TECHNIUM

Technium Vol. 3, Issue 7 pp.110-120 (2021)
ISSN: 2668-778X
www.techniumscience.com

Vertical Equating Accuracy Using Kernel Method

Muhammad Ali Gunawan[1], Heri Retnawati[1], Badrun Kartowagiran[1]

[1] Education Research and Evaluation Study Program, Yogyakarta State University, Jl. Colombo Yogyakarta No. 1, Karang Malang, Caturtunggal, Depok District, Sleman Regency, Special Region of Yogyakarta 55281, Indonesia

Email : guns12380@gmail.com

Abstract. This study aims to reveal: (1) the methods that provide more accurate results on vertical equating with the classical approach (linear and equipercentile) with the kernel method; (2) the methods that provide more accurate results on the vertical equating of the IRT approach with the kernel method. This research is simulation research using the Markov Chain Monte Carlo (MCMC) method. The experiment used a factorial design 3 factorial (treatment) and 40 repetitions. The three controlled factors are: 1) ability parameters (θ = 0.314; 0.431; and 0.451); 2) the number of samples (500, 1000, and 5000) and 3) the length of the test (10, 20 and 30). The data were generated using the WinGen3 application with one parameter logistic model (1PL / Rasch Model). The analysis was carried out by comparing the mean of standard error equating (MSE). The results of this study are as follows. (1) The vertical equating method using the Item Response Theory (IRT) approach, especially the Haebara and Stocking-Lord methods, is more accurate than the classical method (equipercentile and linear). (2) The non-kernelled IRT approach method is more accurate than the kernelled IRT method (smoothing with the Gaussian kernel).

Keywords: Classical Equating Method, IRT Method, Kernel Method, Test Equating Accuracy, Vertical Equating.

1. INTRODUCTION

The quality of assessment instruments will continue to be needed in all fields, especially in the field of education which carries out periodically assessments (von Davier, 2011:21), hence the issue of test security (Suhardi, 2020; Gunartha, et.al., 2020) and the accuracy of information on student learning outcomes (Retnawati, 2016) through the test package used is very urgent (Khasanah, Ramli and Dwiastuti, 2020). So it is necessary to scores equating from various test packages that are used interchangeably and meet the principle of fairness in the realm of measurement and assessment (Dorans, Moses, and Eignor, 2010:1). In the sense that score equating seeks to eliminate the effect of unintentional difference in scores on the difficulty parameters of different test packages, especially on vertical equating (Nisa and Retnawati, 2018). Equating of scores is necessary to be fair to examinees and to provide users with scores that have the same meaning across different editions or test packages used. Thus, the package of questions to be tested is able to realize the uniqueness of each school and region in the learning materials taught in schools (Andrian, Kartowagiran and Hadi, 2018) and especially regarding the accuracy of the assessment instruments used (Retnawati, et.al, 2017).

Efforts to develop quality instruments (valid and reliable) through comprehensive assessments using test equivalency methods have been carried out by many previous researchers, both with the classical theory approach and item response theory (IRT). Most of these studies were conducted with a relatively large number of samples (n = 20,000 and above). Meanwhile it in a very limited number of students

schools (Kartowagiran, et.al. 2019). This is where the results of developing tests or theories on the equating of ICT or IRT have not been maximally practiced by teachers in schools, because after all, parametric statistical assumptions in analyzing sample data such as the normal distribution are very difficult to obtain. When these parametric assumptions are not met, the estimation accuracy becomes very weak (Yu & Zuan, 2012:190-213).

On the other hand, the practice of a test equating contains *errors*, Kolen and Brennan (2014: 247) mention two sources of estimation error of test equating, namely random errors and systematic errors. Random equating error is a sampling error that occurs when a group of examinees is randomly drawn from a population, and their scores are used to estimate the population equating. Systematic equating errors can occur in the following situations: (1) smoothing equating; (2) violation of statistical assumptions; (3) the data collection design is not implemented properly; and (4) the sample of examinees used for equating is very different from the study population. In fact, most teachers have not been able to properly compile and analyze instruments according to the existing measurement theory (Kartowagiran and Jaedun, 2016).

To overcome the problems encountered in the implementation of equating tests, especially with regard to small samples, Kim, von Davier and Haberman (2006) which was inspired by the research of Hanson, Zeng, and Colton (1991) which examined the function of various equating methods with samples ranging from 100 up to 3,000 observations. They found that, with a sample of 100, the identity function gave a lower equating error (including bias and sample error) than the others. Badrun, et.al (2019), Asiret and Sunbul (2016), Pramudita, Rosnawati, and Mam (2019), also conducted research with a relatively small sample size of 320 test participants, this study proves that with a small sample size, parallel tests can be developed/tested. Therefore, further research on the equating of tests with small samples is needed so that it can be applied by teachers in schools. At least, at the regional level (one Regency/City) represented by the Subject Teacher Consultation (MGMP) as a question developer with a relatively small number of samples.

