REiD (Research and Evaluation in Education), 4(2), 2018, 164-174
Available online at: http://journal.uny.ac.id/index.php/reid

Comparing the methods of vertical equating for the math learning achievement tests for junior high school students

***[1]Chairun Nisa; [2]Heri Retnawati**
[1]Department of Educational Research and Evaluation, Universitas Negeri Yogyakarta
Jl. Colombo No. 1, Depok, Sleman, Yogyakarta 55281, Indonesia
[2]Department of Mathematics Education, Universitas Negeri Yogyakarta
Jl. Colombo No. 1, Depok, Sleman, Yogyakarta 55281, Indonesia
*Corresponding Author. E-mail: c.nisa.258@gmail.com

Submitted: 12 April 2018 | Revised: 20 December 2018 | Accepted: 21 December 2018

Abstract

Developing the students' mathematical ability needs to be carried out to improve the teaching process. This is very important for continuous education. This study aimed to describe: (1) the characteristics of the mathematics achievement tests for grades VII and VIII; (2) the equity constant of the vertical equating result of the mathematics achievement; (3) the accuracy of the mean & mean method, mean and sigma, Haebara characteristics curve, Stocking & Lord characteristics curve methods in the vertical equating of the tests for grades VII and VIII. The data were the students' scores for the Higher Order Thinking tests collected with the anchor test design. The analysis technique utilized was the descriptive quantitative analysis. The findings of the study show that: (1) the learning achievement tests for grades VII and VIII have the difficulty level (location) in the fair category (0.190 and 0.451), and the discrimination index (slope) in the category of good with the mean of 0.700 and 0.633; (2) the vertical equating result shows an equation of Y' = 0.88X-0.27 with the mean and mean method, Y' = 0.19X-0.02 with the mean and sigma method, Y' = 0.38X-0.12 with the Haebara characteristics curve method, and Y' = 0.57X-0.18 with the Stocking and Lord characteristics curve; (3) the lowest Root Mean Square Different (RMSD) belongs to the mean and mean method, followed by the Stocking and Lord characteristics curve method, mean and sigma method, and the Haebara characteristics curve method.

Keywords: *equating method, vertical equating, HOT, mathematics*

Introduction

Science and technology development in Indonesia has brought about changes in almost every aspect of human life. The development demands that different problems be solved through the effort to master science and technology. In order to be able to contribute to the global competition in the 21st century, human being needs to develop the self-quality so that they can compete with others. The human resource quality is influenced by education.

The improvement of the quality of education becomes an essential pillar for the development of education in Indonesia. Quality education will result in competitive human resources as stated in the Law of Republic of Indonesia No. 20 of 2003 on National Education System. It is stated that the national education functions to develop the capability, character, and civilization of the nation for enhancing its intellectual capacity, and is aimed at developing learners' potentials so that they become persons imbued with human values who are faithful and pious to the one and only God; who possess morals and noble character; who are healthy, knowledgeable, competent, creative, independent; and as citizens, are democratic and responsible.

The Partnership for 21st Century Skills (P21) argues that teaching should focus on developing critical thinking, communication, collaboration, and creativity as students' skills

Copyright © 2018, REiD (Research and Evaluation in Education)
ISSN 2460-6995

in the 21st century. The 4C's skills are a part of the higher order thinking skills. Therefore, students need to develop their higher order thinking skills in every educational process.

Based on the data from different surveys, it is found out that students' achievement in maths in PISA and TIMSS is still low. Thus, holistic and continuous efforts need to be made to improve the quality of education from all parties including students, teachers, principals, and the government. According to Mardapi (2012, p. 12), efforts to improve the quality of education in educational institutions can be made by improving the quality of the teaching and assessment system. This means that teaching is closely related to assessments. Teachers, as an important component in education, should be able to carry out their duties and play their roles as stated in Law No. 14 of 2005 of Republic of Indonesia about Teachers and Lecturers. Teachers are expected to be able to develop students' potentials, through both the teaching process in the class and the assessment model used. The assessment model used by teachers can actually provide information on the teaching process and learning achievement.

A good assessment can be carried out by collecting accurate data related to students' learning achievement and this can make the class assessment process beneficial to the students, that is, it can improve the students' motivation and learning achievement (Stiggins & Chappuis, 2012, p. 3). Therefore, learning achievement assessment is expected to be able to provide information about the students' ability development. The information can be a reference to know the quality of the learning achievement at the class, school, or national levels. This can be used as a study to improve the quality of Indonesian education.

