*Advances in Social Science, Education and Humanities Research, volume 529*
**Proceedings of the International Conference on Engineering, Technology and Social Science (ICONETOS 2020)**

# Equating Test Instruments Using Anchor to Map Student Abilities Through the R Program Analysis

Melly Elvira[1,*], Syamsir Sainuddin[2]

[1]*Universitas Islam Negeri Maulana Malik Ibrahim Malang, Indonesia*
[2]*Universitas Cokroaminoto Palopo, Indonesia*
[*]*Corresponding author. Email:melly@uin-malang.ac.id*

**ABSTRACT**
This study aims at investigating the equivalence of mathematics try out scores of junior high school through anchor. In the study, the 2 Parameters Logistic (2PL) model of Item Response Theory (IRT) analysis was used through the Haebara method. The estimation step of the equating parameter utilized the equateIRT package in the R program. The results show that the relationship between try out questions of package 1 and package 2 cannot be said to be equivalent. After analyzing and calculating the regression equation, it can be concluded that package 2 of Mathematics National Examination (MNE) try out questions has a higher level of difficulty than package 1. Equating package 1 to package 2* through the Haebara method for the 2 PL model will provide equal or fair scores for the test takers who work on package 1 and package 2. The samples of questions were taken from MNE questions distributed in Makassar in 2014 which consisted of 2 question packages. Question package 1 was responded to by 2099 students, while question package 2 was responded to by 2068 students. The equating design used was a common item design. With the equating made, in the future, the test takers' ability can be detected properly if they work on similar tests, even though test packages are relatively different.

***Keywords:*** *Test Equating, Haebara Method, R Program, Student Ability.*

## 1. INTRODUCTION

The results of the try out exam are used as one of the considerations for mapping the quality of the education unit or the basis for determining the graduation of students from the educational unit. The try out scores can be a benchmark in measuring student ability if they use the same items. However, in practice, try outs are administered using more than one package. It will most likely create a difference in the perceived value of ability among participants. The obtained score difference cannot be directly concluded as to participants' different abilities. As it is known, other considerations should be observed first, for example, whether the questions that have been made have the same level of difficulty to measure the ability or it has not. Therefore, it is necessary to equate the scores of students' answers to measure whether their abilities are equal or not through the equating method.

The equating method is a scientific method that can equate the raw score for package 1 to another through value conversion. This method is appropriate to use where there is never a question from two test sets with different items even though they are based on the same question grid and have the same level of difficulty [1]. This means that each student's scores obtained from the different packages have not shown their true ability. Thus, it is necessary to continue with the process of equating their answer scores to distinguish high-ability and low-ability students. Hence, neither party feels disadvantaged.

Sets of tests can be equated when they have several common items for direct equivalence. If the two test sets use different patterns, the conversion result contains the average coefficient of equivalence. The R analysis used on this occasion calculates the direct or indirect coefficient of equating. The package used will generate

 

an estimate of the equating coefficient and standard error of the direct equivalent.

### *1.1. Type of Equating*

There are two types of equating, namely horizontal equating and vertical equating.

#### *1.1.1. Horizontal Equating*

Horizontal equating is the equating of test sets that have a comparable level of difficulty where the test is given to a group of participants who have an equal distribution of abilities [1]. Horizontal equating is an equalization conducted on two or more test sets which have a comparable difficulty level where the test is carried out by test takers with the same ability distribution.

#### *1.1.2. Vertical Equating*

At certain times testing is also needed for students in different classes. This is intended to compare the abilities of students in the upper and lower classes [2]. The vertical equating is also a procedure to control the quality of learning in an area. Two or more equated test scores are tests that measure different levels/grades, of which some are higher or lower than others.

### *1.2. Basic Principles of Equating*

Lord in[1], explains that there are four basic principles in equating the test, namely:

a. The principle of equity, which means each group of test takers has the same ability.

b. The principle of invariance, which means that the equating is by mapping the same score regardless of the group of test-takers.

c. The principle of symmetry, in which equating can be carried out back and forth regardless of which one is labelled X and which one is labelled Y.

d. Unidimensional.

### *1.3. Equating Design*

In equalizing the test sets, one of the designs that can be used for data collection is to connect test scores [1], which is the common item design. This design is used when the two tests to be equated are given to two different groups. Each test contains a set of common items which are test items that are used commonly in the two presented packages. A set of common items is a part of the test items that are called a common item or anchor item.