Equating tests with the kernel method were first introduced by Holland and Thayer (1989) and further developed by von Davier et al. (2004b) with a concentration on the Gaussian Kernel (GK) Davier (2011:160). Kernel equating is an equating method that is carried out to reduce bias and standard error of equating which is usually assumed that smoothing as a characteristic of population score distribution, irregularity indicates sampling error. To reduce sampling error (i.e., random error) and to improve equating performance, Rosenbaum and Thayer (1987), as well as Holland and Thayer (1987) have suggested a process called presmoothing that performs log-linear smoothing on discrete score distributions before applying equating procedure.

Several recent studies related to equating of tests using the kernel method include: Wang, Zhang and You (2020) in a random equating group design, kernel equating of IRT scores was observed to be more accurate and stable than others; Wolkowitz and Wright (2019) As the sample size increases, the amount of bias in equating passing scores decreases ; Arikan (2019) results based on the equipercentile and linear equating methods are consistent with each other, except for a high range of score scales ; Arikan and Gelbal (2018) the results of kernel equating are almost the same as those of the appropriate traditional equating method; De Ayala, Smith and Dvorak (2018); in general, both KE and TCC equating yielded accurate results, although KE tended to outperform TCC on a parametric *true score* scale across conditions ; Liu and Kolen (2018) declared that presmoothing can reduce equating errors random and equating total error ; Netto (2018) the IRT method tends to give better results than KE even though the data are not generated from the IRT process ; Wiberg (2016) local linear (kernel) IRTOSE method produces low bias and low measurement error values.

The existence of differences in research results from the several studies mentioned above, indicates that there is a great opportunity to conduct research on the accuracy of the equating of the kernel method with the current methods, namely the classical method (equipercentile and linear) and the item response theory method (Mean-Mean, Mean-Sigma, Haebara and Stocking-Lord). This study aims to determine (1) the accuracy of the vertical equating method with the classical approach (linear and equipercentile) with the IRT approach (Mean-Mean, Mean-Sigma, Haebara and Stocking-Lord); (2) The accuracy of the kernelized and non-kernelized IRT approach vertical equating method.

2. METHOD

Technium Vol. 3, Issue 7 pp.110-120 (2021)
ISSN: 2668-778X
www.techniumscience.com

This research is a simulation research using the *Markov Chain Monte Carlo* (MCMC) method, namely research conducted with fictitious data using parameters from the results of previous studies (data generated with predetermined conditions), namely the research of Retnawati and Wulandari (2019) which is used as *posterior* data.

This type of research simulations have been selected for this study involved a number of treatments that are not possible to be done on the real conditions in the field (Retnawati, 2008), which is the result should be equal or close to the real conditions, it is used the parameters of previous research results that parameter ability of participants learn and item parameters. The controlled conditions in this study were the parameters of the ability of students, the number of samples, the length of the test and the equating method used (classic, IRT and kernel) with vertical equating method .

The equating design used is the equivalent group design (EG) without using shared items, where vertical equating is carried out with the design conception written by von Davier (2011: 61), namely randomly equal groups of students at the same grade level given different block items and most of the blocks are assigned to groups at adjacent grade levels. In this design, one sample of students at each grade level takes a test form including the same item block as the adjacent class below it, one sample takes an item block for that class only, and a third sample takes the same item block as the adjacent class above it. For the lowest and highest values in the design, they can of course only share a block of items with one contiguous class, as illustrated in the following table.

Table 1. Equivalent Group Design with Joint Item Blocks in Adjacent Classes

	Question Item Level					
Grade	SS8.1	SS8.2	SS9.1	SS9.2	SS10.1	SS10.2
8.A	X	X				
8.B		X	X			
9.A		X	X			
9.B			X	X		
9.C				X	X	
10.A				X	X	
10.B					X	X

Note: SS = Examination Semester Test Form.

The sample group is assumed to be equal based on the grade of the test used, where grade 8 semester 2 takes a test for semester 1 for grade 9, and grade 9 for semester 1 takes a test for semester 2 for grade 8. Furthermore, grade 9 semester 2 takes a test for grade 10 in semester 1 and grade 10 in semester 1 takes a test for semester 2 in grade 9 (von Davier, 2011:62). One important aspect of the content coverage problem for vertical-scale designs is that managing items with students graded above will not be appropriate if students have not been shown the relevant content in item blocks. Therefore it is assumed that the material taught in the higher class is material based on the learning material in the previous class (learning continuum).

The variables controlled in this study were (1) ability based on grade level (grade) (θ); (2) the number of samples (n) and (3) the length of the test. So this research contains 3 factorial or 9 treatments. Simulation data generated by WinGen3 version 3.1.15.455. The assumption test is carried out at the initial stage after the data is generated, namely the item response theory assumption test (unidimensionality, local independence, and invariance) and the equating assumption test is carried out on all data packages (1,080 data package). The assumption test was carried out using a combination of SPSS20, Excel and R.

