Research Article doi: 10.12973/eu-jer.8.4.1073

European Journal of Educational Research
Volume 8, Issue 4, 1071 - 1079.
ISSN: 2165-8714
http://www.eu-jer.com/

A Comparison of Score Equating Conducted Using Haebara and Stocking Lord Method for Polytomous

Risky Setiawan*
Universitas Negeri Yogyakarta,
INDONESIA

Received: July 20, 2019 ▪ Revised: August 31, 2019 ▪ Accepted: October 3, 2019

Abstract: The purposes of this research are: 1) to compare two equalizing tests conducted with Hebara and Stocking Lord method; 2) to describe the characteristics of each equalizing test method using windows' IRTEQ program. This research employs a participatory approach as the data are collected through questionnaires based on the National Examination Administration of 2018. The samples are classified into group A and group B respectively by 449 and 502 respondents. This paper discusses how to equalize shared items using the anchor method with a set of instruments in the forms of 35 questionnaire items and 6 shared items. In addition, the researcher also uses PARSCALE to estimate each respondent's skills and each item's characteristics. The shared items are eventually equalized using IRTEQ program. The results show that there is a significant difference between those conducted using Haebara method (0.592) which produces bigger mean-sigma value and Stocking & Lord (0.00213). Thus, the results show that the shared testing items may improve respondents' discrimination and increase the difficulty level (parameter b). Due to the availability of shared items, it is good and appropriate to equalize two different tests on different theta skills.

Keywords: *Equating, polytomous, graded data.*

To cite this article: Setiawan, R. (2019). A comparison of score equating conducted using Haebara and Stocking Lord method for polytomous. *European Journal of Educational Research*, *8*(4), 1071-1079. http://doi.org/10.12973/eu-jer.8.4.1071

Introduction

Scoring is one of the most important components in education system. Scoring results may reflect the development or progress of educational outputs when compared from time to time, school to school, or district to district. Equalizing educational achievement processes among schools or districts in the measurement theory is called equating.

Based on its coverage, scoring is divided into macro and micro. Macro scoring tends to use samples in analyzing the program and its impacts; the program is called curriculum. Education program is a planned program to improve the quality of education. Meanwhile, micro scoring which is commonly used in a class is intended to figure out the learning outcomes, particularly students' achievements. The target is a learning program in a classroom assisted by the teacher.

Thus, this paper discusses how to equalize the shared items using anchor method, a set of testing instruments in the forms of questionnaire with 35 items and shared items. By using PARSCALE, the researcher may estimate each respondent's skills and item's characteristics that the shared items are eventually equalized using IRTEQ program.

Gronlund (1990, p. 180) suggests several elements to construct a test, as follows: 1) item stem should be meaningful to the item itself and show certain problem; 2) test item stem involves many possible answers and free from irrelevant materials; 3) negative item stem statement is only used when the expected learning results are quite significant; 4) All answer alternatives are grammatically consistent to those item stems; 5) an item clearly contains only one best right answer; 6) Test items are used to measure understanding which contains several new things, yet should be carefully selected; 7) All distractors should be logical or reasonable; 8) Verbal association between stem and the right answer should be avoided; 9) the length of answer choices should not relatively show the right answer; 10) The right answer should appear on each position of choices with several approach similarities, yet in randomly orders; 11) efficiently use particular choices as if there is no the right answer or all the answers above are right; and 12) Do not use multiple choice items if the other items are more appropriately used.

* **Correspondence:**
Risky Setiawan, Universitas Negeri Yogyakarta, Graduate School, Colombo Street No.1 Karangmalang Depok Sleman, Yogyakarta, Indonesia
✉ riskysetiawan@uny.ac.id

A test should measure learning results with the same scale while the possible approach may be conducted as follows: 1) use anchoring items (common items) for several testing sets; 2) use the calibrated items (items which characteristics at one common scale are acknowledged); and 3) combine both approaches by selecting the anchoring items from the calibrated items (Kumaidi, 2000, p. 105). In this case, the Item Response Theory (IRT), role is quite essential to equalize the scale. After the qualified items are selected based on professional adjustment of experts in certain field of study and measurement specialists as well as supported with empirical data of the trial resulted items, the following activity is making a scale and determining where each item should be located in that scale (Setiadi, 1998, p. 10). Naga (1992, p. 394) states that from time to time item banks keep experiencing continuous development with new item inputs and old item omissions. This research uses a modern scoring theory. To create good item instruments, many elements are highly required, especially from the essential aspects which require deeper and more fundamental studies either from classic or modern measurement points of view that the test utilization may result in higher function of testing item or examination information. Thus, there is no one perfect test as long as those various requirements explained above are not completely fulfilled.