**Table 1.** Common Item Design (Anchor)

| Population | Sample | Item X | Anchor | Item Y |
|---|---|---|---|---|
| P | 1 | √ | √ | |
| Q | 2 | | √ | √ |

The number of anchors is usually adjusted according to the length of the test. Having 20% of anchor items results in less error of equivalent measurement than using 10% of anchor items. The anchor plays an important role in equating process [3]. It is good practice to build a test anchor by adjusting it to the test specifications, thus anchor is a mini version of the two equalized tests. That means, they must have the same difficulty level and contain the same content.

### *1.4. Equating of Tests Based on Item Response Theory Approach*

Various methods that can be applied to correlate scores between two or more tests. When we viewed from the calibration technique, the test equating method is divided into two methods, namely the separate calibration method and the simultaneous calibration method [4],. In the separate calibration method, the two tests are calibrated independently while in the simultaneous calibration method both tests are calibrated simultaneously. In the simultaneous calibration, there is no calculation of the equating constant and the results of the calibration of the two tests automatically show that the item parameters and capabilities are on the same scale.

Relationship between item parameters and ability if the scale on test 1 is equalized to the scale of test 2 where the test sets are done by two different groups as follows [5].

$$a_{2j} = \frac{a_{1j}}{\alpha} \quad (1)$$

$$b_{2j} = \alpha b_{1j} + \beta \quad (2)$$

Where,

$b_{1j}, b_{2j}$ : difficulty index of test 1 and test 2

$a_{1j}, a_{2j}$ : discriminant index of test 1 and test 2

$\alpha$ and $\beta$ : constant value in equating

One method that can be used in equating this test is the Haebara method. The Haebara method is a characteristic curve method equating the item parameters based on item characteristic functions. The sum squares of the difference between the value of the function for the same abscissa on each item of the characteristic curve of the two equated scales is expressed as $H(\theta_i)$:

$$H(\theta_i) = \sum_{j=1}^{n} \left(T_{ij} - T_{ij}^{*}\right)^2 \quad (3)$$

$$T_{ij} = P_j(\theta_i) \quad (4)$$

$$T_{ij}^{*} = P_{ij}^{*}(\theta_i) \quad (5)$$

Where

n : number of items

$P_j(\theta_i)$ : probability of correctly answered item j by participants with $\theta_i$ ability

### *1.5. Equating Design*

With the application of more than one method in the equating process, it is necessary to know how the accuracy of each equating method is. The accuracy of the equating results can be seen by comparing the average Root-Mean-Square Different (RMSD) values of the item characteristics before and after being equated. Kilmen & Demirtasli (2012) made an equivalent using four methods in the IRT approach. The accuracy of the four methods was calculated by looking at the smallest RMSD value. To calculate the equating accuracy, the following formula can be used [7].

$$RMSD(a) = \sqrt{\frac{\sum_{i=1}^{N}\left(a_2^{*} - a_1\right)^2}{N}} \quad (6)$$

$$RMSD(b) = \sqrt{\frac{\sum_{i=1}^{N}\left(b_2^{*} - b_1\right)^2}{N}} \quad (7)$$

Where

N : number of items

$a_2^{*}$ : discrimination index of test 1 after being equated with test 2

$a_1$ : discrimination index of test 1

$b_2^{*}$ : difficulty index of test 1 after being equated with test 2

$b_1$ : difficulty index of test 1

The equating method used is the Haebara and Stocking-Lord methods. The try out exam that was carried out consisted of two question packages. The problem was whether the question packages could measure objectively against students. To avoid the feelings of disadvantages or advantages among students due to different question packages, a score equating process is done to measure students' abilities objectively.

## 2. METHOD

### *2.1. Type of Research*

This research is a descriptive exploratory study investigate the equivalence of two question packages based on the results of the 2014 Junior High School mathematics try out scores in Makassar. This was done since the question packages were not able to measure students' abilities equally. Students obtaining high scores in package 1 did not necessarily mean that they have higher abilities than students who worked on package 2. It could be caused by package 2 being more difficult. Therefore, an equating is administered based on the student test scores to see how the students' real abilities.

### *2.2. Research Subjects and Objects*

The subjects consisted of groups of students from the Junior High School of Makassar who participated in the Mathematics MNE try out for the 2013/2014 academic year. The objects chosen were multiple choice of math questions consisting of 40 items with a dichotomy score.

### *2.3. Data Collection Techniques*

The data collection technique in this study used documentation techniques, by collecting student responses to the mathematics MNE try out in Makassar which consisted of 2 question packages, namely question package 1 which was responded to by 2099 students, and question package 2 which was responded to by 2068 students.