The equating method used is the classical equating method (Linear and Equipercentile), the item response theory (IRT) method, namely: Mean-Mean, Mean-Sigma, Haebara and Stocking-Lord and Kernel equating on the classical and IRT methods. Equating analyzed using R package **SNSequate** and **kequate.** Equating results were evaluated by looking at the standard *Error of Equating* (SEE) value for all pairs of data form/test form. The mathematical formula for standard error of equating in the *Equivalent Group* (EG) design is.

Technium Vol. 3, Issue 7 pp.110-120 (2021)
ISSN: 2668-778X
www.techniumscience.com

$$SEE_Y(x) = \hat{\sigma}_Y(x) = \sqrt{Var(\hat{e}_Y(x))}$$

With the provisions, the smaller of standard error equating indicating the more accurate the equating result, on the contrary, the greater the standard error of equating, the less accurate the equating result.

3. RESULTS

The presentation of the results of vertical equating consists of: (1) equating with the classical method (equipercentile and linear), (2) equating using the IRT approach, and (3) equating with the Kernel method, including the classic method and the kernelized IRT.

a. Equating Accuracy With Classical Method (Equipercentile and Linear)

Vertical equating with classical methods (equipercentile and linear) obtained the following results.

Table 2. Mean of Standard Error Equating Using the Classical Method (Equipercentile and linear)

Code	Mean of Standard Error Equating	
	Equipercentile	Linear
D8.3-D9.8	0,19794	0,03268
D8.1-D9.7	0,21049	0,03284
D9.6-D9.9	0,25909	0,03590
D9.1-D10.1	0,22134	0,03546
D9.7-D10.2	0,20580	0,03609
Rata-rata	**0,21893**	**0,03459**

Based on the table 2. above, obtained the mean of standard error of vertical equating by using classical methods equipercentile amounted to 0.21893 and the linear method is 0.03459, it indicates that the linear method is more accurate than the equipercentile method. As also shown in the following graph, where in all data packages, we found the mean of standard error equating with the linear method is much smaller than the equipercentile method.

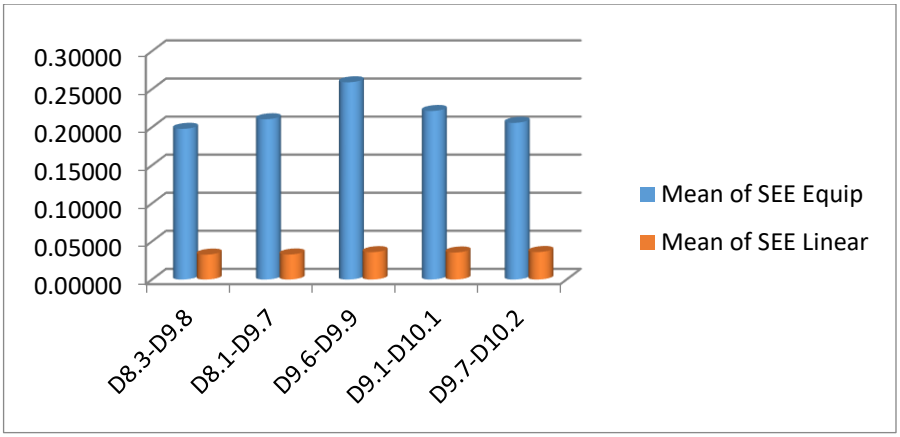

Figure 1. Mean of Standard Error Equating for Classical method (equipercentile and linear)

b. Vertical Equating with the IRT Method

The coefficient equating and standard error equating using the Item Response Theory method is obtained as presented in the following table.

Table 3. Mean Standard Error Equating with IRT Method

Code	Mean of Standard Equating Error			
	Mean-Mean	**Mean-Sigma**	**Haebara**	**Stocking-Lord**
D8.3-D9.8	0,0220	0,0275	0,0218	0,0221
D8.1-D9.7	0,0220	0,0265	0,0216	0,0220
D9.6-D9.9	0,0221	0,0276	0,0222	0,0221
D9.1-D10.1	0,0221	0,0282	0,0214	0,0221
D9.7-D10.2	0,0220	0,0274	0,0213	0,0220
Average	**0,0220**	**0,0274**	**0,0216**	**0,0221**

Table 3. above shows that the smallest of mean of standard error equating is Haebara method (0.0216) compared with three other IRT equating method, followed by Mean-Mean method (0.0220), the Stocking-Lord method (0.0221), and the last is the Mean-Sigma method (0.0274). As also shown in the following graph.