Naga (1992, p. 2) states that testing scores based on each item's checking process results is conducted as the test takers show the correct an incorrect answers which are classified into two: 1) single score, as the answer is from one test taker, and 2) composite score, as the answer is combination of single score. Barnard (2011), also mentions that there is no definition of score equating which may be universally accepted. Peterson, Kolen (1989, p. 221) define equating as a process used to ensure the scoring results from the testing administration to use in turns. Meanwhile, Crocker and Algina (1986, p. 457) assert that equation may be defined as a process to set equivalent scores with two instruments. Score equating is an empirical procedure required to transform scores of one testing instrument to the others that the score equating should be conducted based on the testing scores.

Creating an equal test for two or more question packages is not easy or probably impossible as there must be differences. It is almost impossible to organize a real parallel multi-package test (Petersen, Kolen & Hoover, 1989). Although the test is made using similar testing specifications in writing each item and by changing the numbers, there is no guarantee that the difficulty level of each item will be the same. Moreover, the answer keys or choices are different. Angoff (1971); Kolen (1995) explain that the equating methods are divided into 2 categories: 1) equipercentile equating and 2) linear equating.

Implementation of item responsive theory in testing equating activity should meet two basic assumptions which consist of unidimension and local independence (Brennan & Kollen, 2004). Meanwhile, some procedures to conduct a testing equating activity are based on item responsive theory as follows: 1) conducting item parameter estimation and ability parameter; 2) Estimating scale of item responsive theory using linear transformation; and 3) Equalizing the scores. If using scores of the right answer, the conversion is conducted to the right answer scale and then continued to the scoring scale. As testing equating activity has an empirical procedure and this activity then requires a certain design to be well considered.

There are three kinds of testing equating to use, that is, single group design, equivalent group design, and anchor testing item design. In single group design, one group of participants is used to provide responses on two testing instruments (X and Y). Item parameter of both testing instruments is separately estimated by calibrating test takers' ability or item parameter. Based on the design which calibrates the test takers' ability parameter, the item parameter of the testing instrument X and Y is at the same scale.

Ideally, to equalize scores of several testing instruments, those should be given to the same respondents. By comparing the test takers' ability from those two or more testing instruments, equating of those two testing instruments may be conducted. Facts in the field, this design is not easy to conduct as there are some exhaustion, learning, and exercise factors for the second or further tests. In addition, there will be a difficulty in designing adequate time for respondents to attend the test for more than once (Miyatun & Mardapi, 2000).

Lord (1990) suggests that the equating concepts or ideas are implied as follows: 1) Testing measurement with different characteristics may not be equalized; 2) as the raw scores are on consistently different test, the equating processes may not be conducted; 3) Raw scores on test with various difficulty levels may not be equalized as the test is not consistently the same with that of difficulty levels; 4) Scoring errors on test A and test B may not be equalized, unless both tests are completely parallel; and 5) equating may be applied in test with a complete reliability.

Equating is conducted by converting one package to the others, from which measures the same abilities. Testing instrument equating is score decision making obtained from a package adjusted to different forms of difficulty levels. If package X is more difficult than package Y, then the equating of package X to package Y results in higher or more valuable package X if equalized into package Y (Crocker & Algina, 1986). There are three data designing bases to take or analyze in conducting a testing equating (Brennan & Kolen, 2004), including: 1) Data design collected from two groups tested with different packages with the same content outlines, in which both packages are randomly distributed; 2) For equating process, one of the tested groups is given package A, then package B, and package A once again; and 3) Different testing instrument is also given to different test takers. However, in both packages, there is an anchor test

given to all test takers. Anchor test is one testing criterion to conduct equating. The test takers may not be randomly distributed although random distribution may also not influence this model.