### *2.4. Data Analysis Techniques*

The equating design used common items design with the following model:

**Table 2.** Equating Views with Common Items

| Population | Sample | Package 1 | *Anchor* | Package 2 |
|---|---|---|---|---|
| P | 1 | 37 | 3 | |
| Q | 2 | | 3 | 37 |

The analysis used was the IRT approach referring to the Haebara method. The step of estimating the equating parameters utilized the equateIRT package in the R program.

## 3. RESULT & DISCUSSION

### *3.1. Result*

#### *3.1.1. The Analysis of Question Package Items 2PL Model of Package 1 and Package 2*

The data analysis used the Item Response Theory (IRT) or modern theory approach. It was intended that the score obtained does not depend on the ability of students or the sets of tests used. The use of the 2-parameter logistic model (2PL) refers to the number of fit items and the accuracy of the information. Researchers had analyzed 1PL, 2PL, and 3PL. The results show that the 2PL model is better than other models.

The total information of package 1 based on the estimation with the 2PL model in the range of capability of -4 to 4 was 16.72 (77.97%), and the total information package 2 based on the estimation with the 2PL model in the range of capability of -4 to 4 was 10.61 or equal to (73.31%). In general, the information functions of package 1 and package 2 can be seen in the following graph:

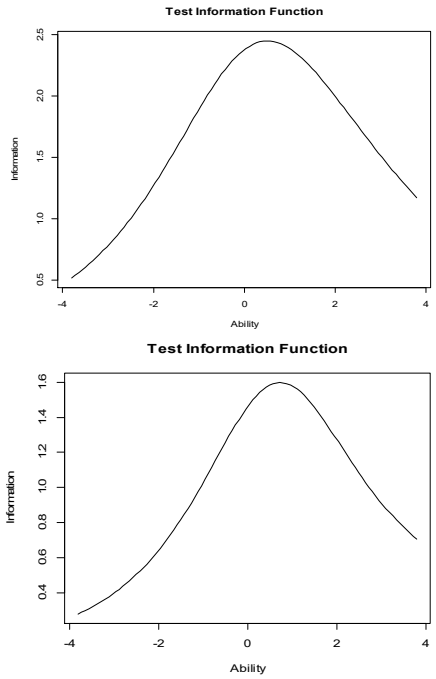


**Figure 1.** Information Functions on Package 1 and Package 2

Figure 1 shows that the test sets, both package 1 and package 2, can measure the ability of test-takers from -4 logit to 4 logit well. This means that this test can be used to measure the ability of students from low to high levels properly. In addition, the displayed graph follows a normal distribution with the maximum value being the average ability of 0 logit. The following is the comparison of item characteristics of each package based on the 2PL model and the total information from each package:

**Table 3.** Number of fit items and Test Information for Package 1 and Package 2

| Package 1 | | Package 2 | |
|---|---|---|---|
| Number of Item Fit | Test Information | Number of Item Fit | Test Information |
| 28 | 77.97% | 22 | 73.31% |

#### *3.1.2. The Analysis of Question Package Items 2PL Model of Package 1 and Package 2*

The item parameter estimation was done through the IRT approach using the 2PL model. The following are the steps for estimating parameters using 2 packages test:

a. Estimated using the 2PL model

```
> m1<-ltm(paket1~z1)
> m2<-ltm(paket2~z1)
```

b. Conducted Variant and Covariance for all IRT Models

```
> estm1 <- import.ltm(m1, display = FALSE)
> estm2 <- import.ltm(m2, display = FALSE)
```

c. Listed the Coefficients and Covariance matrices for all IRT Models

```
estc <- list(estm1$coef, estm2$coef)
estv <- list(estm1$var, estm2$var)
paket <- paste("paket", 1:2, sep = "")
```

d. Created a class object with modIRT syntax

```
> mod2PL<-modIRT(coef = estc, var = estv,
   names = paket, display = FALSE
```

e. Estimated all direct equivalent coefficients for Package 1 and Package 2 by involving the items together with the Haebara method

```
> HE2PL<-alldirec(mods = mod2PL, method =
   "Haebara")
> HE2PL
Direct equating coefficients
Method: Haebara
Links:
paket1.paket2
```

```
paket2.paket1
> summary(HE2PL)
Link: paket1.paket2
Method: Haebara
Equating coefficients:
   Estimate  StdErr
A 1.2614 0.18168
B 1.3890 0.20738
Link: paket2.paket1
Method: Haebara
Equating coefficients:
  Estimate   StdErr
A 0.92365   0.14256
B -0.84544 0.15378
```

### *3.1.3. Equating Regression Coefficient*

Based on the previously obtained output, the equating coefficient and direction of the equating are briefly presented in the following table:

**Table 4.** Equating Estimation of Difficulty and Difference Indices

| Model | Direction | a | b |
|---|---|---|---|
| 2PL | 1 to 2 | 1.2614 | 1.389 |
| | 2 to 1 | 0.92365 | -0.84544 |

**Table 5.** Regression Equations for Estimation of Difficulty and Discriminant Index

| Model | Direction | Regression Equations |
|---|---|---|
| 2PL | 1 to 2 | bi*= 1.26bi + 1.39 |
| | | ai* = ai/1.26 , ci* = 0 |
| | 2 to 1 | bi*= 0.92bi -0.85 |
| | | ai* = ai/0.92 , ci* = 0 |

Based on Table 4, it can be concluded that package 2 tends to be more difficult than package 1. This can be seen that the coefficient of 1 to 2 results in positive A and B values. Hence, if the item parameter is equated, it will produce a larger and positive value. Based on Table 5, the equating of the mathematics MNE try out question package 1 to package 2 or vice versa which can be done by referring to the new equation to equate the test. With the help of MS. Excel, the item parameter estimation can be done until the results are obtained in the 2PL model to determine the equivalence of package 1 and package 2.

### *3.1.4. Accuracy of Equalization Estimates*

After determining the regression equation for each direction and the IRT model used, the next step was to determine the estimation accuracy based on the Haebara method by calculating the smallest RMSD value. The following table presents a comparison of the RMSD values of the equating directions.

**Table 6.** Comparison of Root Mean Square Different (RMSD)

| Model | Direction | Characteristic | RMSD |
|---|---|---|---|
| 2 PL | 1 to 2 | b | 7.432 |
| | | a | 0.086 |
| | 2 to 1 | b | 18.354 |
| | | a | 0.034 |

## *3.2. Discussion*

The following is an estimation of the students' ability after the Haebara method was administered in the 2PL model.

**Table 7.** Comparison of Package 2 and Package 1 Scores After Equating

| Theta | The Score of Package 2 | The Score of Package 1*HA |
|---|---|---|
| -4.0 | 5.44 | 5.86 |
| -3.5 | 5.80 | 6.34 |
| -3.0 | 6.22 | 6.91 |
| -2.5 | 6.72 | 7.60 |
| -2.0 | 7.32 | 8.44 |
| -1.5 | 8.03 | 9.44 |
| -1.0 | 8.88 | 10.62 |
| -0.5 | 9.88 | 11.94 |
| 0.0 | 11.01 | 13.35 |
| 0.5 | 12.26 | 14.75 |
| 1.0 | 13.57 | 16.09 |
| 1.5 | 14.86 | 17.32 |
| 2.0 | 16.09 | 18.45 |
| 2.5 | 17.24 | 19.49 |
| 3.0 | 18.29 | 20.44 |
| 3.5 | 19.26 | 21.32 |
| 4.0 | 20.16 | 22.14 |

Based on Table 7, it can be seen that the capabilities with a logit or theta scale in each package have differences. However, after it was equated using Haebara method, the values are relatively similar even

though package 2 looks a bit low. This is because package 2 has a higher index of difficulty compared to package 1.

Visually, the following is the comparison graph of the test-takers' ability who answered package 2 and test-takers' ability who answered package 1 which has been equated with the Haebara method.

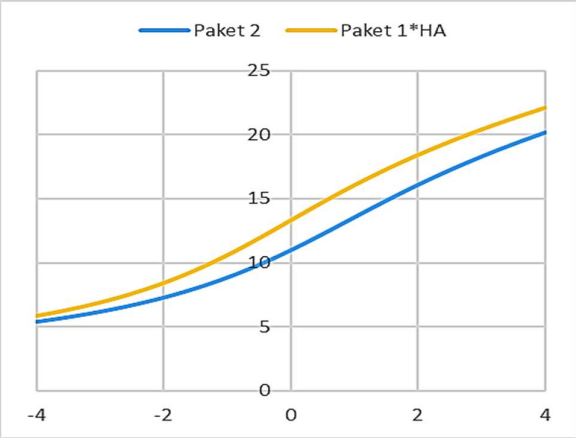


**Figure 2.** The Ability Comparison of the Test Participants of Package 1 and Package 2

Based on Figure 2, it can be seen that equating package 1 to package 2* based on 2PL with the Haebara method has a very high level of precision. This means that by equating package 1 to package 2 using the Hebara method, a relatively equal score will be obtained. Thus, the scores of students who worked on package 1 and package 2 would be fairer. This is in accordance with the findings of Retnawati who suggested the characteristic curve method or IRT and the Haebara method, tend to produce scores with small errors [8]. This is confirmed by Yusron, Retnawati, & Rafi [9] who state that equating using the Haebara method results in better equality than other methods. Although it differs from the results found by Kilmen and Demirtasli [6], who recommend the Stocking-Lord model as the model of equating, however, it applies to a sample size of 500-1000, while the sample size used in this study was around 2000 participants.