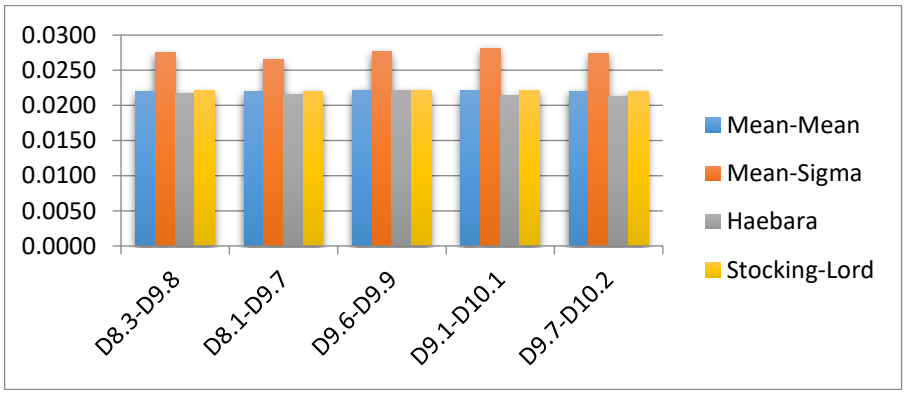

Figure 2. Mean of Standard Error Equating Using IRT Method

Based on the graph above, it can be said that the Haebara and Mean-Mean methods have much higher accuracy than the other two methods, namely the Stocking-Lord and Mean-Sigma methods. This result is compared with the classical equating method (equipercentile and linear), the results will be obtained as listed in the following table.

Table 4. Mean of Standard Error Equating with Classical and IRT Method

Code	Classical Test Theory (CTT)		Item Response Theory (IRT)			
	Equip.	**Linear**	**MM**	**MS**	**Hbr**	**SL**
D8.3-D9.8	0,19794	0,03268	0,02203	0,02746	0,02175	0,02207
D8.1-D9.7	0,21049	0,03284	0,02195	0,02646	0,02162	0,02197
D9.6-D9.9	0,25909	0,03590	0,02215	0,02765	0,02216	0,02213
D9.1-D10.1	0,22134	0,03546	0,02207	0,02815	0,02138	0,02208

D9.7-D10.2	0,20580	0,03609	0,02203	0,02740	0,02129	0,02202
Mean	**0,21893**	**0,03459**	**0,02204**	**0,02742**	**0,02164**	**0,02205**

Note: Equip = Equipercentile, MM = Mean-Mean, MS = Mean-Sigma, Hbr = Haebara, SL = Stocking-Lord

Table 4. above shows that vertical equating using the *item response theory* (IRT) method is more accurate than the classical method, where the Haebara and Mean-Mean methods have the smallest mean of standard error equating (Haebara = 0.02164 and Mean-Mean = 0.02204) compared to the linear method (more accurate than the equipercentile method) in the classical method, with the mean of standard error equating is 0.03459.

If the order of accuracy of all the equating methods is made, then the most accurate method is the Haebara method, followed by the Mean-Mean method, the Stocking-Lord method, the Mean-Sigma method, then the classical (linear) method in fifth order, and Equipercentile method occupy the last position. For more details, see the following graph.

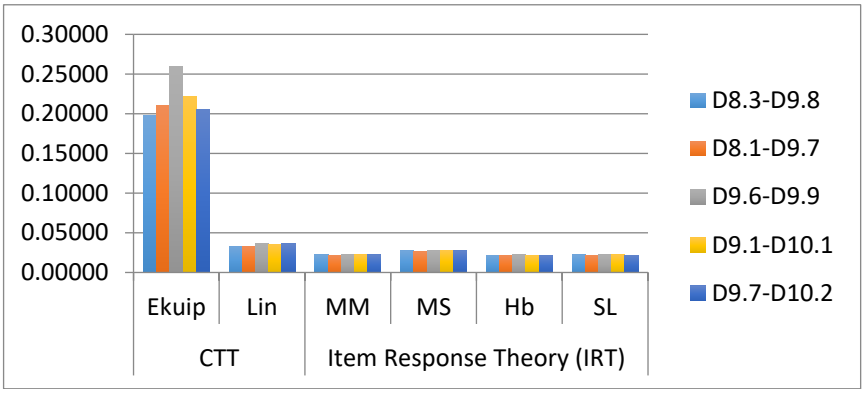

Figure 3. Mean of Standard Error Equating with Classical and IRT method

Based on the graph, it can be clearly seen that the Haebara and Mean-Mean methods are much smaller than the classical methods, both linear and equivalent percentile methods. In other words, the IRT method (Haebara and Mean-Mean) is much more accurate than the vertical equalization method with the classical method (equipercentile and linear).

c. The Accuracy of Vertical Equating with IRT Method and IRT-OSE Kernel.

The difference between true-score and observed score equating is that true-score IRT equating based on the mean of the conditional score distribution whereas observed-score IRT equating based on the marginal score distribution and uses the IRT model to define the probability conditional scores involved (Gonzalez and Wiberg, 2017:120). On the R software, there are so many packages that can be *downloaded* and used as needed as equating package with ***kequate*** developed by Bjorn Anderson, Kenny Branberg and Marie Wiberg; and ***SNSequatue*** developed by Jorge Gonzalez.

The results equating using the IRT and IRT-OSE (kernelized IRT) methods are shown in the summary table regarding the mean of standard error equating juxtaposed with the results of the following non-kernelized IRT equivalence.