This research is fundamental because in order to standardize items on a broad scale of a country such as Indonesia; a proper analysis of equality needed. Equivalence provides information that equating package questions between provinces can increase so that biases and disparities between regions can be reduced. As research conducted by Rahmawati (2015), which analyzes the equivalent with the results of the 0.5 point score raw TCC difference criteria leads to 100% consistency in the graduation classification.

Methodology

Research Goal

This research employs a participatory approach as the data are collected through questionnaires based on the National Examination Administration of 2018. This paper discusses how to equalize shared items using anchor method with a set of instruments in the forms of 35 questionnaire items and 6 shared items. In addition, the researcher also uses PARSCALE to estimate each respondent's skills and each item's characteristics. The shared items are eventually equalized using IRTEQ program.

Sample and Data Collection

Sampling with stratified random sampling of 12 high schools in Yogyakarta province. Each school is represented by one class so there are 24 groups of students. The samples are classified into group A and group B respectively by 449 and 502 respondents. The instrument used has been validated in content by experts in measuring and evaluating educators. The construct validation was analyzed and the loading factor value was more than 0.5. While the reliability of the resulting Cronbach's Alpha is 0.83. So the instrument is feasible to use for the equating process.

Analyzing of Data

The first category is a completed scoring conducted using the comparison between X and Y testing score which may be equivalent as the percentage ranking orders of each group is the same. Furthermore, to equalize the scoring of 2 different tests, the same examining test should be given to both. Meanwhile, the second category, it is assumed that score x on test X and score y on test Y have a linear relationship. Tumilisar (2006) states that equating method is ways to figure out the equating relationship of 2 testing scores of two different research instruments using particular statistical methods, while the data are collected using particular data collection designs. The equipercentile equating Method is divided into two, as follows:

Table 1. Frequencies and Percentages of Shared-items alternative answer choices

Item	**Categories**			
	1	**2**	**3**	**4**
0003				
frequent	24	125	324	29
percentage	4,8	24,9	64,5	5,8
0010				
frequent	147	84	213	58
percentage	29,3	16,7	42,4	11,6
0013				
frequent	81	70	196	155
percentage	16,1	13,9	39	30,9
0023				
frequent	54	76	308	64
percentage	10,8	15,1	61,4	12,7

Based on above table, it shows that the total frequencies and percentages of the graded data on 2 different tests have similar shared items. The first category is a complete assessment carried out using a comparison between test scores X and Y which may be equivalent to the ranking order of percentages of each group is the same. Furthermore, to equalize the scores of 2 different tests, the same examination test must be given to both. Meanwhile, in the second category, it is assumed that the x score on the X test and the y score on the Y test have a linear relationship. The equalization method in this research is to find out the equation of 2 test scores from two different research instruments using certain statistical methods, while the data are collected using a specific data collection design.

The equalization method is carried out with the following steps: 1) Using the equicate equating equipercentile method is a way to find out the equivalence of two test scores with two different research instruments. Data is collected using

unequal anchor test designs, while the anchor test is an internal anchor test that uses certain statistical methods. Equipercentile equivalents are calculated using the direct equipercentile method which is divided into two assessment instruments, each based on an anchor test with real populations. The equalization process of various testing instruments can be done in two ways: horizontal and vertical equalization. Horizontal equalization is to equalize the process of two different testing instruments but the measurements are the same. Meanwhile, vertical equalization is the process of equalizing the two groups of testing at different levels of education, but on the same testing instrument.

The equalization aims to equalize the scores by comparing the scores obtained from one testing instrument with the other through the assessment of the tanning process; mean is a step in the standard equation test; it the two test instruments scores obtained from tests A and tests B can be compared when fulfilling four requirements: 1) Measuring abilities or characteristics that are similar. Thus, testing composed of a variety of different content cannot be equated; 2) After being equalized, the frequency distribution of scores obtained from test A must be the same as those from test B, that the scores obtained from test A and test B can be exchanged after being compared; 3) Testing the equalization must be free from the data or work of the test takers, as well as conversions originating from the equation that applies to all similar situations; and 4) Transformations must be the same without considering which tests can be used as a basis or conversion reference. This means that the interpretation of scores must have the same equation from test A to test B or from test B to test A.