The test in mathematics has different characteristics from that in other subjects. The mathematics materials are hierarchical and closely related to each other. This means that the students' mastery of previous materials becomes the basis for continuing to and understanding of the materials in the next level. Teachers are expected to be able to write a good test and also to use the test to connect the students' learning achievement in different grades so that the information about the students' ability development can be known.

In addition to knowing the characteristics of the test items used, teachers are expected to make sure that, in order that the information about the students' ability development is accurate, the test items should be in the students' ability level (Gagné, 1977, p. 158). The use of test items which are beyond the students' ability will make the students unable to answer the questions so that teachers will not be able to find out the information about the students' development. Students of the same age and grade may not have the same development.

The scores of two different tests from two or more different groups can be compared when the items are equal and are based on the same scale (Kolen & Brennan, 2004, p. 5). The equating between scores can be done statistically. A statistical analysis is carried out to the scores of two different tests to be adjusted on the same scale. The statistical process used to produce a single scale from the scores of two different tests with the same scale is called equating (Kolen & Brennan, 1995, p. 5). Hambleton, Swaminathan, and Rogers (1991, p. 123) state that equating is a process to transform the score X to the test score matrix Y or vice versa, so that the result of the equating process can be compared.

There are two kinds of equating process which can be conducted to test scores: horizontal equating and vertical equating. Horizontal equating is the equating carried out to test scores which have equal difficulty index at the same grade, while vertical equating is the equating process carried out to reveal the students' ability measured by test instruments which have different difficulty index and on different grades, but they measure the same trait (Crocker & Algina, 2008, p. 456; Hambleton & Swaminathan, 1985, p. 197). Thus, the vertical equating can be used by teachers to reveal the students' ability development although the students are in different grades and they have different abilities provided that the tests measure the same traits.

The equating using the Item Response Theory approach can be carried out using different methods. They are the mean-and-mean method, the mean-and-sigma method, and the characteristic curve transformation (Kolen & Brennan, 2004). Several previous studies carried out used the classical approach and Item Response Theory on elementary school students by Antara and Bastari (2015); equating using the IRT approach with the mean-and-mean method, mean-and-sigma, Haebara, and Stocking-and-Lord methods for the mixed model by Kartono (2008), and equating using the IRT approach on mixed tests by Uysal and Kilmen (2016). Previous studies showed the accuracy of different methods. The utilization of different equating methods resulted in different equating results, so to find out an accurate result, it is necessary to choose the appropriate design and method in accordance with the condition. Therefore, teachers will be able to find accurate information about the students' ability development.

The scoring model used was Generalized Partial Credit Model (GPCM). This is because GPCM is an alternative scoring in the teaching assessment (Istiyono, 2016).

Method

This is a study of vertical equating in general using the quantitative approach. In the instrument development part, the researchers developed a mathematics HOTS instrument using the mixed model for grades VII and VIII of junior secondary schools administered in the even semester. Revision based on the expert's suggestions was carried out after the readability testing and content validation by an expert. The revised instrument was then tried out in one junior secondary school which was not the sample of the study, that is, SMPN 3 Lubuk Pakam. The data from the try out were analyzed using the IRT approach using the Parscale program to find out the characteristics of the developed items so that the items would be good items. The good items were then set into a mathematics test for grades VII and VIII.

The research was carried out in Deli Serdang District, Indonesia, especially in public junior secondary schools in the district in the academic year of 2016/2017. The study was conducted from May to June 2017. The population was students' response on mathematics test. There were 51 schools taken as the sample using the stratified proportional random sampling technique. This was done because the population had levels, that is, grade VII and grade VIII. The students in each grade were then selected proportionally. The schools were categorized into high, middle, and low categories based on the national examination scores in the previous year (data obtained from *Dinas Pendidikan Pemuda dan Olahraga*). Five schools were selected to be the sample. The sample consisted of 1009 students, including 505 grade VII students and 504 grade VIII students.