Table 5. Average Error of Equalization Standards on Equalization of IRT and IRT-Kernel

Code	**Mean of Standard Error Equating**				
	MM	**MS**	**Hbr**	**SL**	**IRT-OSE**
D8.3-D9.8	0,0220	0,0275	0,0218	0,0221	0,14425

D8.1-D9.7	0,0220	0,0265	0,0216	0,0220	0,14449
D9.6-D9.9	0,0221	0,0276	0,0222	0,0221	0,14600
D9.1-D10.1	0,0221	0,0282	0,0214	0,0221	0,14540
D9.7-D10.2	0,0220	0,0274	0,0213	0,0220	0,14597
Rata-rata	**0,0220**	**0,0274**	**0,0216**	**0,0221**	**0,14522**

Note: MM = Mean-Mean, MS = Mean-Sigma, Hbr = Haebara, SL = Stocking-Lord, IRT-OSE = Kernel IRT

Based on the table above, the mean of standard error equating of kernelized IRT is much larger than non kernelized IRT. Where the mean of SEE is 0.147826. With such mean standard error of equating, the kernelized IRT equating method ranks third after the Stocking-Lord and Haebara methods as shown in the following graph.

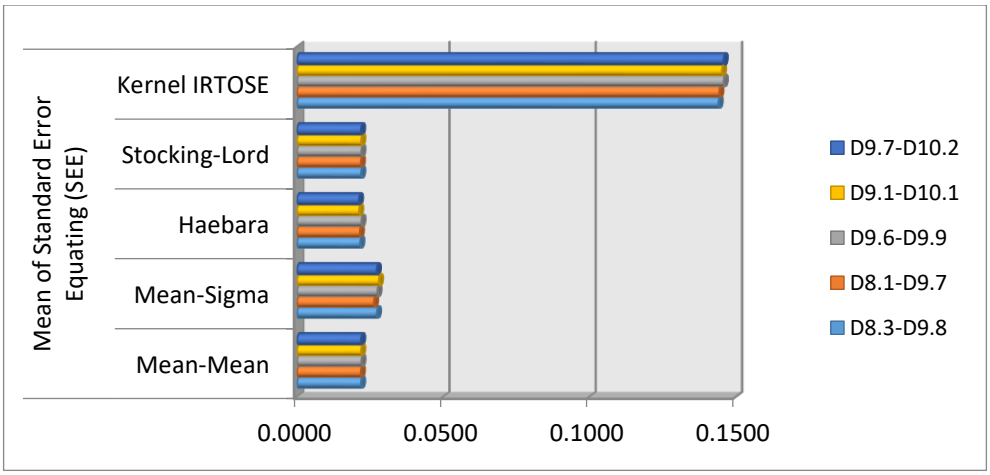

Figure 4. Mean of standard error Equating with IRT and IRT-Kernel method

In the graph above, it can be seen that the IRT method is more stable and smaller than the kernelized IRT method.

d. The Accuracy of Classical-Kernel Method

The kernelized classical method is carried out by the steps namely: 1) pre-smoothing 2) scoring probability estimation, 3) discrete score distribution continuity, 4) calculating and diagnosing the equating results, and 5) evaluate the accuracy of the equating with the standard error of equating (SEE). Table below shows the results of the classical and kernel-classical method.

Table 6. Mean of Standard Error Equating with Classical and Kernel-Classic Method

Code	Equipercentile	Linear	Kernel-Equip	Kernel-Lin
D8.3-D9.8	0,19794	0,03268	0,19471	0,17102
D8.1-D9.7	0,21049	0,03284	0,19687	0,17148
D9.6-D9.9	0,25909	0,03590	0,21062	0,17774
D9.1-D10.1	0,22134	0,03546	0,20809	0,17681
D9.7-D10.2	0,20580	0,03609	0,20487	0,17842
Rata-rata	**0,21893**	**0,03459**	**0,20303**	**0,17509**

Technium Vol. 3, Issue 7 pp.110-120 (2021)
ISSN: 2668-778X
www.techniumscience.com

According to the table above, we found the mean of standard error equating using classical methods equipercentile is 0.21893 and linear method is 0.03459. While the mean of standard error equating using kernel on classical methods decreased to 0.20303 and the linear kernel methods increased to 0.17509.

It can be said that the kernel equating can reducing the standard error equating for equipercentile method, but it cannot be applied to the linear method. This is in accordance with von Davier (2011:141-142) that the traditional equating procedure (CTT) based on percentile rank can be viewed as an advanced procedure for uniform kernels. Procedure uniform kernel also has a linear cumulative distribution, will however not an ideal procedure.