Findings / Results

On equivalent group design, the participants of two equivalent groups (K1 and K2) and two testing instruments (X and Y) are employed. Participants of group K1 do the testing instrument X and participants of group K2 do the testing instrument Y. As group K1 and K2 are equivalent, both groups are then considered single. Constanta determination of further conversion is due to the single group design. The advantage of this design may avoid the negative influence caused by test takers' exercise and exhaustion, while the disadvantage is that there is a bias possibility as it is not easy to distribute test takers' abilities from those completely equivalent groups (Sukirno, 2007, p. 310).

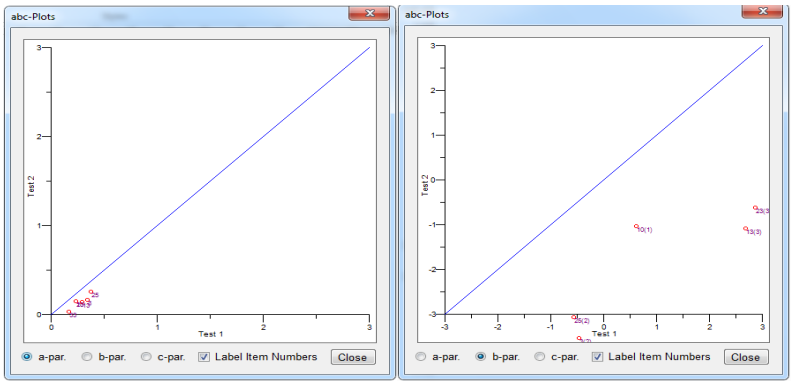


Figure 1. abc - plot Graphics on IRTEQ

On equivalent group design, the participants of two equivalent groups (K1 and K2) and two testing instruments (X and Y) are employed. Participants of group K1 do the testing instrument X and participants of group K2 do the testing instrument Y. As group K1 and K2 are equivalent, both groups are then considered single. Constanta determination of further conversion is due to the single group design. The advantage of this design may avoid the negative influence caused by test takers' exercise and exhaustion, while the disadvantage is that there is a bias possibility as it is not easy to distribute test takers' abilities from those completely equivalent groups (Sukirno, 2007, p. 310).

Anchor testing design is generally used when there is a testing security problem which becomes one important consideration to implement several tests at the same time. At this design, each testing instrument has several common items and each group does different testing instruments. There are two variations in this design, such as (Chong and Osborn, 2005) if common item is calculated when giving the score, it is called as internal common item; and 2) If common item is not calculated when giving the score, it is called as external common item.

Petersen et.al (1989), state that anchor test consists of several items which are considered as the miniature of both equalized tests (having most equating similarity, both content and material depth with both equalized tests). Livingston, Dorans, and Wright (1990, p. 75) state that a method employing anchor testing score is used to adjust different abilities between the new and old testing samples. Based on practice experiences (rule of thumb), Kolen and Brennan (1995, p. 248) state that the same item number should be at least 20% of the whole testing items (40 items or more) and in this case, as the test is quite long, 30 items are then considered adequate.

Anchor testing design uses two testing instruments (X and Y) and the participants of two groups (K1 and K2). Each testing instrument is added with anchor testing items Z that both testing instruments become (X+Z) and (Y+Z). The test

takers of group K1 do the testing instruments (Y+Z) that anchor test items Z are completed by both groups of test takers. The equating scale similaritation is conducted by caliberating the ability parameter or anchor testing item parameter (Camilli & Shepard, 1994). If the anchor testing design is caliberated with item parameter, then the test takers' ability parameter on both groups is at the same scale.

During the estimation, item parameter for anchor testing items is identically maintained for both groups, yet the evaluated item parameter is different. The scale for those two evaluated item parameter sets are the same as they are closely related through their similarity on anchor testing item parameter (Susongko, 2005). Thus, the equating is thereby formally completed by manipulating the model of item responsive theory than the particularly separated procedure. When equating is required by one testing item to be evaluated, the other items are considered as anchor testing items. It means that each item is individualized stated as the evaluated item on the separated calibration analysis.