The HOTS test instrument on mathematics used was the GPCM analysis model. The test consisted of 15 items each set consisting of 10 items in the form of multiple choice and five items in the form of essay items. The multiple choice items were used as this kind of items is more objective and reliable in finding out the students' response, nor influenced by the subjectivity of the scorers. Meanwhile, the essay items were used to find out the students' higher order thinking skills. The equating design used was the common item non-equivalent groups (Hambleton & Swaminathan, 1985). Both tests had the same items as the anchor. Four multiple choice i-tems (26.7%) were used as the anchor. It was based on the theory which states that the minimum items for the anchor is 20% (Kolen & Brennan, 1995, p. 248).

Unidimension testing was conducted by the factor analysis in SPSS 22. Data can be analyzed using the factor analysis when they meet two criteria: Kaiser-Meyer-Olkin Measure Sample of Adequacy (KMO-MSA) and Bartlett's Sphericity Test. KMO-MSA test was needed to see the sample adequacy and Bartlett's test was used to see the normality of the analyzed data. Field (2000, pp. 453–469) states that further analysis can be carried out when the KMO has the sig. < 0.05 and the MSA is > 0.05. Hambleton and Swaminathan (1985, p. 16) state that the unidimension testing was met when the test only measures one dominant dimension, that is, the same ability. The unidimension aspect can be seen from the eigenvalue obtained from each test and the unidimension criterion can be seen from the scree plot formed. The local independence assumption testing functions to find out that the students' ability is independent from the items. This means that the students' answer to one item is not influenced by the answer to another item. The conformity of the model was done to know the appropriate model with the analyzed data. The conformity testing was meant to know that the items used were appropriate with the model used. The way used to know the conformity of the model was by comparing the chi-square observed and the chi-square table with a certain degree of freedom. Then, the parameter estimation and ability estimation were carried out with the appropriate model. The parameter estimation and the ability estimation were analyzed using the Parscale program.

Vertical equating was carried out based on the result of the item characteristics analysis using IRTEQ program (Han, 2009). The test equating was done by making an equation using the mean-and-mean, mean-and-sigma, Haebara characteristics curve, and Stocking-and-Lord characteristics curve methods to see the equating of the test for grades VII and VIII based on the difficulty index and the discrimination index in the test anchor.

The next step was finding the smallest error of the used equating methods. The accuracy of the method can be seen by calculating the RMSD for each method.

$$RMSD(\theta) = \sqrt{\frac{\sum_{i=1}^{N}\left(\hat{\theta}_i - \theta_i\right)^2}{N}}$$

Notes:
N = the number of the testees,
$\hat{\theta}_i$ = the first students' ability after being equated
θ_i = the first students' ability before being equated.

Findings and Discussion

Findings

Based on the item response theory used, it is necessary to find out the assumption of the item response theory. When the assumptions were met, it is possible to conduct further item response theory analysis. There are two assumptions, i.e. unidimension assumption, and local independence assumption.

The Unidimension and Local Independence Assumptions

Before testing the unidimension assumption, it is necessary to find out the adequacy of the sample through KMO-MSA and Bartlett's test of Sphrericity for the normality of the data which were used. The empirical analysis for the test for Grade VII shows that the KMO-MSA value is 0.933 with the Bartlett's test significance of 0.000. Meanwhile, for the test for grade VIII, the KMO-MSA value is 0.867 with the Bartlett's test significance of 0.000. Based on the result of the analysis, it is indicated that both instruments for grade VII and grade VIII have the KMO-MSA >0.05 and the Bartlett's test significance of <0.05, so that both tests meet the assumptions. This means that the unidimension test can be carried out.

The result of the analysis shows that the factor formed having the eigenvalue of > 1 in the test for grade VII is only one factor with the value of 5.121. The factor formed having the value > 1 is a factor that can be maintained and can be used as an indicator of a trait (Wagiran, 2014, p. 302). The eigenvalue is the highest value among the other eigenvalues so that it is indicated that the mathematics higher order thinking skill test instrument for Grade VII is unidimensional.

For the test for Grade VIII, there are four components having the eigenvalue >1, so that it is indicated that the mathematics test for Grade VIII formed four factors. The test scree plot of Grade VIII test shows that the eigenvalue became slopy starting from the second factor. Other information from the result of the analysis shows that the one dominant factor has the highest eigenvalue, that is, 4.936, so that it is indicated that the mathematics higher order thinking skill test for Grade VIII is unidimensional.

