

## Which one is more accurate, BILOG or R program? (a comparison for score test equating)

Dian Normalitasari Purnama<sup>1</sup>, Samsul Hadi<sup>2</sup>, Sukirno<sup>1</sup>, Heri Retnawati<sup>2</sup>, Rizki Nor Amelia<sup>3</sup>

<sup>1</sup>Department of Accounting Education, Faculty of Economics and Business, Universitas Negeri Yogyakarta, Yogyakarta, Indonesia

<sup>2</sup>Department of Educational Research and Evaluation, Graduate School, Universitas Negeri Yogyakarta, Yogyakarta, Indonesia

<sup>3</sup>Department of Integrated Sciences, Faculty of Mathematics and Natural Sciences, Universitas Negeri Semarang, Semarang, Indonesia

### Article Info

#### Article history:

Received Jan 26, 2023

Revised Aug 15, 2023

Accepted Sep 30, 2023

#### Keywords:

Accuracy

BILOG

Equating test

Item response theory

R program

Root mean square deviation

### ABSTRACT

Evaluation may be carried out using different tests that are not necessarily parallel. Students with lower abilities may get higher scores while those with higher abilities get lower scores. Measurement errors caused by this condition require test equating. Several computer programs, including Bilog and the R program, can be used for test equating. Each program has a different level of accuracy, and the accuracy of the equating results will affect the standard errors of equating. This study aimed to find out the most accurate equating test method and the accuracy of the estimated BILOG and R program. This research used two sets of tests with equivalent group designs. The determination of the most accurate equating method was based on the root mean square deviation (RMSD) value. Equating test packages, A to package B with BILOG program estimation on the mean and Sigma method resulted in RMSD value of 0.320. In the mean and mean method, RMSD value is 0.250. Meanwhile, equating package A to package B using R program on the mean and sigma resulted in RMSD value of 0.300, and the mean and mean method with the RMSD value is 0.272. The mean and mean yield RMSD values smaller than the mean and sigma methods. Therefore, the mean and mean method is more accurate and applicable. Moreover, the estimation results of the BILOG program are more accurate and can be used in test equating.

*This is an open access article under the [CC BY-SA](#) license.*



### Corresponding Author:

Dian Normalitasari Purnama

Accounting Education Department, Faculty of Economics and Business, Universitas Negeri Yogyakarta

Caturtunggal, Depok, Sleman 55281, Special Region of Yogyakarta, Indonesia

Email: dianpurnama@uny.ac.id

## 1. INTRODUCTION

Assessment and evaluation of student learning outcomes are important things to do. This is one of the efforts made to monitor the achievement of student competencies after the learning process is carried out. Assessment informs students about their abilities, strengths, and weaknesses and shows the strategies used in the learning process [1]. Assessment is not just a task that is assessed from a series of questions [2], but it is rather a method of examining students' knowledge and behavior during the teaching and learning process [3] in order to investigate what students already know [4]. Thus, an effort is needed to develop a strict and strong measurement tool in the educational context [5]. A good assessment enables teachers to identify whether students can achieve learning outcomes [6]. Each assessment aims to gauge and categorize test-takers' proficiency in a certain area [7]. Assessment and evaluation are also carried out to control the quality of national education as a form of accountability for education providers to interested parties, including

students, institutions, and educational programs. When compared periodically, from school to school, or district to district, student scores may show the progress or advancement of educational outcomes.

One of the main challenges in education is developing test items to determine the extent of students' understanding after carrying out the learning process. Teachers, schools, or educational institutions in each region use different test instruments to carry out the assessment. In standardized testing, test equating is important to ensure fairness for test-takers [8]. The different test forms which are not necessarily parallel means that the level of difficulty or distinguishing power of the two test packages is sometimes different [9]. This can cause students with lower abilities to get higher scores when working on easy questions, while students with higher abilities get lower scores when taking more difficult tests. This situation can cause measurement errors when examinee's ability and the item's features could predict or explain the success of an examinee [10]. If there is a measurement error, the evaluation results do not show the actual level of student competence. Therefore, it is necessary to make an equal test.

Testing programs that employ numerous exam forms or that introduce new test formats must demonstrate that the results and the conclusions drawn from them are equivalent [11]. Making an equal test for two or more question packages is not easy or probably impossible, as there must be differences. In addition, organizing a parallel multi-package test is almost impossible [12]. The main challenge in administering several test packages standard is related to the assurance that these packages are equal and measure the same indicator [13]. There is no guarantee that the difficulty level of each item will be the same, even though the test was created using the same testing specifications. To compare assessment results, equalizing scores on different test instruments is necessary. The results of this comparison can be used to determine the quality of education between schools or regions.

Two parallel test packages are considered "equated" for one group of examinees when the standard deviations and means of the two test forms are the same [8]. An equating technique will be used when the parallelism of the two test packages has been established. Equating is a statistical procedure that converts the raw test results onto a standard scale that is used for all test formats. It is based on the supposition that the testing procedures have been followed quite precisely to guarantee that the same build is being assessed [14]. Equating is a statistical procedure used to equalize test results so they can be compared across tests. [15]. Comparing test results from various test formats is done via test score equating [16]. Even when students used various test packages, equating can compare their scores [17]. The process of equating two different forms of the same test is often briefly referred to as test equating [18]. This will prevent test takers from being adversely affected by the easier or harder test packages they are given. Going through equating, even though test packages are relatively varied, the test takers' aptitude can be detected correctly in the future if they work on similar exams [19].

Test equating can be done using either the item response theory (IRT) or the classical test theory (CTT) methods [20], [21]. The equivalence test in CTT must have the same reliability index. According to IRT, which makes use of a mathematical model, the likelihood that test takers will correctly respond to a question depends on both their ability and the item characteristics [22], [23]. One of the most important properties of IRT is item-free person measurement and person-free test calibration [24]. One of the most important properties of IRT is parameter invariance or item independent and sample independent [25]–[28]. The CTT is selected due to its ease of application despite its limitations in measuring the item difficulty level and discrimination since both indicators' calculation is based on the test taker's total score [29]. IRT model is understood to be better at revealing the score of the test than CTT model because the result of the IRT model is not affected by the characteristic of test taker groups [29].