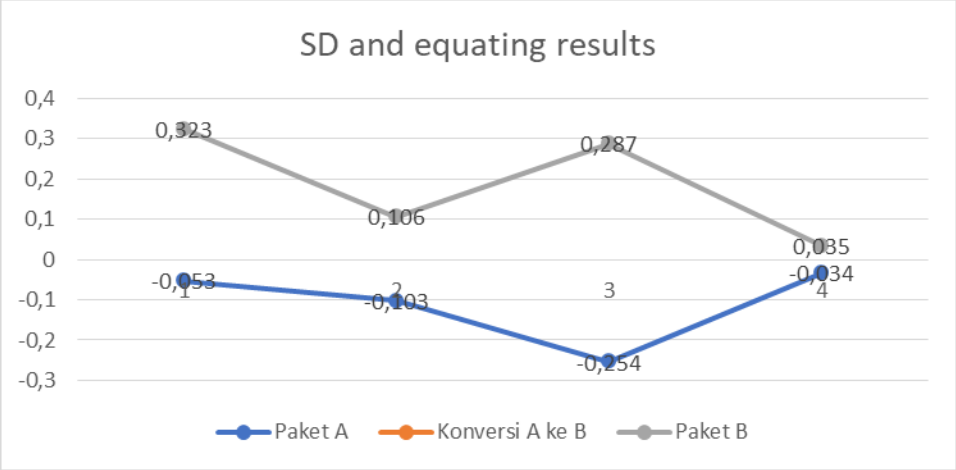
It has been explained before, that the vertical equalization with the classical approach uses linear and equilateral methods. Vertical equalization with the linear method gets the conversion equation: (1) group A is lY (x) = 1,051 X + 0.082; (2) group B is lY (x) = 1,172 X + 0.014. The presentation data means that the two equalization linear equations have fulfilled the equating requirements so that the next process can be carried out. The mean score and standard deviation of the equalization results are presented in a graph, as shown in Table 2 — the amount of SEE equalization with the linear method obtained for 0.725.

Tabel 2. Rerata Location dan Simpangan Baku pada Dua Sub test

Equating	Mean- Location (b)	Deviations
Location A	-0.053	0,323
Location B	-0.254	0,287

The result of equalizing the location with the average sigma method shows that the two sub-tests made have met the rules and standards with the amount of RMSD of 0.28. The following is a graph of the equalization of theta ability in the two sub-tests.

Figure 2. Theta Average of Equalization Results



In the equivalent of the linear method (Figure 2) it appears that the mean score of students will rise if converted to the grade level above. It also means that students' abilities have increased along with class grades. However, if we look at the conversion equation from package A to package B, it can be seen that students who have low scores and high scores have different conversion directions. For example, the conversion of grade III to class VI gives information that students who get under 20 in package A will be lower if the score is converted to package B, while students who score 20 and above will get a higher score. The same problem can be investigated in other conversion equations. This study shows the inconsistency of the conversion results using the linear method.

Equivalence of scores by the equi-precentil method provides more consistent information than the linear method. This can be seen from the average score of the conversion results, which all have the same tendency after the equalization. The average score of package A is higher than the average score of package B. This can be interpreted that the ability of students has increased in accordance with the increase in grade ranking. This is certainly very reasonable because student knowledge should increase according to the learning experience.

Discussion and Conclusion

Equating method based on item responsive theory has the function to determine conversion Constanta. It means that equating between two or more testing instruments may be conducted when conversion Constanta is already known. The resulted conversion value is then substituted in scale equation on the equating design used. Several testing equating methods may be used as well as factors influencing the accuracy of the testing equating method. Item responsive theory has four testing equating methods: regression, sigma mean, mean and robust sigma, as well as characteristic curve (Angoff, 1982).

The first testing equating method is called a regression method. The conversion Constanta a and b are determined with a regression method by considering the test takers' responses on testing instrument X and Y. The participants' item parameter estimation and ability parameter should meet the linear regression equation, as follows:

$y = ax + b + e$ *with* $a = rxy\ Sy/Sx$ *and* $b = \hat{y} - ax$

Description:

y : ability estimation or item parameter estimation on testing instrument Y

x : ability estimation or item parameter estimation on testing instrument X

rxy : Correlational Coeficient between X and Y

y, x : mean of y and x

Sy, Sx : standard deviation of x and y

e : error in regression instrument estimation

The use of this method is considered asymmetric that it is inadequate to determine the conversion Constanta. Moreover, the equating of two or more testing instruments highly requires invariance and asymmetric requirements from the equalized testing instruments.