Local Independence

After being proven that the test is unidimensional, the local independence assumption is automatically proven, too (Retnawati, 2014, p. 7). Therefore, the local independence assumption for the tests for grades VII and VIII is met.

Test Item Analysis

The item analysis was carried out using the item response theory. The fitness of the model on the output PH2 in the Parscale program can be seen from the item fit statistics. In order to determine the appropriate model, the data were analyzed using the 3-logistic parameter model.

The analysis was carried out by comparing the 1-parameter logistic model, the 2-parameter logistic model, and the 3-parameter logistic model. An item is said to fit with a model when the chi-square observed is lower than the chi-square table or the significance level <α. The result of the test instrument model fitness analysis for the test for grade VII and grade VIII can be seen in Table 1.

The result of the analysis using the Parscale program shows that the mathematics test instrument for grades VII and VIII is most appropriate using the IRT analysis with the 2-parameter logistic model. This is based on the fact that the highest number of the items fitting the model is in the 2-parameter model.

The result of the analysis using the Parscale program provides information about the item parameter based on the item difficulty index (b) and the discrimination index (a). The difficulty index can be said to be good when it is in the range of -2 to +2 (Baker, 2001, p. 22; DeMars, 2010, p. 21; Hambleton et al., 1991, p. 5).

The result of the analysis of the item difficulty index for grade VII and grade VIII are presented in Table 2. In addition, the result of the item discrimination index for grade VII and grade VIII is presented in Table 3.

From Table 2, it can be seen that the item having the highest difficulty index for Grade VIII is item no 1 with a logit of 0.792 while the item having the lowest difficulty index is item no 6 with a logit of -0.807. The anchor items in the tests both for grades VII and VIII are used for further analysis. The highest difficulty index for the grade VIII test is on item no 9 with a logit of 1.456, while the item with the lowest difficulty index is item no 6 with a logit of -1.095.

Table 1. Model fitness analysis result

No	Model	Grade VII test	Grade VIII test
		Number of items (prop > 0.05)	Number of items (prop > 0.05)
1	1 PL	11	9
2	2 PL	14	13
3	3 PL	7	7

Table 2. The analysis of the item difficulty index (location)

Grade VII test			Grade VIII test		
Item	**Difficulty index**	**Category**	**Item**	**Difficulty index**	**Category**
1	0.792	Good	**1***	0.114	Good
2	-0.017	Good	**2***	-0.424	Good
3	0.105	Good	**3***	0.114	Good
4	-0.005	Good	**4***	0.777	Good
5	0.570	Good	5	0.538	Good
6	-0.807	Good	6	-1.095	Good
7*	0.044	Good	7	0.219	Good
8*	0.072	Good	8	1.145	Good
9*	0.108	Good	9	1.456	Good
10*	-0.036	Good	10	0.092	Good
11	0.498	Good	11	0.863	Good
12	0.393	Good	12	1.043	Good
13	0.592	Good	13	0.552	Good
14	0.069	Good	14	1.164	Good
15	0.470	Good	15	0.207	Good
Mean	0.190			0.451	

*: *Anchor*

Table 3 shows that all items in the test for grade VII have good discrimination index. The item having the highest discrimination index is item no 6 with a logit of 0.943, and the item having the lowest discrimination index is item no 11 with a logit of 0.041. The item having the highest discrimination index of the test for grade VIII is item no 1 with a logit of 1.377, and the item with the lowest discrimination index is item no 12 with a logit of 0.147. The item anchor in the tests for grades VII and VIII is used for further analysis.

The test set function would be higher when the test items had high information function. Standard error measurement (SEM) is closely related to the information function. The higher the information function, the smaller the SEM, and vice versa. The relation between the information function and the SEM is presented in Figure 1 and Figure 2.

Figure 1 shows that the mathematics higher order thinking skill test for Grade VII has a low score in the range between -1.4 and +1.9. This means that the test would provide higher information when it was used to measure the students' ability in the range between -1.4 and +1.9. Figure 2 shows that the test has a higher information function compared with the standard estimation error in the range between -1.2 and +2.5. Therefore, the mathematics tests were appropriate for students having the ability in the range between -1.2 and +2.5.