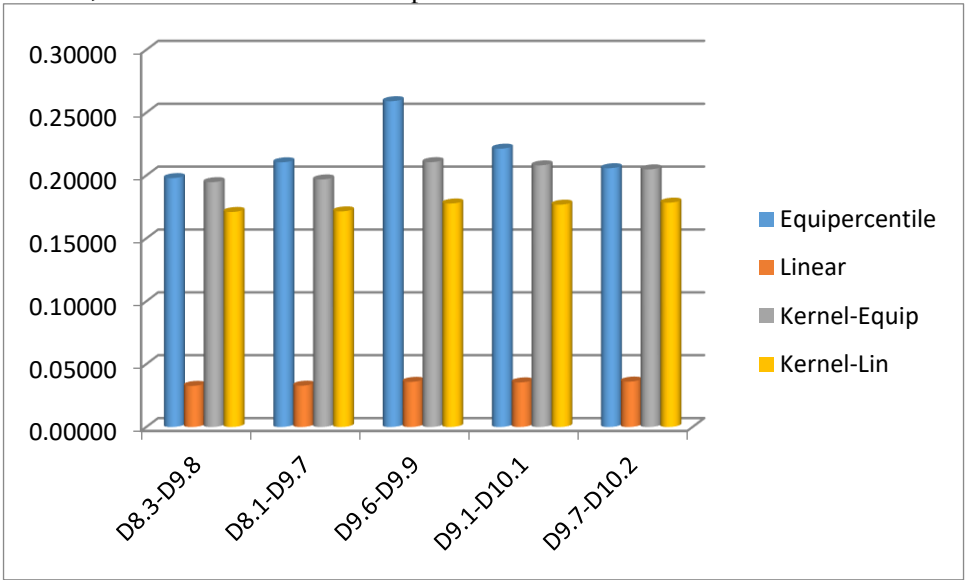

Figure 5. Mean of standard error equating with classic and kernel-classic method

4. DISCUSSION AND CONCLUSSION

The accuracy of vertical equating in the IRT method (especially Haebara and Stocking-Lord) is higher than the classical equivalence method (equipercentile and linear). This is evidenced by the mean standard error of equating produced by the two methods, where the smallest standard error of equating produced by the Haebara method is 0.0223, followed by the Stocking-Lord method with an SEE 0.0224, at third place is occupied by the Mean-Mean method which produces an SEE 0.0227, the fourth position is occupied by the Linear method with SEE 0.0307. The fifth position is occupied by the Mean-Sigma method with SEE 0.034. While the sixth and last is kernelized IRT method (IRT-OSE) SEE 0.1372 and equipercentile method with SEE 0.2347.

That is, if the comparison uses the smallest mean of standard error equating in each method, namely the Haebara method with SEE 0.0223 and a linear method with SEE 0.0307. So it can be said that the vertical equating with IRT method (Haebara and Stocking-Lord) is more accurate than classical method (equipercentile and linear).

Furthermore, the equating with the non-kernelized IRT method (especially the Haebara and Stocking-Lord methods) results in more accurate results than the kernelized (smoothing) IRT method. Where the mean of standard error equating using Haebara obtained a value of 0.0223 and by the method of Stocking-Lord of 0.0224. While using kernelized IRT gained mean of standard error equating is 0.1372.

Based on the results of the data analysis as above, it can be said that the application or use of the kernel method in the IRT equating method is a special case or the same as the classical linear equating that was kernelized, because the kernel method (smoothing) can cause an increasing the standard error of equating or reduce the accuracy of the equating results**.** The results of this study are the same as the findings of Liu and Kolen (2018) in their research entitled Comparison of Parameter Smoothing Strategy Selection in Mixed Format Tests under Randomized Group Design. Where the research shows that the effectiveness of *presmoothing* in reducing the total equating error depends on the model selection strategy, and an inappropriate model selection

strategy can lead to higher equating errors. This finding is also similar to the results of Mao's (2006, in Arikan and Gelbal, 2018), that the higher error at the end point with the Kernel equating method is due to the fact that the point scale in the Gauss Kernel continuation method is in the range from +∞ and – .

This finding is different from the results of research conducted by Wang, Zhang and You (2020) that in a random equivalent group design with a sample size of 10,000. The IRT kernel equating score was observed to be more accurate and stable than the others. In the non-equivalent group with the anchor test design, equating of IRT scores was observed to show the lowest systematic and random error among the equating methods studied. This difference in results could be caused by the smoothing model used and the number of parameters in the item response theory (3 logistic parameter model) and the number of samples, while in this paper, we used the Rasch model (1 logistic parameter IRT) with a relatively small sample size of 500, 1000 and 5000, as revealed in the research by Liu and Kolen (2018) above. These different results can also be caused by different definitions (Harjo, Kartowagiran, and Mahmudi, 2019) regarding the kernel equating itself.

This study result also differs from that of De Ayala, Smith and Dvorak (2018) which showed that in general, both KE and TCC equivalence yielded accurate results, although KE tended to perform better than TCC on a parametric *true score* scale in all conditions. This difference in results can be said to be reasonable because the data used in this study are observed scores and even though using the same parameters may not guarantee that identical simulation results will be obtained because the random number generating mechanism may differ from one computer to another, or in different versions with the same software program (Bulut and Sunbul, 2017). Moreover, the research of De Ayala, Smith and Dvorak (2018) uses true scores. It is also explained by Liu and Kolen (2018) that this may occur because post smoothing directly smooth the equivalence function, while presmoothing smoothes the observed score distribution with different C parameters for Form X and Form Y (test equating form).