Because IRT has invariance properties in its parameter, test equating using IRT is more representative than test equating using CTT. Both the ability parameter and the test parameter are invariant [30]. The comparison of the equating test between the CTT and IRT revealed that the IRT's equating standard error is lower than that of the CTT's [31]. There are two types of test equating, namely vertical equating and horizontal equating [32]. Vertical equating compares exams across levels of difficulty, while horizontal equating compares two different test versions [24]. Horizontal equating is performed on test instrument scores of the same difficulty level within the same level participant group [33]. It can also be defined as determining the equal score for differences [34]. Horizontal equating is proper when it is used for the security of a test, so several forms of tests are needed. These forms are not the same, but it is expected that they are similar in their content and difficulty. There are few equating ways that can effectively function when the difficulty, reliability, and content of tests vary so greatly between different forms [35].

Based on IRT, there are three testing equivalence designs that were employed: i) Anchor testing design; ii) Equivalent group design; and iii) Single group design [36]. In an equivalent group design, two tests were done to two equivalent groups drawn randomly from the same population (presumed that they had the same level of ability) [37], [38]. The equivalent group design may not only lessen the impact of activity and boredom, but it may also cause bias since the groups' levels of ability were probably not equivalent [39]. The use of a large sample was advised to lessen the likelihood of bias. The non-anchor item equating

utilizing the equivalent group design was previously introduced in a prior study [40]. Equating the item response theory consists of mean and sigma methods, mean and mean, and characteristic curves [15].

In equating with the item response theory approach, there is a relationship between item parameters and a person's ability [22]. When the scale in the first test is equated with the scale of the second test, tests are carried out by two different groups as (1) and (2).

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

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

$b_{1j}$ ,  $b_{2j}$  are item difficulty indexes in test 1 and test 2.  $a_{1j}$ ,  $a_{2j}$  are discrimination indexes of test 1 and test 2.  $\alpha$ ,  $\beta$  is equating constant with  $\alpha$  and  $\beta$  are constants in equating. The relationship between parameter estimates of the items from the two test tools that are equated based on the mean sigma method with:

$$b_2 = \alpha b_1 + \beta, \text{ the value is obtained}$$

$$\bar{b}_2 = \alpha \bar{b}_1 + \beta \text{ and } \sigma(b_2) = \sigma(b_1)$$

$$\text{then } \alpha = \frac{\sigma(b_2)}{\sigma(b_1)}$$

$$\beta = \bar{b}_2 - \alpha \bar{b}_1$$

The formula for the mean and mean methods is as:

$$b_2 = \alpha b_1 + \beta$$

$$a_2 = \frac{a_1}{\alpha}, \bar{b}_2 = \alpha \bar{b}_1 + \beta \text{ is obtained}$$

$$\text{so } \alpha = \frac{a_1}{a_2}$$

$$\beta = \bar{b}_2 - \alpha \bar{b}_1$$

Where:

$\bar{b}_1$ ,  $\bar{b}_2$ : parameter estimate of the item level difficulty in 1st and 2nd test

$a_1$ ,  $a_2$ : estimation of item discrimination index in 1st and 2nd test

$b_1$ ,  $b_2$ : the Mean level of difficulty of the item in 1st and 2nd test

$\sigma(b_1)$  and  $\sigma(b_2)$ : standard deviation of the difficulty level of test 1 and 2  $\alpha$  and  $\beta$  are constants used to equating the test [15].

After  $\alpha$  and  $\beta$  are known, the results of the grain parameter estimation and ability parameters of the test device 1 are equated on the same scale as the test 2 device using:

$$b_2^* = \alpha b_1 + \beta$$

$$a_2^* = \frac{a_1}{\alpha}$$

$$\theta_2^* = \alpha \theta_1 + \beta$$

where:

$b_2^*$ : the difficulty level of the item on the 1st test 1 after equate on the scale of 2nd test

$a_2^*$ : discrimination index on the 1st test 1 after equate on the scale of 2nd test

$\theta_2^*$ : student's ability on the 1st test 1 after equate on the scale of 2nd test [22]

Psychometrics has become interested in equating methods, and the development of new methods in equating test [41]. Besides the various methods that can be used for the equating process, there are also various kinds of programs that can be used to estimate the parameters of the problem and then find the equations. Several methods and programs produce different accuracy. This will result in different results in

measurement of the same package of questions. Therefore, it is important to examine the equating method and what programs are most accurate in the equating process.

Related to the existence of two test devices and several methods and software that can be used for equating processes, a comparison is needed to see what methods and software produce the least errors. Any IRT program, including Winstep and BILOG may be used for equating [42]. BILOG is the program used for alternative form equating [43]. It is an IRT application that is widely used in equating tests. There have been many studies on equating using BILOG program [44]–[47]. A newly developed application that can be used for equating tests using the IRT is the R program. There are several studies on equating tests using this program [19], [48]–[50].

Currently, there has never been a study that reveals the equivalence and accuracy problems between the mean/mean method and mean/sigma based on the estimation results between the BILOG program and the R program. There are studies that compared equating methods, but they used the IRTEQ Program [32], [51] or Rasch model [52], [53]. Other studies focused more on comparing equating methods that focused on the ability distribution estimation and sample size [54], comparing equating methods with the use of different test formats [55], investigating the effect of the number of biased items and their distribution in the equated form on equating error [56] and examining the impact of anchor items in the mean/sigma method on true score equating using IRT [57]. This study will describe the equating of the problem with the mean/sigma methods and the mean/mean based on the estimation results of the BILOG and R program as well as find out the level of accuracy.

Equating accuracy can be measured by comparing of the value of root mean square of error (RMSE) or commonly called root mean square of deviation (RMSD). The accuracy of equating is calculated by looking at the smallest RMSE value [54]. To calculate equating accuracy the formula can be used [58].

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

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

$$\text{RMSD}(\theta) = \sqrt{\frac{\sum_{i=1}^N (\theta_2^* - \theta_1)^2}{N}}$$

## 2. RESEARCH METHOD

This descriptive-quantitative study aimed to compare the vocational high school (VHS) accounting exams which are offered in packages A and B. This study was carried out in Yogyakarta, Indonesia. The research participants were grade XII (third grade) students of VHS. The 650 test takers from six vocational high schools were chosen using the stratified random selection technique, and they were given test instruments in the form of answer sheets. The data collected through documentation. The Aiken formula was used to determine the instrument's content validity [59].