The second testing equating method is sigma mean method. In this method, the determination of conversion Constanta α and β based on mean and sigma method is conducted by considering the estimation value of testing item difficulty level parameter on testing instrument bx and by. Hambleton & Swaminathan (1985: 26) state that the relationship between testing item parameter estimation or test takers' ability parameter on the second testing instruments may be equalized, while the determination of its conversion Constanta should meet the following equation:

$y = ax + b$ *with* $a = Sy/Sx$ *and* $b = \hat{y} - ax$

Mean and sigma method are considered asymmetric that the relationship of y to x may be determined using the same method. However, Hambleton & Swaminathan (1991), suggest that those mean and sigma equating methods are not considering the error standard variations of item parameter estimation. The third testing equating method is called mean and robust sigma method. Hambleton and Swaminathan (1991) state that the mean and sigma equating method is not considering item parameter estimation variations but the existence of item parameter estimation error standard variations. Steps in determining conversion Constanta of testing equating use this method, as follows (Sukirno, 2007: 312):

Item parameter weight (wi) on each couple (bxi and byi) is determined with: wi = [max{v(xi), v(yi)}]-1, where: i = 1,2,3,4....k, v(xi) and v (yi) are parameter estimation variants of difficulty testing level X and Y.

wi scaling weight is determined using the following formula: wi`= k = the number of anchor items on testing instrument X and Y. Calculation of the weighted testing estimation X and Y employ the following formula: *xi`=wi`xi and yi`=wi`xi*. Mean and standard deviation of the weighted testing estimation X and Y is determined by *x, y, Sx`, Sy`*. Conversion Constanta α and β is determined using mean and standard deviation of the weighted estimation which is conducted by substituting the mean and standard deviation of the weighted estimation at equating scale equation.

Stocking and Lord (Hambleton, 1985) state that in mean and sigma equating method, the process of conversion Constanta is determined without considering the possibility of extreme group scores, while the mean and robust sigma equating method may be improved by considering the extreme group scores.

The fourth method used in testing equating is the characteristic curve method. The conversion Constanta α and β with the characteristic curve method is determined by considering the value of item parameter testing estimation of item instrument x and y. The mean and sigma equating as well as the mean and robust sigma method in determining conversion Constanta only consider the existing relationship among item difficulty parameters on one testing instrument to the others. The relationship among the differential power parameters on both testing instruments has not yet been considered.

Rahayu (2008) states that the characteristic curve method considers information obtained from item differential power parameter and item difficulty level in determining conversion Constanta. Thus, in the characteristic curve equating

method, the relationship among differential power difficulty parameters, and the testing item difficulty parameters which may be equalized should be considered. In addition, the test takers' true scores in the characteristic curve method on both testing instruments should be considered.

The test takers' True Score (txa) with θa ability responding to k item at instrument X and Y is as follows:

xa = (θa, bxi, , cxi) and *ya =(θa, byi, ayi, cyi)*

Each item at testing intrument χ and ϒ should meet the following eqution:

byi = αbxi + β *αyi = atau α =* *cyi = cxi* *β = byi – αbxi*

Constanta α and β are selected in such a way that the function F as mentioned below may reach its minimum score.

F = (xa – yi)

Description:

F : the function derived from α and β, showing dissimilarities between xa and ya

N : number of test takers

xa : test takers' true score at ability a of testing intrument X

ya : test takers' true score at ability a of testing intrument Y

Similar with study by Chong and Osborn (2005) suggest that there are four equivalent aspects that should be considered in testing equating: 1. Interference, how far the scores of both tests may be used to measure the same purposes, such as measuring the accounting achievements and the counting abilities. 2. Construct, how far both test packages may measure the same constructs. 3. Population, how far the population used is homogeneous or the same. In addition, quality and quantity factors related to learning system that should be equalized. It means that a school which students with low economic and social background, poor school infrastructure and facility, and common teachers may not be compared with the unequal school conditions. 4. characteristics or measuring conditions, how far the measuring condition similarity may be conducted for both testing packages, either from the length of testing items, testing forms, testing administration, testing period of time, item types, and testing procedures. This is the same as Antara's (2015) study which shows that the mean & sigma method shows that the mean ability of students experiences an increase as class grades increase.

