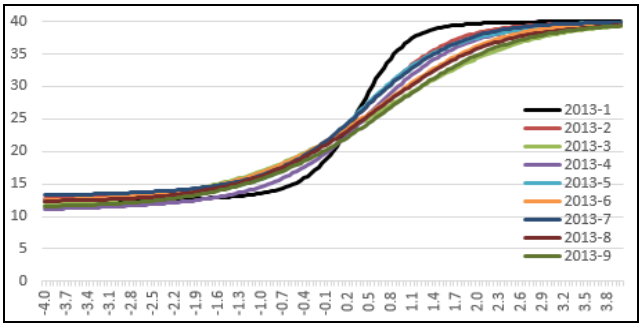

Fig. 1. The TCC of the MNE 2013 Test Instruments

By looking at Figure 1 the researchers found information that the equating that had been conducted resulted in mutually approaching curves. In a closer look, there was one curve that had been inclined to less coinciding namely test package 1 (2013-1). On the contrary, the curves of the other test packages (2013-2 to 2013-9) were inclined to more coinciding from one to another. This finding implied that the test instruments that had been administered in MNE 2013 had been equal from one to another.

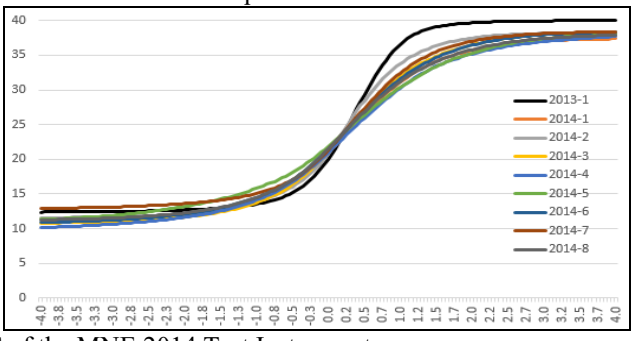

Fig. 2. The TCC of the MNE 2014 Test Instruments

Looking at Figure 2, it was apparent that the curves that had been resulted were also inclined to coincide one to another. In a closer look, it was also apparent that the inter-test package curves (2014-1 to 2014-8) were also coinciding from one to another. These findings implied that the test packages that had been administered in the MNE 2014 for Junior High School had been equal. Then, in comparison to the characteristics curve of test package 1 from the MNE 2013 (2013-1) it was apparent that these curves had been less coinciding although they had been close to each other. Such curves indicated that the test instruments that had been administered in the MNE 2014 had almost been equal to those that had been administered in the MNE 2013.

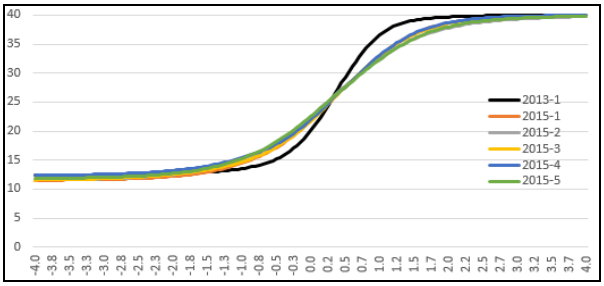

Fig. 3. The TCC of the MNE 2015 Test Instruments

Looking at Figure 3 it was apparent that the curves that had been yielded were also mutually closer to each other. In a closer look, it was also apparent that the inter-test package curve in MNE 2015 (2015-1 to 2015-5) had been very coinciding. Such coincidence implied that the test packages that had been administered in the MNE 2015 had been equal. In comparison to the test package 1 from the MNE 2013, it was apparent that the curve of the MNE 2015 had been inclined to coincide only on certain ability and to move away on the other ability. This condition also indicated that the test instruments that had been administered in the MNE 2015 had been equal to those that had been administered in the MNE 2013.

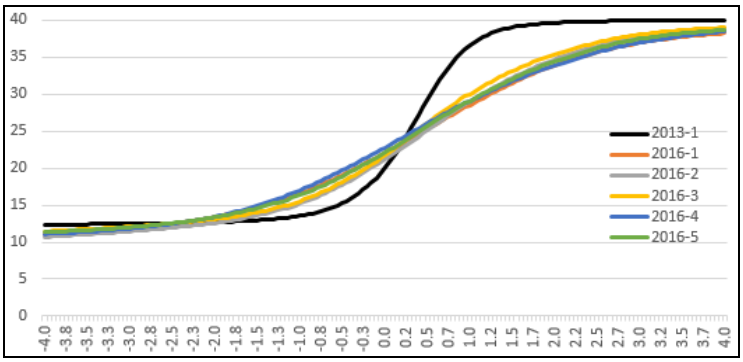

Fig 4. The TCC of the MNE 2016 Test Instruments

By looking at the TCC in Figure 4 the researcher found that the yielded curves were closer to each other as well. In a closer look, it was apparent that the inter-test package curve from the MNE 2016 test packages (2016-1 to 2016-5) had been mutually coinciding. This coincidence implied that the test packages that had been administered in the MNE 2016 had been equal. Next, in comparison to the item characteristics curve package 1 from the MNE 2013 (2013-1), the researchers could see that the curves had been less coinciding on certain ability scales; however, these curves still have similar tendency. Thereby, based on the information of the TCC there was an indication that the test instruments that had been administered in the MNE 2016 had been equal to those that had been administered in the MNE 2013.