Item response theory with BILOG program application was used to analyze the test instrument and produce three phases of output. In the first phase, it revealed the number of test participants answering test items correctly, the ratio of correct answer probability divided by wrong answer probability, and the biserial coefficient. The second phase obtained the data on item parameters according to the item response theory model used. The 1PL model covers the data on the difficulty level, the 2PL model covers information on the difficulty level and discrimination index, and the 3PL model measures the difficulty level, discrimination index, and guessing factor. In estimating the parameter, the logistic model with the highest number of fit items was used. Fit items have a calculated Chi-square value that is less than the table's Chi-square value or a p-value that is higher than 5%. The purpose of the goodness of fit test was to determine whether the items chosen were consistent with the model that was utilized.

After knowing the characteristics of the problem, the equivalence of two question packages was tested. The test was aimed at examining whether package A and B contained parallel problems or not. If, based on the results of the equality test, it was proven that the two package questions were not parallel, it would be necessary to equate the question package. Conversely, if the test results showed that the two packages were parallel questions, then there would be no need to do an equating process. Allen and Yenn stated that two sets of questions are said to be parallel if they have the same mean and variance [60]. Therefore, parallel testing was carried out with the SPSS program to see the difference in mean and variance.

There are two methods used in the equating process of this problem, namely mean & sigma and mean & mean methods. The process of equating the test instruments was carried out after equality testing that

aim to prove the tests to be parallel. The equating process was carried out based on the results of parameter estimation from the BILOG program analysis and R program. From the equating process, the constant conversion information of the test device was obtained. The determination of the convergence constant used for equating the test was based on a predetermined method. This equating process was done using equivalent group designs. The reason why the researchers chose this design was based on the data obtained in the field. The data in the form of students' answers were obtained from two different test instruments given to two different groups of students but with equivalent abilities. On both test instruments, there was no anchor.

### 3. RESULTS AND DISCUSSION

#### 3.1. The validity of instrument

This research involved five raters to estimate the validity using Aiken formula. The content validity of an instrument is the extent to which the items in the instrument represents the components in the over-all area of the contents of the object to be measured and the extent to which the items reflect behavioral traits that will be measured [61]. In sum, the validity of the test items in both packages according to Aiken formula is relatively good (65% item on package A, and 67.5% on package B).

#### 3.2. Characteristics of accounting test items based on item response theory

Data analysis under the three-parameter logistics was conducted to determine the model that fits the items (1-PL, 2-PL, and 3-PL). The fit model analysis was assisted by BILOG software. The most suitable logistic model is the model that produces the most fit items with the criteria of chi-square value greater than 0.05 [62].

##### 3.2.1. Results question analysis

Based on the results of the BILOG phase 1 analysis, nine items in test package A (items 1, 3, 15, 22, 28, 35, 38, 39, and 40) and package B (items 4, 7, 10, 15, 19, 25, 26, 32, and 35) were not included in the subsequent analysis because the biserial value was less than 0.3. In phase 1, an item that had a biserial value less than 0.3 was not included in the subsequent analysis because it could interfere with the analysis process [40]. The results of the goodness of fit analysis Table 1 shows that the item analysis based on the Item Response Theory fits the 2-PL model for both packages. Then, based on the logistic parameter two (2PL), the model adds completion parameters to the difficulty level [63], [64]. Further analysis carried out to see the good or poor characteristics of the questions with the 2-PL model show the following results: package A had 27 good items and 13 poor items, while package B had 24 good item and 16 poor items. It was found that the problem was caused by the level of difficulty and index of discrimination that exceeded the criteria (i.e., the value exceeded +2 for item difficulty).

Table 1. Goodness of fit (p-value) on package A and B

Category	Package A			Package B		
	1-PL	2-PL	3-PL	1-PL	2-PL	3-PL
Fit	15	32	31	11	30	18
Misfit	25	8	9	29	10	12

#### 3.3. Equating of accounting test

To determine whether both packages are parallel, it is necessary to verify the accounting test equation for both packages A and B. The t-test can be used to test the equation for the test instruments. The t-test result demonstrates the significant value for assumed equal variances at  $0.000 < 0.05$ . Therefore, equating is required since the mean score generated from packages A and B differs (mean difference=3.092).

After packages A and B proved to be not parallel, the equating between the two test instruments must be done. In the equating process using this 2PL model,  $c_j$  is assumed to be 0 for all  $j$  [16] and which package will serve as the standard for equating must be decided. This study equates the package A test to the package test B. The parameter estimate is based on the response of the data to the response of students to the logistical model. Based on the results of the analysis with the BILOG program, it is proven that the 2-PL model has item numbers with the most item characteristics and is the most suitable for application. Table 2 presents the summary of the item parameters of both test packages.

Table 2. Summary of the parameters

Item	Package A		Package B	
	Difficulty level	Discrimination index	Difficulty level	Discrimination index
5	-0.320	1.364	0.843	1.084
8	-0.608	1.444	-0.677	1.719
9	-0.136	1.842	-0.282	1.793
12	-0.369	1.634	-0.949	1.606
13	-0.484	1.548	-0.307	1.709
16	-0.262	1.197	-0.154	1.167
17	-0.743	1.508	0.927	0.628
20	0.297	1.199	0.066	1.165
21	1.511	0.573	1.529	0.630
23	-0.103	1.552	0.169	1.787
24	-1.116	0.606	-0.270	1.848
27	-0.035	1.591	0.229	1.619
30	0.624	0.981	0.682	1.049
33	0.602	1.272	0.738	0.910
34	0.427	1.601	0.791	1.402
36	-0.468	1.317	0.391	0.887
37	-0.730	0.6110	-1.375	0.905
$\mu$	-0.1125	1.2847	0.1383	1.2887
$\Sigma$	0.6413	0.3855	0.7490	0.4222