Item Equating Analytical Results

Polytomous item is one equalized test, that is, Graded Partial Credit Model, which consists of a graded data system in one permanent bound with a total number of 35 questionnaire items and Test takers for A = 502 and test takers for B = 449. The data are then analyzed using PARSCALE program to figure out each testing instrument estimation and equalized using IRTEQ program. The analysis is conducted by comparing two sigma mean methods: *Haebara* Method as well as *Stocking and Lord* Method.

The results show that the mutual testing items may improve test takers' differential power and add the difficulty level (parameter b) that mutual testing items may be appropriate to equalize two different test with different theta abilities as well.

Table 3. Item Parameter Estimation's Descriptive Statistics of Pre and Post Equlization Tes A and B

Parameter Item	**Measurement**	**Tes A**	**Tes B**	**Equating result**
Slope (a)	Mean	0.358	0.391	0.866
	SD	3.568	4.259	3.931
	N	35	35	
Threshold (b)	Mean	-1.076	-0.546	0.763
	SD	818,98	866.59	6.314
	N	35	35	
RMSD Chi-Square	0.00213			

Sigma mean of each testing instrument and TTC analytical results of each method are then shown as follows:

Table 4. Item Parameter Estimation's Equating Accuration using TCC Method on Test A and B

TCC		**Haebara**	**Stocking & Lord**
A			
Mean-Mean	0.55	0.68	0.61
Mean-Sigma	0.62	2.4	2.49
Value		0.592	0.00213
B			
Mean-Mean	0.62	0.68	0.61
Mean-Sigma	3.2	2.4	2.49
Value		0.592	0.00213

From the TCC recapitulation of two methods above, it can be concluded that the use of Linking testing method and Haebara method result in greater Value Mean-Sigma than that when using the *Stocking & Lord* method. Thus, Equating and Linking is more effective and optimum when using *Haebara* method. TCC (*Test Characteristic Curve*) Graphic between Linking items and all items is shown below:

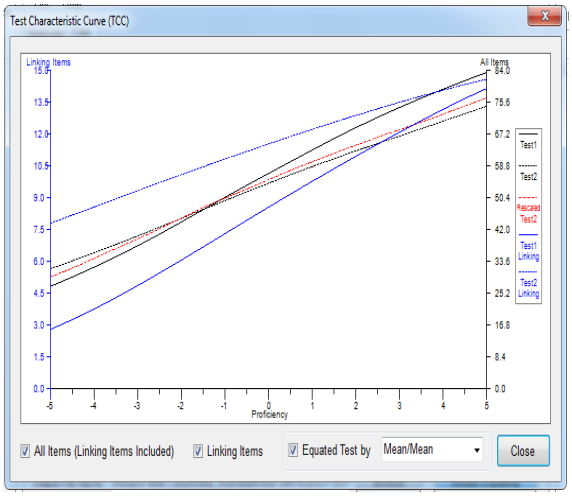


Figure 3. Test Characteristic Curve Graphic

Conclusions

Item responsive theory is an alternative choice aiming to be free from a testing dependency which is given with samples of test takers. In this case, although the items are completed by the high or low competence students, the indication of each item difficulty level has no changes. Moreover, the implementation of item responsive theory on testing equating should be fulfilled with unidimension assumption and local independence. There are some ways made to conduct the testing equating activity based on the item responsive theory as follows: 1) Estimating parameter, 2) Estimating scale of item responsive theory using a linear transformation, and 3) equalizing score. Meanwhile, there are three designs of testing equating used based on item responsive theory as follows: 1) Single group design, 2) Equivalent group design, and 3) Anchor testing design. Methods which are recently developed to conduct testing equating based on item responsive theory are as follows: 1) Regression method, 2) mean and sigma methods, 3) Robust mean and sigma methods, and 4) Characteristic curve method. The average score of package A is higher than the average score of package B. This can be interpreted that the ability of students has increased in accordance with the improvement in class rank. Thus, students' knowledge must increase according to the learning experience.

Suggestions

Referring to the results of the research that Referring to the results of research that the equalization carried out to provide specific information can be taken with several options. Each choice has different advantages that are used to develop good test kits. The suggestion for researchers in the field of measurement is that the selection of an appropriate equalization method can improve the representation of the characteristics of test participants to optimize student learning outcomes.