Table 3. The analysis of the discrimination index parameter

Grade VII test			**Grade VIII test**		
Item	**Discrimination Index**	**Category**	**Item**	**Discrimination Index**	**Category**
1	0.703	Good	**1***	1.377	Good
2	0.670	Good	**2***	0.313	Good
3	0.767	Good	**3***	0.157	Good
4	0.820	Good	**4***	0.756	Good
5	0.596	Good	5	0.736	Good
6	0.943	Good	6	0.171	Good
7*	0.606	Good	7	0.844	Good
8*	0.786	Good	8	0.745	Good
9*	0.743	Good	9	0.478	Good
10*	0.812	Good	10	1.015	Good
11	0.401	Good	11	0.928	Good
12	0.834	Good	12	0.147	Good
13	0.745	Good	13	0.352	Good
14	0.683	Good	14	0.696	Good
15	0.405	Good	15	0.791	Good
Mean	0.700		Mean	0.633	

*: Anchor

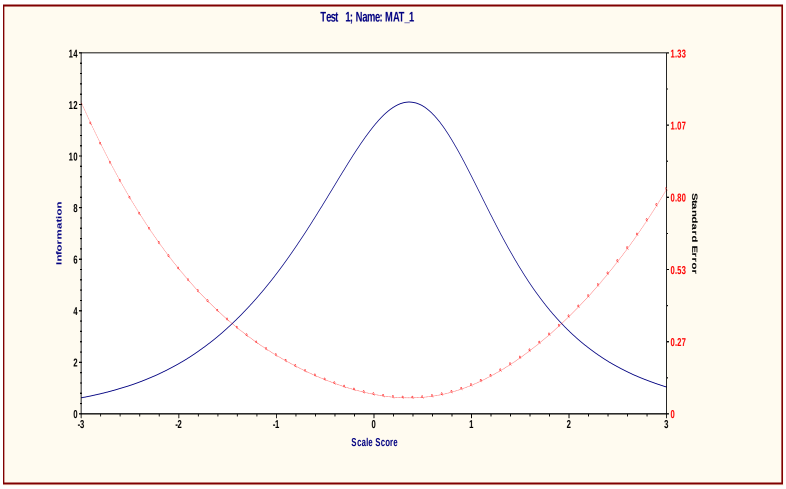

Figure 1. The relation between the information function and SEM of the test for Grade VII

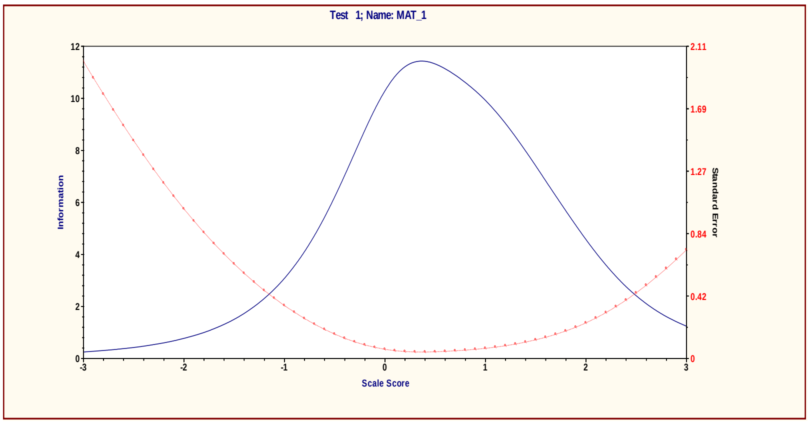

Figure 2. The relation between the information function and SEM of the test for Grade VIII

Note: -------- = *SEM*
——— = Test information

Table 4. The mean of the slope and the standard error of the equating result

Method	Equating	***Slope* mean**	SD
Mean & mean	*Slope* of classVII (X)	0.700	0.150
	Slope of class VII (Y*)	0.797	0.171
	Slope of class VIII (Y)	0.634	0.357
	WITs scale (Y*)	536.241	7.780
	WITs scale (Y)	528.835	16.233
Mean & sigma	*Slope* of classVII (X)	0.700	0.150
	Slope of class VII (Y*)	3.689	0.792
	Slope of class VIII (Y)	0.634	0.357
	WITs scale (Y*)	667.855	36.033
	WITs scale (Y)	528.835	16.233
TCC Haebara	*Slope* of class VII (X)	0.700	0.150
	Slope of class VII (Y*)	1.845	0.396
	Slope of class VIII (Y)	0.634	0.357
	WITs scale (Y*)	583.928	18.017
	WITs scale (Y)	528.835	16.223
TCC Stocking & Lord	*Slope* of classVII (X)	0.700	0.150
	Slope of class VII (Y*)	1.230	0.264
	Slope of class VIII (Y)	0.634	0.357
	WITs scale (Y*)	555.952	12.011
	WITs scale (Y)	528.835	16.223