### 3.3.1. Mean/sigma and mean/mean method based on BILOG

In the mean/sigma method, the calculation of constants  $\alpha$  and  $\beta$  uses the Mean and standard deviation of the level of difficulty, and the resulted constants are  $\alpha = 1.168$  and  $\beta = 0.270$ . From the values of constants  $\alpha$  and  $\beta$ , the equality of package A (x) to package B (y) is obtained.

$$\theta_2^* = 1.168\theta x + 0.270$$

$$b_2^* = 1.168bx + 0.270$$

$$a_2^* = \frac{a_1}{1.168}$$

In mean/mean method, the calculation of constants  $\alpha$  and  $\beta$  uses the mean of difficulty level and discrimination index, which resulted in constants  $\alpha = 0.997$  and  $\beta = 0.250$ . From the constants  $\alpha$  and  $\beta$ , it is found that the equation of package A (x) to package B (y) is as:

$$\theta_2^* = 0.997\theta x + 0.250$$

$$b_2^* = 0.997bx + 0.250$$

$$a_2^* = \frac{a_1}{0.997}$$

By using the  $\alpha$  and  $\beta$  constants, the transformation of item parameters is carried out so that the equating item parameter for both methods by Biologist obtained in Table 3. After being equalized, the item difficulty level and item discrimination index show that the mean/sigma method provides a higher mean ( $0.1386 > 0.1378$ ) and a higher standard deviation ( $0.7491 > 0.639$ ) than the mean/mean method; meanwhile the item discrimination index show that the mean/sigma method provides a lower mean ( $1.0999 < 1.2886$ ) and a lower standard deviation ( $0.3301 < 0.3867$ ) than the mean/mean method.

### 3.3.2. Mean/sigma and mean/mean method based on r program

In the mean/sigma method, the calculation of the constants  $\alpha$  and  $\beta$  uses the Mean and standard deviation of the level of difficulty, and the constants resulted are  $\alpha = 1.241$  and  $\beta = 0.2946$ . From the values of constants  $\alpha$  and  $\beta$ , the equality of package A (x) to package B (y) is obtained.

$$\theta_2^* = 1.241\theta x + 0.295$$

$$b_2^* = 1.241bx + 0.295$$

$$a_2^* = \frac{a_1}{1.241}$$

In the mean/mean method of calculating the constants  $\alpha$  and  $\beta$  using the mean of the level of difficulty and the discrimination power, and the constants are  $\alpha = 1.0729$  and  $\beta = 0.2743$ . From the values of constants  $\alpha$  and  $\beta$ , the equality of package A (x) to package B (y) is obtained.

$$\theta_2^* = 1.073\theta_x + 0.274$$

$$b_2^* = 1.073bx + 0.274$$

$$a_2^* = \frac{a_1}{0.274}$$

Using the  $\alpha$  and  $\beta$  constants, the transformation of item parameters is carried out so that the equating item parameter for both methods by R program is obtained in Table 4. After being equalized, consistent with the BILOG results, the item difficulty level and item discrimination index show that the mean/sigma method result in a higher mean (0.1445>0.1444) and a higher standard deviation (0.7593>0.6565) than the mean/mean method. Meanwhile the item discrimination index show that the mean/sigma method provides a lower mean (1.1832<1.3684) and a lower standard deviation (0.3818<0.4416) than the mean/mean method.

Table 3. Conversion of package A to package B using BILOG

Item number	Mean/sigma method				Mean/mean method			
	Package A		Package B		Package A		Package A <sub>2</sub> <sup>*</sup>	
	b Initial	a Initial	(b <sub>2</sub> <sup>*</sup> )	(a <sub>2</sub> <sup>*</sup> )	b Initial	a Initial	(b <sub>2</sub> <sup>*</sup> )	(a <sub>2</sub> <sup>*</sup> )
5	-0.320	1.364	-0.1038	1.1678	-0.320	1.364	-0.0690	1.3681
8	-0.608	1.444	-0.4401	1.2363	-0.608	1.444	-0.3562	1.4483
9	-0.136	1.842	0.1112	1.5771	-0.136	1.842	0.1144	1.8475
12	-0.369	1.634	-0.1610	1.3990	-0.369	1.634	-0.1179	1.6389
13	-0.484	1.548	-0.2953	1.3253	-0.484	1.548	-0.2325	1.5527
16	-0.262	1.197	-0.0360	1.0248	-0.262	1.197	-0.0112	1.2006
17	-0.743	1.508	-0.5978	1.2911	-0.743	1.508	-0.4908	1.5125
20	0.297	1.199	0.6169	1.0265	0.297	1.199	0.5461	1.2026
21	1.511	0.573	2.0348	0.4906	1.511	0.573	1.7565	0.5747
23	-0.103	1.552	0.1497	1.3288	-0.103	1.552	0.1473	1.5567
24	-1.116	0.606	-1.0335	0.5188	-1.116	0.606	-0.8627	0.6078
27	-0.035	1.591	0.2291	1.3622	-0.035	1.591	0.2151	1.5958
30	0.624	0.981	0.9988	0.8399	0.624	0.981	0.8721	0.9840
33	0.602	1.272	0.9731	1.0890	0.602	1.272	0.8502	1.2758
34	0.427	1.601	0.7687	1.3707	0.427	1.601	0.6757	1.6058
36	-0.468	1.317	-0.2766	1.1276	-0.468	1.317	-0.2166	1.3210
37	-0.730	0.611	-0.5826	0.5231	-0.730	0.611	-0.4778	0.6128
$\mu$	-0.1125	1.2847	0.1386	1.0999	-0.1125	1.2847	0.1378	1.2886
$\Sigma$	0.6413	0.3855	0.7491	0.3301	0.6413	0.3855	0.6394	0.3867