Acknowledgements

Great appreciation and gratitude are due to the Education Assessment Center (PUSPENDIK), the Ministry of Education and Culture of Indonesia, and the school where the data was collected, which provided research facilities and permits.

References

Angoff, W. H. (1971). *Scales, norms and equivalent scores. Educational measurements*. Washington, DC: American Council on Education.

Antara, A., & Bastari, B. (2015). Vertical equalization with classical and item response theories in elementary school students. Journal of Educational Research and Evaluation*, 19*(1), 13-24. https://doi.org/10.21831/pep.v19i1.4551

Brennan*,* R. L., & Kolen, M. J. (2004), *Test equating, scaling, and linking*. Iowa City, IO: American Council on Education and Springer Publisher.

Camilli, G., & Shepard, L. A. (1994). *Methods for identifying biased test items*. Thousand Oaks, CA: Sage Publications.

Dafflon, B. (2011). Solidarity and the design of equalization: Setting out the issues. *eJournal of Tax Research, 10*(1), 1-26.

Yu, C. H., & Osborn Popp, S. E. (2005). Test equating by common items and common subjects: Concepts and applications. *Practical Assessment Research & Evaluation*, *10*(4), 1-19.

Crocker, L. M., & Algina, J. (1986). *Introduction to classical and modern test theory*. New York, NY: Holt, Rinehart, and Winston.

Gronlund, N. E., & Linn, R. .L (1990). *Measurement and evaluation in teaching* (6th ed). New York, NY: Macmillan/ London, UK: Collier Macmillan.

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

Hambleton, R.K., & Swaminathan, H. (1985). *Item response theory principles and applications.* Boston, MA: Kluwer.

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

Kolen, M. J., & Brenan, R. L. (2004). *Test equating, scaling, and linking.* Iowa, IO: Springer.

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

Kumaidi. (2000). Standardization of problem items. *Journal of Education and Culture, 5*, 132-143.

Livingstone, S. A., Doran, N. J., & Wright, N. K. (1990). What Combination of Sampling and Equating Methods Work Best? *Applied Measurement in Education*, *3*, 73-95.

Lord, F. M. (1990). *Aplications of item response theory to practical testing problems*. Mahwah, NJ: Lawrence Erlbaum Associates.

Miyatun, E., & Mardapi D. (2000). Comparison of test equalization methods according to item response theory*. Jurnal Penelitian dan Evaluasi, II*(3), 124-132.

Naga, D. S. (1992). *Introduction to scoring theory on educational measurement*. Jakarta, Indonesia: Besbats.

Peterson, N. S., Kolen, M. J., & Hoover, H. D. (1989). Scaling, norming, and equating. In R.L. Linn (Ed), *Educational measurement*. New York, NY: Macmillan.

Rahayu, W. (2008). The effect of the linking method on many false positive items on dif detection based on item responsiveness theory. Dissertation. Jakarta, Indonesia: Universitas Negeri Jakarta.

Rahmawati, R., & Mardapi, D. (2015). Modified Robust Z method for equating and detecting item parameter drift. *REiD (Research and Evaluation in Education), 1*(1), 100-113. https://doi.org/10.21831/reid.v1i1.4901

Setiadi, H. (1998). Question bank calibrated with the IRT concept solve problems of systematic exams held in specified periods. *Jurnal Kajian Dikbud, IV*, 13.

Sukirno, D. S. (2007). National test equalization: Why and how? *Jurnal Cakrawala Pendidikan, XXVI*(3), 305-321.

Susongko, P. (2005, May). *Matching item parameters concurrently to test statistically the existence of item function (DIF).* Paper presented at the National Seminar on Research Results on Evaluation of Learning Outcomes and Management, Yogyakarta, Indonesia.

Tumilisar, A. V. J. (2006). Relative accuracy of equalization tests for 300-size samples judging from the equalization Method and Refining Technique. *Jurnal Pendidikan Penabur*, *6*, 1-19.

Weimo, Z. (1998). Test equating: What, why, how? *Research Quarterly for Exercise and Sport, 69*(1), 11-23. http://doi.org/10.1080/02701367.1998.10607662