The Equating Result

This vertical equating employed a mixed model with a 2-logistic parameter. The 2-logistic parameter includes equating the difficulty index parameter (b) and the discrimination index parameter (a). The equating method used in this study was the mean-and-mean method, mean-and-sigma method, Haebara characteristics curve, and Stocking-and-Lord characteristics curve. The vertical equating using IRTEQ produced the conversion equation: (1) Y'=0.88X-0.27 in the mean-and-mean method; (2) Y'=0.19X-0.02 in the mean-and-sigma method; (3) Y'=0.38X-0.12 in the Haebara characteristics curve method; and (4) Y'=0.57-0.18 in the Stocking-and-Lord characteristics curve.

The equating result with the parameter of slope (a) and location (b) in the mean-and-mean method, mean-and-sigma method, the Haebara characteristics curve, and also the Stocking-and-Lord characteristics curve is presented in Table 4 and Table 5, while the result of the calculation of the equating accuracy is presented in Table 6.

Table 5. The location mean and the standard error of the equating result

Method	Equating	Mean of *Location*	SD
Mean & mean	*Location* of class VII (X)	0.190	0.385
	Location of class VII (Y*)	-0.103	0.339
	Location of class VIII (Y)	0.451	0.669
	WITs scale (Y*)	495.317	15.406
	WITs scale (Y)	520.521	30.448
Mean & sigma	*Location* of class VII (X)	0.190	0.385
	Location of class VII (Y*)	0.061	0.073
	Location of class VIII (Y)	0.451	0.669
	WITs scale (Y*)	500.731	3.326
	WITs scale (Y)	520.521	30.448
TCC Haebara	*Location* of class VII(X)	0.190	0.385
	Location of class VII(Y*)	-0.048	0.146
	Location of class VIII(Y)	0.451	0.669
	WITs scale (Y*)	497.823	6.653
	WITs scale (Y)	520.521	30.448
TCC Stocking & Lord	*Location* of class VII(X)	0.190	0.385
	Location of class VII(Y*)	-0.072	0.219
	Location of class VIII(Y)	0.451	0.669
	WITs scale (Y*)	496.734	9.979
	WITs scale (Y)	520.521	30.448

Table 6. The calculation result of RMSD

Equating	Equating method	*RMSD*
Class VII to Class VIII	*Mean & Mean*	0.2955
Class VII to Class VIII	*Mean & Sigma*	0.8102
Class VII to Class VIII	*TCC* Haebara	0.631
Class VII to Class VIII	*TCC* Stocking & Lord	0.466

Discussion

The equating in this study used the mean-and-mean method, the mean-and-sigma method, the Haebara characteristics curve, and Stocking-and-Lord characteristics curve. The sample used consist of 505 grade VII students and 504 grade VIII students. This was based on the minimum sample measure in item response theory with the 2-logistic parameter, that is, 500 respondents (DeMars, 2010, p. 34). The item anchor used was four items or 26.7%. The number of the anchor influences the test equating result (Kartono, 2008, p. 303). The anchor must be at least 20% of the number of test items (Kolen & Brennan, 2014, p. 288). The test characteristics based on item response theory resulted in the mean of the parameter of the item difficulty index or location (b) in the good category in the range between $-2 < b < 2$, that is 0.190 and 0.451 successively. The mean of the discrimination index or slope (a) for grade VII and Grade VIII was 0.701 and 0.634 successively. Based on the item difficulty index, these items were in a good category because they lied in the range $-2 < b < 2$.

The calculation result of the equating constant based on anchor items results in some equations. The equations obtained using the mean-and-mean, mean-and-sigma, Haebara characteristics curve, and Stocking-and-Lord characteristics curve methods are Y'= 0.88X–0.27, Y'=0.19X-0.02, Y'=0.38X–0.12, and Y'=0.57X–0.18 successively.