Table 4. Conversion of package A to package B using R program

Item number	Mean/sigma method				Mean/mean method			
	Package A		Package B		Package A		Package A <sub>2</sub> <sup>*</sup>	
	b Initial	a Initial	(b <sub>2</sub> <sup>*</sup> )	(a <sub>2</sub> <sup>*</sup> )	b Initial	a Initial	(b <sub>2</sub> <sup>*</sup> )	(a <sub>2</sub> <sup>*</sup> )
5	-0.309	1.588	-0.0887	1.2797	-0.309	1.588	-0.0572	1.4801
8	-0.588	1.511	-0.4349	1.2177	-0.588	1.511	-0.3566	1.4083
9	-0.140	1.973	0.1210	1.5900	-0.14	1.973	0.1241	1.8389
12	-0.353	1.866	-0.1433	1.5038	-0.353	1.866	-0.1044	1.7392
13	-0.447	1.924	-0.2599	1.5505	-0.447	1.924	-0.2053	1.7933
16	-0.258	1.427	-0.0254	1.1500	-0.258	1.427	-0.0025	1.3300
17	-0.686	1.739	-0.5565	1.4014	-0.686	1.739	-0.4617	1.6208
20	0.237	1.373	0.5888	1.1065	0.237	1.373	0.5286	1.2797
21	1.494	0.559	2.1486	0.4505	1.494	0.559	1.8772	0.5210
23	-0.118	1.932	0.1483	1.5569	-0.118	1.932	0.1477	1.8007
24	-1.119	0.615	-1.0938	0.4956	-1.119	0.615	-0.9263	0.5732
27	-0.051	1.825	0.2315	1.4707	-0.051	1.825	0.2196	1.7010
30	0.546	1.059	0.9723	0.8534	0.546	1.059	0.8601	0.9870
33	0.513	1.419	0.9313	1.1435	0.513	1.419	0.8247	1.3226
34	0.352	1.825	0.7315	1.4707	0.352	1.825	0.6520	1.7010
36	-0.431	1.656	-0.2401	1.3345	-0.431	1.656	-0.1881	1.5435
37	-0.730	0.668	-0.5739	0.5383	-0.730	0.668	-0.4767	0.6226
$\mu$	-0.1211	1.4682	0.1445	1.1832	-0.1211	1.4682	0.1444	1.3684
$\Sigma$	0.6119	0.4737	0.7593	0.3818	0.6119	0.4737	0.6565	0.4416

The results of equating test items based on the results of the BILOG program parameter estimation on the mean/sigma methods produce equating  $b_2^* = 1.168bx+0.60$ , while the mean/mean methods equating are  $b_2^* = 0.997bx-0.250$ . Equating based on the results of analysis with program R on the mean/sigma methods produce the equation  $b_2^* = 1.241bx+0.295$ , while the mean/mean method yields the equation  $b_2^* = 1.073bx+0.274$ .

### 3.4. Accuracy of equating results based on root mean square deviation

In line with the findings of previous study which shows that the mean/mean method gives the smallest RMSD value compared to mean/sigma method [51], [65]. This study also proves that all RMSD values in the mean/mean method are lower than mean/sigma method, both in the estimation by BILOG and R program as shown in Table 5. Smaller RMSD values indicate more accurate equation results, so it can be concluded that the mean/mean method is better at detecting test takers' abilities when taking similar tests using relatively different test packages, compared to the mean/sigma method.

Table 5. Summary of RMSD based on the BILOG and R program

Parameter	BILOG program		R program	
	RMSD mean/mean	RMSD mean/sigma	RMSD mean/mean	RMSD mean/sigma
Item difficulty (b)	0.251	0.272	0.272	0.300
Discrimination index (a)	0.004	0.192	0.192	0.298
Ability ( $\theta$ )	0.250	0.320	0.279	0.355

RMSD value in equating using the BILOG program is smaller compared to equating based on the results of the analysis of the R program. In the equating test by calculating RMSD from the mean/mean method and the mean/sigma method, it can be seen that in general the RMSD value of the mean/mean method is smaller than the RMSD value of the mean/sigma method [54], [56], [65]. RMSD is calculated by utilizing the estimation result ability parameter and parameters of the equation's results. This result is determined by the capability of parameter estimates, which are scaled to equating the item parameter. This equating strategy made a big contribution to RMSD. The smaller the RMSD, the better and accurate equating method used [65].

The difference in the accuracy of the equating results between the BILOG and the R program is due to the difference in the analysis results by the two programs. Score equating is impacted by a variety of things. It is required to estimate the parameters initially before utilizing a few approaches for equating the parameters. Several variables have an impact on parameter estimate outcomes. These factors include the estimate model, the sample size, the number of test items, as well as the estimation technique for item and ability characteristics. The distribution of item parameters, distribution of ability parameters, comparison of estimate techniques, total number of items in the test, and software used are the variables that need to be focused on while equating [65].

So far, many studies have compared the accuracy of equating tests regarding the methods and sample size used [14], [66]. No research has been found to compare the accuracy of equating test results based on the analysis program used. Even though many analysis programs can be used for the equating tests process, each program has a different level of accuracy. The limitation of the results of this study is to find support for the results of previous studies. Other research needs to examine the accuracy of equating tests based on the analysis program used since the accuracy of the results of equating tests affects the equating scores. An accurate equating tests score will be able to measure test takers' ability more accurately [67].

## 4. CONCLUSION

The results of this study indicate that the two sets of accounting tests are different (not parallel), so an equation is needed. Equating was carried out using the BILOG program and the R program. Of the two equating methods used, the mean/mean method gave more accurate results than the mean/sigma method, which was indicated by a low RMSD value. The equating results generated from BILOG program and R program showed that the RMSD value is smaller in the BILOG program than the R program. As previous research comparing the accuracy of the BILOG program and the R program on equating tests has yet to be found, this research is something of a novelty. However, similar research with different data needs to be conducted to support the results of this study.

## ACKNOWLEDGEMENTS

Great appreciation and gratitude are due to the Accounting Subject Teachers in Yogyakarta, Indonesia for providing research facilities and permission.