The results of the analysis showed that the test instruments that had been administered were almost equal, both in terms of package and in terms of year, according to the standards of test instruments 2013. One of these results were in concordance to a study by [2] which showed that one of the test instruments that had been administered in the MNE, namely the 2014 one, which consisted of 20 test packages, had a tendency to be equal. Moreover, the presence of these results clearly indicated that the test instruments that had been administered, especially the test packages, in the NE had been considered fair. As a result, the test participants did not consider themselves being in disadvantages. This matter certainly could be confirmed empirically using the equating procedures that had been performed. [13] proposed the importance of test equating procedures in order that the NE test participants would not be in disadvantages and the scores that would be awarded to the NE test participants became fair although each participant completed different test package and experienced different difficulty level.

Looking at inter-year equating results, it appears that TCC of test instruments of the MNE 2015 and 2016 are less closely aligned with test instruments of MNE 2013 (see figure 3 and 4). This is because the NE 2015 and 2016 began to accommodate Higher Order Thinking Skills (HOTS) items which refers to PISA and TIMSS items. This is inseparable from the Indonesian government's policy to implement the Curriculum 2013, where one of the main focus in this curriculum is the development of students' HOTS [14, 15]. In

addition, in the NE 2015 and 2016 has been implemented Computer Based National Examination (CBNE), where the CBNE score is more accurate than Paper and Pencil Test score [16].

4 Conclusions

Based on the results of the analysis, the researchers conclude that the instruments that have been administered in the MNE (2013-2016) are almost equal, both from one package to another and from one year to another, with the standards of test instrument 2013. On the other hand, the researcher propose the following suggestions namely (1) for the future studies, it is necessary to carry out the equating procedure of test instruments of the NE that employ wider-scale object, both in terms of coverage (province) and of sample, so that the information that can be attained from the equating procedures will be more accurate ; (2) the equating procedures are very important to perform in each test instruments especially for the implementation of Final Examination and National Examination so that information regarding the quality of the test instruments can be attained. This information can serve as the foundation for the government to make policies regarding the design of Final Examination and National Examination test items.

References

1 H. Retnawati, B. Kartowagiran, J. Arlinwibowo, E. Sulistyaningsih, Int. J. Instr. **10(3)**, 257–276, (2017)

2 H. Retnawati, J. Kependidikan, **46(2)**, 164–178, (2016)

3 T. Rijanto, J. Penelit. dan Eval. Pendidik. **16(1)**, 365–383, (2012)

4 A. P. Herkusumo, J. Pendidik. dan Kebud. **17(1)**, 455–471, (2011)

5 N. J. Dorans, J. Appl. Psichol. Meas, **28(4)**, 219–226, (2004)

6 J. Ryan, F. Brockmann, *A practitioner's introduction to equating with primers on classical theory and item respons theory*. (Council of Chief State School Officers, 2009)

7 İ. Uysal, S. Kilmen, Int. Online J. Educ. Sci. **8(2)**, pp. 1–11, (2016)

8 H. Retnawati, *Teori respons butir dan penerapannya* (Parama, Yogyakarta, 2014)

9 R. K. Hambleton, H. Swaminathan, H. J. Rogers, *Fundamental of item response theory* (Sage Publication, Newbury Park, CA, 1991)

10 M. J. Kolen, R. L. Brennan, *Test equating: Methods and practices* (Springer, New York, NY, 2004)

11 X. Pang, E. Madera, N. Radwan, and S. Zhang, *A comparison of four test equating* (Methods Research Report, 2010)

12 C. H. Yu, S. E. O. Popp, Pract. Assessment, Res. Eval. **10(4)**, 1–19, (2005)

13 D. S. Sukirno, Cakrawala Pendidik. **26(3)**, 305–321, (2007)

14 H. Retnawati, S. Hadi, A. C. Nugraha, Int. J. Instr. **9(1)**, 33–48, (2016)

15 E. Apino, H. Retnawati, J. Phys. Conf. Ser. **812**, pp. 1–7, (2017)

16 H. Retnawati, Turkish Online J. Educ. Technol. **14(4)**, 135–142, (2015)