The findings of the research show that the score conversion of the parameter of location and slope indicate consistent results in the mean-and-mean method, mean-and-sigma method, Haebara characteristics curve method, and also Stocking-and-Lord characteristics curve method. An examples of the equating result using the mean-and-mean method can be seen in item no 5 for grade VII. The equation of the parameter location of Grade VII to Grade VIII is:

$$b^* = 0.88\,(b_x) - 0.27.$$

Item no 5 for grade VII has the difficulty index (location) of 0.57 logit. Thus, after being equated to Grade VIII location, it becomes b*=0.232. This means that item no 5 Grade VII has the location of 0.57 logit being equal with the location value of 0.232 in the item for Grade VIII. The result of the equating of the item difficulty index (location) shows that the item difficulty index for Grade VIII experiences a decrease after being converted to Grade VIII. If the item no 5 for Grade VII was done by students in Grade VIII, the VIII grade students would find it easier.

The result of the equating of location in the four methods shows that the test for Grade VIII is more difficult than the test for Grade VII. This information can be seen from the comparison table of the parameter scores of the item difficulty index before and after being equated in each method. The mean score of the item difficulty index parameter decreases when converted to a higher scale. This means that the test for Grade VII is easier when it is done by Grade VIII students. On the other hand, the test for Grade VIII students would be more difficult when it was done by Grade VII students.

The equating result of the discrimination index parameter scores (slope) can be illustrated in one of the items of the test for Grade VII. The discrimination index parameter conversion equation from Grade VII to Grade VIII with the mean-and-mean method is:

$$a^* = \frac{a_x}{0.88}.$$

Item no 5 for Grade VII has the discrimination index of 0.596 logit. Therefore, after being equated to Grade VIII, the discrimination index would be b*=0.677. This means that the discrimination index for Grade VII, that is, 0.596 is equal with the discrimination index of 0.677 for Grade VIII. The equated discrimination index of item no 5 increases after being equated to Grade VIII. All the four methods provide consistent information. This can be seen on the comparison of the mean of the discrimination index of the items for Grade VII and for Grade VIII which has been equated (in WITs scale). It is indicated that after being equated, the test instrument for Grade VII has a higher discrimination index than the test instrument for Grade VIII.

The method which provides the smallest error in the equating using 2-logistic parameter is the mean-and-mean method. This, then, is followed by the Stocking-and-Lord method, the Haebara method, and the mean-and-sigma method successively. Baker and Al-Karni (1991) state that the mean-and-mean method has better accuracy than the Stocking-and-Lord method. A study which was conducted by Kartono (2008) concludes that the equating using the mean and mean method is one level better than the mean and sigma method. Sugeng (2010, p. 289) states that the mean and mean method tends to provide more accurate information than the IRT vertical equating using the partial credit model. Uysal and Kilmen (2016) present their study stating that the Stocking-and-Lord method has a smaller error than the Haebara method. The mean-and-sigma method results in the biggest error, while the sample size and the distribution of the students' ability influences the RMSD value (Kilmen & Demirtasli, 2012, p. 130).

Conclusion and Suggestions

Conclusion

The characteristics based in the item response theory results in the mean of the parameter value of the item difficulty index or location (b) which is categorized as good in the range of -2 < b < 2. The parameter values are 0.190 and 0.451. The mean of the discrimination index parameter value or slope (a) for the tests for Grades VII and VIII are 0.701 and 0.634 successively.

The equating results in four equations based on the method used, that are, Y'= 0.88X–0.27 using the mean-and-mean method, Y'=0.19X-0.02 using the mean-and-sigma method, Y'=0.38X–0.12 using the Haebara characteristics curve method, and Y'=0.57X–0.18 using the Stocking-and-Lord characteristics curve method.

The calculation of the equating accuracy results in the Root Mean Square Difference (RMSD) of the mean-and-mean method, the mean-and-sigma method, the Haebara characteristics curve method, and the Stocking-and-Lord characteristics curve method of 0.2955, 0.8102, 0.6315, and 0.466 successively. The mean-and-mean provides the smallest RMSD, followed by the Stocking-and-Lord characteristics curve method, the Haebara characteristics curve method, and the mean-and-sigma method.

Suggestions

The study related to the students' mathematics higher order thinking skill development is still limited, that is, it is only concerned with the ability development of grade VII and grade VIII students. Further studies need to be carried out for the ability development from grade VII to grade VIII and from grade VIII to grade IX. In addition, it is suggested that the use of the methods be studied further with different logistic parameters and different lengths of tests to get more accurate information.