## REFERENCES




- [1] A. Gullickson, "Review of Practical Assessment, Research, & Evaluation, Volume 10," *Journal of MultiDisciplinary Evaluation*, vol. 3, no. 4, pp. 199–203, 2007, doi: 10.56645/jmde.v3i4.94.
- [2] G. T. L. Brown, "The past, present and future of educational assessment: a transdisciplinary perspective," *Frontiers in Education*, vol. 7, 2022, doi: 10.3389/educ.2022.1060633.
- [3] A. Shaw *et al.*, "Thinking critically about critical thinking: validating the Russian HEIghten® critical thinking assessment," *Studies in Higher Education*, vol. 45, no. 9, pp. 1933–1948, Oct. 2020, doi: 10.1080/03075079.2019.1672640.
- [4] H. Goldstein, "A response to 'assessment and learning: fields apart?'" *Assessment in Education: Principles, Policy and Practice*, vol. 24, no. 3, pp. 388–393, 2017, doi: 10.1080/0969594X.2017.1319338.
- [5] G. Richmond, M. del C. Salazar, and N. Jones, "Assessment and the future of teacher education," *Journal of Teacher Education*, vol. 70, no. 2, pp. 86–89, 2019, doi: 10.1177/0022487118824331.
- [6] E. Care, H. Kim, A. Vista, and K. Anderson, "Education system alignment for 21st century skills: focus on assessment," Center for Universal Education at the Brookings Institution, 2018.
- [7] I. Himelfarb, "A primer on standardized testing: history, measurement, classical test theory, item response theory, and equating," *Journal of Chiropractic Education*, vol. 33, no. 2, pp. 151–163, 2019, doi: 10.7899/JCE-18-22.
- [8] C. Z. Fong and T. Y. Chuen, "Test score equating and item anchoring for high stakes examination," *Vietnam Journal of Educational Sciences*, vol. 18, no. 4, pp. 1–11, 2023, doi: 10.15625/2615-8957/22210401.
- [9] B. Kartowagiran, D. Mardapi, D. N. Purnama, and K. Kriswanto, "Parallel tests viewed from the arrangement of item numbers and alternative answers," *Research and Evaluation in Education*, vol. 5, no. 2, pp. 169–182, 2019, doi: 10.21831/reid.v5i2.23721.
- [10] L. H. Haw, S. B. Sharif, and C. G. K. Han, "Analyzing the science achievement test: perspective of classical test theory and Rasch analysis," *International Journal of Evaluation and Research in Education (IJERE)*, vol. 11, no. 4, pp. 1714–1724, 2022, doi: 10.11591/ijere.v11i4.22304.
- [11] B. Babcock and K. J. Hodge, "Rasch versus classical equating in the context of small sample sizes," *Educational and Psychological Measurement*, vol. 80, no. 3, pp. 499–521, Sep. 2020, doi: 10.1177/0013164419878483.
- [12] N. S. Petersen, M. J. Kolen, and H. D. Hoover, "Scaling, norming, and equating IRT," in *Educational Measurement*, American Council on Education/Macmillan Publishing, 1989.
- [13] B. Kartowagiran, S. Munadi, H. Retnawati, and E. Apino, "The equating of battery test packages of mathematics national examination 2013-2016," *SHS Web of Conferences*, vol. 42, 2018, doi: 10.1051/shsconf/20184200022.
- [14] B. O'Neill, B. Sorhaindo, J. J. Xiao, and E. T. Garman, "Financially distressed consumers: their financial practices, financial well-being, and health," *Journal of Financial Counseling and Planning*, vol. 16, no. 1, pp. 73–87, 2005.
- [15] M. J. Kolen and R. L. Brennan, *Test equating, scaling, and linking: methods and practices: third edition*. Springer New York, 2014, doi: 10.1007/978-1-4939-0317-7.
- [16] J. González and M. Wiberg, *Applying Test Equating Methods Methodology of Educational Measurement and Assessment*. Cham: Springer International Publishing, 2017.
- [17] S. Sukirno, "National Final Examination Test Equivalency: Why and How?" (in Indonesian), *Cakrawala Pendidikan*, vol. 26, no. 3, pp. 305–321, 2007, doi: 10.21831/cp.v3i3.3983.
- [18] W. J. van der Linden, "What is actually equated in 'test equating'? a didactic note," *Journal of Educational and Behavioral Statistics*, vol. 47, no. 3, pp. 353–362, Feb. 2022, doi: 10.3102/10769986211072308.
- [19] M. Elvira and S. Sainuddin, "Equating Test Instruments Using Anchor to Map Student Abilities Through the R Program Analysis," in *Proceedings of the International Conference on Engineering, Technology and Social Science (ICONETOS 2020)*, 2021, vol. 529, doi: 10.2991/assehr.k.210421.095.
- [20] M. Battauz, "Testing for differences in chain equating," *Statistica Neerlandica*, vol. 77, no. 2, pp. 134–145, Aug. 2023, doi: 10.1111/stan.12277.
- [21] W. Leônico, M. Wiberg, and M. Battauz, "Evaluating equating transformations in IRT observed-score and Kernel equating methods," *Applied Psychological Measurement*, vol. 47, no. 2, pp. 123–140, Oct. 2023, doi: 10.1177/01466216221124087.
- [22] R. K. Hambleton, H. Swaminathan, and H. J. Rogers, *Fundamentals of Item Response Theory*. SAGE Publications, 1991.
- [23] S. E. Stemler and A. Naples, "Rasch measurement v. item response theory: knowing when to cross the line," *Practical Assessment, Research and Evaluation*, vol. 26, pp. 1–16, 2021, doi: 10.7275/v2gd-4441.
- [24] R. K. Hambleton and H. Swaminathan, *Item Response Theory: Principles and Applications*. Springer Science, 1985.
- [25] O. O. Adedoyin, "Investigating the invariance of person parameter estimates based on classical test and item response theories," *International Journal of Educational Sciences*, vol. 02, no. 02, Jun. 2010, doi: 10.31901/24566322.2010/02.02.07.
- [26] K. Hori, H. Fukuhara, and T. Yamada, "Item response theory and its applications in educational measurement part I: item response theory and its implementation in R," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 14, no. 2, Oct. 2022, doi: 10.1002/wics.1531.
- [27] Ç. Reyhanlioğlu and N. Doğan, "An analysis of parameter invariance according to different sample sizes and dimensions in parametric and nonparametric item response theory," *Journal of Measurement and Evaluation in Education and Psychology*, vol. 11, no. 2, pp. 98–112, Jun. 2020, doi: 10.21031/epod.584977.
- [28] J. Brzezińska, "Item response theory models in the measurement theory with the use of LTM package in R," *Econometrics*, vol. 22, no. 1, pp. 11–25, 2018, doi: 10.15611/eada.2018.1.01.
- [29] E. Istiyono, W. S. B. Dwandaru, Y. A. Ledo, F. Rahayu, and A. Nadapdap, "Developing IRT-based physics critical thinking skill test: a CAT to answer 21st century challenge," *International Journal of Instruction*, vol. 12, no. 4, pp. 267–280, Oct. 2019, doi: 10.29333/iji.2019.12417a.
- [30] N. S. Aminah, "The Characteristics of Test Equating Methods for Dichotomous Data," (in Indonesian), *Jurnal Penelitian dan Evaluasi Pendidikan*, vol. 16, pp. 88–101, Jan. 2013, doi: 10.21831/pep.v16i0.1107.
- [31] Q. Zhang, J. He, J. Ren, Y. Zhang, and Y. Lu, "A comparative study of equating methods applied in standardized competence test for clinical medicine undergraduates," *Chinese Journal of Medical Education*, vol. 42, no. 7, pp. 577–580, 2022.
- [32] R. Setiawan, "A comparison of score equating conducted using Haebara and Stocking Lord method for polytomous," *European Journal of Educational Research*, vol. 8, no. 4, pp. 1071–1079, Oct. 2019, doi: 10.12973/eu-jer.8.4.1071.
- [33] S. Suharyono and A. Jaedun, "The portrait of high school student physics competence with vertical equating lens," in *Proceedings of the 5th International Conference on Current Issues in Education (ICCIE 2021)*, 2022, vol. 640, doi: 10.2991/assehr.k.220129.023.
- [34] L. Crocker and J. Algina, *Introduction to classical and modern test theory*. Cengage Learning, 2008.