The difference in results in several studies referred to is also explained by the reasons for the research conducted by Luecht and Ackerman (2018), which is caused by the residual factor of the parameters used in equating tests such as discriminatory power parameters, level of difficulty and ability of test participants. In this study, the Rasch model (IRT 1 parameter) was used so that the residue in the analysis results was not too large (only on the ability level factor with the difficulty of the test). In addition, the modeling of the data generation process on the learning outcomes test (education) is more adapted to the log-linear model, spline function and IRT model which have different assumptions. For example, in observed score equating (OSE), the distribution of model estimation test scores is linked using the equipercentile function. As it is known, that the equipercentile function is a composition of mathematical functions that requires continuous data, while in reality the test scores (multiple choice) are not, so that the learning outcomes data (test) need to be carried out continuously. Meanwhile, the continuity of the score distribution function involves approaches commonly used in probability theory and statistical theory (von Davier, 2011:4). So that in terms of this continuity process, it must also be based on certain assumptions, this is what causes the difference in results.
Overall, based on the results of the analysis and discussion described above, it can be concluded that (1) the vertical equating method using the item response theory approach (IRT, especially the Haebara and Stocking-Lord methods) is more accurate than the classical method (equipercentile and linear); (2) The non-kernelized IRT approach method is more accurate than the kernelized IRT method (smoothing with Gaussian kernel).

From the results of this study, several things were found, especially regarding the effect of the kernel equating design (smooth distribution of scores) both with the classical method and with the item response theory approach, however, there are still many shortcomings or limitations that require further and in-depth study, including: 1) This research is only at the stage of using the Gaussian kernel equating method to compare the accuracy of the equating results, has not led to comparisons with other kernel methods such as the uniform kernel, triangle kernel, Epanechnikov kernel, quartic kernel, triweight kernel, and cosine kernel. Further research is needed regarding the accuracy of the results equating differences in terms of the method of smoothing kernel; (2) The accuracy of the equating results is proven theoretically and statistically influenced by the equating design (data collection) and method used, while in this study it is still limited to comparing the accuracy of the equating results in terms used of the methods of equivalent group design. examine the effects of using various equating designs such as single group design, counter balanced groups, Nonequivalent groups with

anchor test; (3) This research is limited to simulation research which has quite a number of weaknesses including the weakness of differences in the use of software/applications and the time of data generation, although the parameters used are derived from the results of previous studies, but the use of different software or the implementation of data generation at the same time. Different results can also produce different final results, so further study is needed on the accuracy of the software used based on existing theories. Software development in the field of assessment and measurement of simple and integrated much needed.

Thanks to:
PUSPENDIK BALITBANG Kemendikbud RISTEK of the INDONESIAN REPUBLIC which has provided National Examination response data to PEP Yogyakarta State University to be analyzed as student learning materials.

5. REFERENCE

[1]. Andrian, D., Kartowagiran, B., & Hadi, S. (2018). The Instrument Development to Evaluate Local Curriculum in Indonesia. *International Journal of Instruction, October 2018, Vol.11, No.4*

[2]. Arikan, C. A., & Selahattin, G. (2018). A Comparison of Traditional and Kernel Equating Methods. *International Journal of Assessment Tools in Education, 2018, Vol. 5, No. 3, 417–427*

[3]. Arikan, C. A. (2019). A Comparison of Kernel Equating Methods Based on Neat Design. *Eurasian Journal of Educational Research 82:27-44*

[4]. Asiret, S., & Sunbul, S. O. (2016). Investigating Test Equating Methods in Small Samples through Various Factors. *Educational Sciences: Theory & Practice, 16,* 647-668.

[5]. De Ayala, R.J., Smith, B., & Dvorak, R. N. (2018). A Comparative Evaluation of Kernel Equating and Test Characteristic Curve Equating. *Applied Psychological Measurement, 2018, Vol. 42(2) 155–168*

[6]. Dorans, N. J., Moses, T. P., & Eignor, D. R. (2010). Principles and Practices of Test Score Equating. *ETS Research Report. ETS RR-10-29.*

[7]. Gunartha, I. W., Sulaiman, T., Suardiman, S. P., & Kartowagiran, B.. (2020). Developing Instruments For Measuring The Level Of Early Childhood Development. *REiD (Research and Evaluation in Education), 6(1), 2020, 1-9*

[8]. Harjo, B., Kartowagiran, B., & Mahmudi, A. (2019). Development of Critical Thinking Skill Instruments on Mathematical Learning High School. *International Journal of Instruction, October 2019, Vol.12, No.4*

[9]. Isnaini., Utami, W. B., Susongko, P., & Lestiani, H. T. (2019). Estimation Of College Students' Ability On Real Analysis Course Using Rasch Model. *REiD (Research and Evaluation in Education), 5(2), 2019, 95-102*

[10]. Kartowagiran, B., & Jaedun, A. (2016). Authentic Assessment Model for Assessing Learning Outcomes of Junior High School Students (SMP): Implementation of Authentic Assessment in Junior High Schools. *Journal of Educational Research and Evaluation, Vol. 20, No. 2, December 2016 (131-141).*

[11]. Kartowagiran, B., Mardapi, D., Purnama, D. N., & Kiswantoro. (2019). Parallel Tests Viewed From The Arrangement Of Item Numbers And Alternative Answers. *REiD (Research and Evaluation in Education), 5(2), 2019, 169-182*

[12]. Khasanah, M. M., Ramli, P., & Dwiastuti, S. (2020). Developing A Dynamic Assessment Instrument To Assess Reasoning Skills About Bacteria. *Journal of Educational Research and Evaluation, Volume 24, No 1, June 2020 (62-75)*

[13]. Kolen, M. J., & Brennan, R. L. (2014). *Test Equating: Scaling and Linking Methods and Practices.* (3rd Ed.). New York: Springer Science Business Media.