The students' mathematics higher order thinking skill ability in Deli Serdang District in this study is still low. Teachers are expected to teach materials with varied cognitive domain as suggested by the curriculum implemented in the schools.

The school principals play an important role in the advancement of the educational institution so that it is suggested that every year a test to know the students' higher order thinking skill development be administered. In addition, it is also necessary for the school to provide a kind of training for the teachers to analyze test items using the classic and modern analyses so that they can develop better items to depict the students' ability in different grades.

References

Antara, A. A. P., & Bastari, B. (2015). Penyetaraan vertikal dengan pendekatan klasik dan item response theory pada siswa sekolah dasar. *Jurnal Penelitian Dan Evaluasi Pendidikan*, *19*(1), 13–24. https://doi.org/10.21831/pep.v19i1.4551

Baker, F. B. (2001). *The basics of item response theory* (2nd ed.). College Park, MD: ERIC Clearinghouse on Assessment and Evaluation.

Baker, F. B., & Al-Karni, A. (1991). A comparison of two procedures for computing IRT equating coefficients. *Journal of Educational Measurement*, *28*(2), 147–162. https://doi.org/10.1111/j.1745-3984.1991.tb00350.x

Crocker, L. M., & Algina, J. (2008). *Introduction to classical and modern test theory*. Mason, OH: Cengage Learning.

DeMars, C. (2010). *Item response theory: Understanding statistics measurement*. New York, NY: Oxford University Press.

Field, A. P. (2000). *Discovering statistics using SPSS for Windows: Advanced techniques for the beginner*. London: Sage Publications.

Gagné, R. M. (1977). *The conditions of learning*. New York, NY: Holt, Rinehart, and Winston.

Hambleton, R. K., & Swaminathan, H. (1985). *Item responsse theory*. Newburg Park, LA: Sage Publication ICC.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage Publications.

Han, K. T. (2009). IRTEQ: Windows application that implements item response theory scaling and equating. *Applied Psychological Measurement*, *33*(6), 491–493. https://doi.org/10.1177/0146621608319513

Istiyono, E. (2016). The application of GPCM on MMC test as a fair alternative assessment model in physics learning. In *Proceeding of the 3rd International Conference on Research, Implementation and Education of Mathematics and Science (ICRIEMS), 16-17 May 2017* (pp. 25–30). Yogyakarta: Universitas Negeri Yogyakarta. Retrieved from http://seminar.uny.ac.id/icriems/sites/

seminar.uny.ac.id.icriems/files/prosiding/PE-04.pdf

Kartono, K. (2008). Penyetaraan tes model campuran butir dikotomus dan politomus pada tes prestasi belajar. *Jurnal Penelitian Dan Evaluasi Pendidikan*, *12*(2), 302–320. https://doi.org/10.21831/pep.v12i2.1433

Kilmen, S., & Demirtasli, N. (2012). Comparison of test equating methods based on item response theory according to the sample size and ability distribution. *Procedia - Social and Behavioral Sciences*, *46*, 130–134. https://doi.org/10.1016/J.SBSPRO.2012.05.081

Kolen, M. J., & Brennan, R. L. (1995). *Test equating: Methods and practices*. New York, NY: Springer New York.

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

Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking: Methods and practices*. New York, NY: Springer.

Law No. 14 of 2005 of Republic of Indonesia about Teachers and Lecturers (2005).

Law of Republic of Indonesia No. 20 of 2003 on National Education System (2003).

Mardapi, D. (2012). *Pengukuran, penilaian, dan evaluasi pendidikan*. Yogyakarta: Nuha Medika.

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

Stiggins, R. J., & Chappuis, J. (2012). *An introduction to student-involved assessment for learning*. Boston, MA: Pearson.

Sugeng, S. (2010). Penyetaraan vertikal model kredit parsial soal matematika SMP. *Jurnal Penelitian Dan Evaluasi Pendidikan*, *14*(2), 289–308. https://doi.org/10.21831/pep.v14i2.1083

Uysal, I., & Kilmen, S. (2016). Comparison of item response theory test equating methods for mixed format tests. *International Online Journal of Educational Sciences*, *8*(2), 1–11.

Wagiran. (2014). *Metode penelitian pendidikan: Teori dan implementasi*. Yogyakarta: Deepublish.