- [35] L. L. Cook and D. R. Eignor, "IRT equating methods," *Educational Measurement: Issues and Practice*, vol. 10, no. 3, pp. 37–45, Sep. 1991, doi: 10.1111/j.1745-3992.1991.tb00207.x.
- [36] S. Wang, M. Zhang, and S. You, "A Comparison of IRT Observed Score Kernel Equating and Several Equating Methods," *Frontiers in Psychology*, vol. 11, p. 308, Mar. 2020, doi: 10.3389/fpsyg.2020.00308.
- [37] N. J. Dorans, T. P. Moses, and D. R. Eignor, "Principles and practices of test score equating," *ETS Research Report Series*, vol. 2010, no. 2, Dec. 2010, doi: 10.1002/j.2333-8504.2010.tb02236.x.
- [38] E. Yusron, H. Retnawati, and I. Rafi, "How are the results of the equating of test packages of mathematics USBN with item response theory?" (in Indonesian), *Jurnal Riset Pendidikan Matematika*, vol. 7, no. 1, pp. 1–12, Jul. 2020, doi: 10.21831/jrpm.v7i1.31221.
- [39] A. Moghadamzadeh, K. Salehi, and E. Khodaie, "A comparison method of equating classic and item response theory (IRT): a case of Iranian study in the university entrance exam," *Procedia - Social and Behavioral Sciences*, vol. 29, pp. 1368–1372, 2011, doi: 10.1016/j.sbspro.2011.11.375.
- [40] E. Miyatun and D. Mardapi, "Comparison of Test Equalization Methods According to Item Response Theory," (in Indonesian), *Jurnal Penelitian dan Evaluasi Pendidikan*, vol. 2, no. 3, Jan. 2004, doi: 10.21831/pep.v2i3.2083.
- [41] C. Akin Arikan, "A comparison of kernel equating methods based on neat design," *Eurasian Journal of Educational Research*, vol. 2019, no. 82, pp. 27–44, Aug. 2019, doi: 10.14689/ejer.2019.82.2.
- [42] C. H. Yu and S. E. O. Popp, "Test equating by common items and common subjects: concepts and applications," *Practical Assessment, Research and Evaluation*, vol. 10, no. 4, pp. 1–19, 2005.
- [43] M. F. Zimowski, E. Muraki, R. J. Mislevy, and R. D. Bock, "Bilog-MG [computer software]," Scientific Software International, 2003.
- [44] R. J. De Ayala, B. Smith, and R. Norman Dvorak, "A comparative evaluation of Kernel equating and test characteristic curve equating," *Applied Psychological Measurement*, vol. 42, no. 2, pp. 155–168, Jun. 2018, doi: 10.1177/0146621617712245.
- [45] C. Liu and D. Jurich, "Application of sampling variance of item response theory parameter estimates in detecting outliers in common item equating," *Applied Psychological Measurement*, vol. 46, no. 6, pp. 529–547, Jun. 2022, doi: 10.1177/01466216221108122.
- [46] A. S. M. G. Allah, M. S. M. Ali, and I. S. A. Abdel-Wahed, "The effect of the difference in the equation method on equating two forms of achievement test according to item response theory," *Al-Azhar Journal of Education (AJED)*, vol. 40, no. 192, pp. 599–640, 2021.
- [47] Y. He and Z. Cui, "Evaluating robust scale transformation methods with multiple outlying common items under IRT true score equating," *Applied Psychological Measurement*, vol. 44, no. 4, pp. 296–310, Nov. 2020, doi: 10.1177/0146621619886050.
- [48] Z. Zhang, "Estimating standard errors of IRT true score equating coefficients using imputed item parameters," *Journal of Experimental Education*, vol. 90, no. 3, pp. 760–782, Apr. 2022, doi: 10.1080/00220973.2020.1751579.
- [49] Y. J. Choi and A. Asilkalkan, "R packages for item response theory analysis: descriptions and features," *Measurement: Interdisciplinary Research and Perspectives*, vol. 17, no. 3, pp. 168–175, Jul. 2019, doi: 10.1080/15366367.2019.1586404.
- [50] M. Battauz, "EquateIRT: An R package for IRT test equating," *Journal of Statistical Software*, vol. 68, no. 7, 2015, doi: 10.18637/jss.v068.i07.
- [51] C. Nisa and H. Retnawati, "Comparing the methods of vertical equating for the math learning achievement tests for junior high school students," *Research and Evaluation in Education*, vol. 4, no. 2, pp. 164–174, Dec. 2018, doi: 10.21831/reid.v4i2.19291.
- [52] M. A. Gunawan, H. Retnawati, and B. Kartowagiran, "Vertical equating accuracy using Kernel method," *Technium Science*, vol. 3, no. 7, pp. 110–120, 2021.
- [53] Z. Uzun and T. Öğretmen, "Test equating with the Rasch model to compare pre-test and post-test measurements," *Journal of Measurement and Evaluation in Education and Psychology*, vol. 12, no. 4, pp. 336–347, Dec. 2021, doi: 10.21031/epod.957614.
- [54] S. Kilmen and N. Demirtashi, "Comparison of test equating methods based on item response theory according to the sample size and ability distribution," *Procedia - Social and Behavioral Sciences*, vol. 46, pp. 130–134, 2012, doi: 10.1016/j.sbspro.2012.05.081.
- [55] İ. Uysal and S. Kilmen, "Comparison of item response theory test equating methods for mixed format tests," *International Online Journal of Educational Sciences*, vol. 8, no. 2, 2016, doi: 10.15345/iojes.2016.02.001.
- [56] M. Yurtçu and C. O. Güzeller, "Investigation of equating error in tests with differential item functioning," *International Journal of Assessment Tools in Education*, vol. 5, no. 1, pp. 50–57, Sep. 2017, doi: 10.21449/ijate.316420.
- [57] M. A. Barri and K. Abdul, "The impact of anchor item exposure on mean/sigma linking and IRT true score equating under the neat design," Doctoral Dissertation, University of Kansas, 2013.
- [58] S. H. Kim and A. S. Cohen, "A comparison of linking and concurrent calibration under item response theory," *Applied Psychological Measurement*, vol. 22, no. 2, pp. 131–143, Jun. 1998, doi: 10.1177/01466216980222003.
- [59] L. R. Aiken, "Content validity and reliability of single items or questionnaires," *Educational and Psychological Measurement*, vol. 40, no. 4, pp. 955–959, Dec. 1980, doi: 10.1177/001316448004000419.
- [60] M. J. Allen and W. M. Yen, *Introduction to measurement theory*. Cole Publishing Company, 1979.
- [61] J. C. Nunnally, *Psychometric theory 3E*. Tata McGraw-Hill education, 1994.
- [62] O. O. Adedoyin, "Using IRT approach to detect gender biased items in public examinations: a case study from the Botswana junior certificate examination in mathematics," *Educational Research and Reviews*, vol. 5, no. 7, pp. 385–399, 2010.
- [63] K. Hori, H. Fukuhara, and T. Yamada, "Item response theory and its applications in educational measurement part II: theory and practices of test equating in item response theory," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 14, no. 3, Dec. 2022, doi: 10.1002/wics.1543.
- [64] R. E. Millsap, "Testing measurement invariance using item response theory in longitudinal data: an introduction," *Child Development Perspectives*, vol. 4, no. 1, pp. 5–9, Mar. 2010, doi: 10.1111/j.1750-8606.2009.00109.x.
- [65] H. Retnawati, "The Comparison of Test Scores Linking Method Using Equating and Concordance," (in Indonesian), *Jurnal Kependidikan*, vol. 46, no. 2, pp. 164–178, 2016, doi: 10.21831/jk.v46i2.10383.
- [66] G. Skaggs, "Accuracy of random groups equating with very small samples," *Journal of Educational Measurement*, vol. 42, no. 4, pp. 309–330, Oct. 2005, doi: 10.1111/j.1745-3984.2005.00018.x.
- [67] S. Powers and M. J. Kolen, "Evaluating equating accuracy and assumptions for groups that differ in performance," *Journal of Educational Measurement*, vol. 51, no. 1, pp. 39–56, Mar. 2014, doi: 10.1111/jedm.12033.