[14]. Luecht, R., & Ackerman, T. A. (2018). A Technical Note on IRT Simulation Studies: Dealing with Truth, Estimates, Observed Data, and Residuals. *Educational Measurement: Issues and Practice, vol. 37, no.3, 65–76*

[15]. Netto, W. L. (2018). Advances in Test Equating: Comparing IRT And Kernel Methods And A New Likelihood Approach To Equate Multiple Forms. *Doctoral Dissertation, Dipartimento di Scienze Statistiche Scuola Di Dottorato Di Ricerca in Scienze Statistiche Ciclo XXXI. Degli University Studies In Padova*.

[16]. Nisa, C., & Retnawati, H. (2018). Comparing The Methods Of Vertical Equating For The Math Learning Achievement Tests For Junior High School Students. *REiD (Research and Evaluation in Education), 4(2), 2018, 164-174*

[17]. Pramudita, K., Rosnawati, R., & Mam, S. (2019). Methods Used By Mathematics Teachers in Developing Parallel Multiple-Choice Test Items In School. *REiD (Research and Evaluation in Education), 5(1), 2019, 10-20*

[18]. Retnawati, H., & Wulandari, N. F. (2019). The Development of Students' Mathematical Literacy Proficiency. *Problems of Education in the 21 st Century, Vol. 77, No. 4*

[19]. Retnawati, H., Kartowagiran, B., Arlinwibowo, J., & Sulistyaningsih, E. (2017). Why are the Mathematics National Examination Items Difficult and What Is Teachers' Strategy to Overcome It? *International Journal of Instruction. Vol. 10 No.3.*

[20]. Retnawati, H. (2016). Proving Content Validity Of Self-Regulated Learning Scale (The Comparison Of Aiken Index And Expanded Gregory Index). *Research and Evaluation in Education, 2(2), 2016, 155-164.*

[21]. Retnawati, H. (2008). Estimation of Test Relative Efficiency Based on Item Response Theory and Classical Test Theory. Yogyakarta: *Dissertation on Educational Research and Evaluation Doctoral Program, Yogyakarta State University.*

[22]. Suhardi, I. (2020). Alternative Item Selection Strategies For Improving Test Security In Computerized Adaptive Testing Of The Algorithm. *REiD (Research and Evaluation in Education), 6(1), 2020, 32-40*

[23]. Von Davier, A. A., Holland, P. W., & Thayer, D. T. (2004). *the Kernel Method of Test Equating.* New York: Springer-Verlag New York, Inc.

[24]. Von Davier, A. A., & Chen, H. (2013). The Kernel Levine Equipercentile Observed-Score Equating Function. *ETS Research Report. ETS RR-13-38.*

[25]. Von Davier, A. A., Holland, P. W., Livingston, S. A., Casabianca, J., Grant, M. C., & Martin, K. (2006). An Evaluation of the Kernel Equating Method: A Special Study with Pseudo tests Constructed from Real Test Data. *ETS Research Report. RR-06-02.*

[26]. Von Davier, A. A., Fournier-Zajac, S., & Holland, P.W. (2007). An Equipercentile Version of the Levine Linear Observed-Score Equating Function Using the Methods of Kernel Equating. *ETS Research Report. RR-07-14.*

[27]. Wang, S., Zhang, M., & You, Sen. (2020). A Comparison of IRT Observed Score Kernel Equating and Several Equating Methods. *Taken from Front. Psychol. 11:308. doi:10.3389/fpsyg.200.00308 on January 25, 2020.*

[28]. Wiberg, M. (2016). Alternative Linear Item Response Theory Observed-Score Equating Methods. *Applied Psychological Measurement 2016, Vol. 40(3) 180–199*

[29]. Wiberg, M., & Gonzales, J. (2016). Statistical Assessment of Estimated Transformations in Observed-Score Equating. *Journal of Educational Measurement, Vol. 53, No. 1.106–125.*

[30]. Wiberg, M., & Branberg, K. (2015). Kernel Equating Under the Non-Equivalent Groups With Covariates Design. *Applied Psychological Measurement, 2015, Vol. 39(5) 349–361*

[31]. Wolkowitz, A. A., & Wright, K. D. (2019). Effectiveness of Equating at the Passing Score for Exams With Small Sample Sizes. *Journal of Educational Measurement, Vol. 56, No. 2,361–390*