The equating is done to produce a fairer and more accountable score, especially if using a different set of questions. From the results of this study, it can be seen that, even though an equating had been carried out, there was still a difference in the score, although the difference was small because only 3 common items/anchors were used. In other cases, more anchor questions can be used to obtain a more precise equating result. According to Retnawati [8] two things that need to be considered in doing the equating, that is, the parameter estimation process including the number of respondents, the number of items, the estimation method, the estimation process for the equating of item parameter distribution, the distribution of ability parameter, method, number of items, and the software. Meanwhile, according to Kolen and Brennan [10], procedures are also taken into consideration to evaluate the extent to which test takers are measured with the same precision for all test sets.

## 4. CONCLUSION

From the discussion, it can be concluded that the relationship between try out questions of package 1 and package 2 cannot be said to be equivalent. After analyzing and calculating the regression equation, it can be concluded that package 2 of MNE tray out questions has a higher level of difficulty than package 1. Equating package 1 to package 2* through the Haebara method for the 2 PL model will provide equal or fair scores for the test takers who work on package 1 and package 2. The results obtained show a relatively equal score after being equated, although there is a slight difference in scores. The results of this study also found that the difference in scores after equating was relatively small. It is suggested that more common questions should be used for writing package questions

## AUTHORS' CONTRIBUTIONS

First author conceived of the presented idea. second author developed the theory and performed the computations. First author encouraged second author to investigate and supervised the findings of this work. All authors discussed the results and contributed to the final manuscript. All authors wrote the manuscript.

## ACKNOWLEDGMENTS

First of all, thanks to ALLAH Subhanahu Wa Taala for his mercy and guidance in giving me full strength to complete this paper. A lot of thanks to my advisor, Prof. Badrun Kwrtowagiran and Prof. Heri Retnawati for all of his support and guidance in helping me. Then, I would like thanks to my husband, for supporting me mentally and physically during finishing this paper. In addition, grateful acknowledgement to all of my friends who never give up in giving their support to me

## REFERENCES

[1] R. K. Hambleton dan H. Swaminathan, *Item Response Theory: Priciple and Applications*, 1 st editi. Springer Science+Business Media, LLC, 1985.

[2] L. Crocker, J. Alglna, M. Staudt, S. Mercurio, K. Hintz, dan R. A. Walker, *I ntroduction to Classical and Modern Test Theory*. Mason, Ohio, 2008.

[3] A. A. V Davier, *Statistics for Social and Behavioral Sciences: Statistical Models for Test Equating, Scaling, and Linking*. New York and London: Springer Science+Business Media, LLC, 2011.

[4] H. Retnawati, "Perbandingan Metode Penyetraan Skor Tes Menggunakan Butir Bersama dan Tanpa Butir Bersama," *J. Kependidikan*, vol. 46, no. 2, hal. 164–178, 2014.

[5] R. K. Hambleton, H. Swaminathan, dan D. J. Rogers, *Fundamentals of Item Response Theory*. California: SAGE Publications, Inc., 1991.

[6] S. Kilmen dan N. Demirtasli, "Comparison of Test Equating Methods Based on Item Response Theory According to the Sample Size and Ability Distribution," *Procedia - Soc. Behav. Sci.*, vol. 46, no. 1980, hal. 130–134, 2012.

[7] S.-H. Kim dan A. S. Cohen, "A comparison of linking and concurrent calibration under item response theory," *Apllied Psychol. Meas.*, vol. 22, no. 2, hal. 131–143, 1998.

[8] H. Retnawati, "Perbandingan Metode Penyetaraan Skor Tes Menggunakan Butir Bersama dan Tanpa Butir Bersama," *J. Kependidikan*, vol. 46, no. 2, hal. 164–178, 2016.

[9] E. Yusron, H. Retnawati, dan I. Rafi, "Bagaimana hasil penyetaraan paket tes USBN pada mata pelajaran matematika dengan teori respons butir ?," *J. Ris. Pendidik. Mat.*, vol. 7, no. 1, hal. 1–12, 2020.

[10] M. J. Kolen dan R. L. Brennan, *Test equating, scaling and linking. Methods and Practice*, Third. New York and London: Springer Science+Business Media, LLC, 2014.