**BIOGRAPHIES OF AUTHORS**

**Dian Normalitasari Purnama**    is a Ph.D. Candidate, Study Program of Educational Research and Evaluation, Graduate School, Universitas Negeri Yogyakarta, Yogyakarta, Indonesia and a Lecturer in Accounting Education Department, Faculty of Economics and Business, Universitas Negeri Yogyakarta, Yogyakarta, Indonesia. She completed her degree in Accounting Education in 2013 at Universitas Negeri Yogyakarta, Indonesia. She did her M.Pd (Master of Education) in Educational Research and Evaluation in 2016 at Universitas Negeri Yogyakarta, Indonesia. Her research areas of interest include accounting education, learning strategies, and learning assessment. She can be contacted at: dianpurnama@uny.ac.id.






**Samsul Hadi**    is a professor and lecturer at the Department of Electrical Engineering Education, Faculty of Engineering, Universitas Negeri Yogyakarta. He received a Ph.D. degree in research and evaluation education from the Universitas Negeri Yogyakarta. He is passionate about measuring, assessing, and evaluating vocational education. His research interests include students' instrument learning and development at various educational levels and areas. His publications cover topics such as vocational learning assessment and computer-based assessment development. He can be reached at email: samsul\_hd@uny.ac.id.






**Sukirno**    is a professor at the Accounting Education Department, Faculty of Economics and Business, Universitas Negeri Yogyakarta, Yogyakarta, Indonesia. He completed his degree in Accounting Education in 1993 at Universitas Negeri Yogyakarta, Indonesia. He did his M. Si (Master of Science) in accounting in 1999 at Universitas Gadjah Mada Indonesia. He obtained his Ph.D. in Business and Administration in 2005 at Asian Institute of Technology, Thailand. His research areas of interest are Accounting Research Methodology, Educational Research Methodology, Accounting information System and Accounting Education. He can be contacted at: sukirno@uny.ac.id.



**Heri Retnawati**    is a professor and lecturer at Universitas Negeri Yogyakarta, Indonesia. She was inaugurated as a professor in 2020. Her scientific field is Mathematics Education Evaluation. Her research interests are related to STEM, computational thinking, mathematics education, psychometrics, and 21st century thinking skills, thinking skills, and assessment for and of learning. She can be contacted at email: heri\_retnawati@uny.ac.id.



**Rizki Nor Amelia**    is a lecturer at the Integrated Sciences Department, Faculty of Mathematics and Natural Sciences, Universitas Negeri Semarang. She completed her degree in Chemistry Education (S.Pd) in 2014 and Master of Education (M.Pd) in 2016 at Universitas Negeri Yogyakarta, Indonesia. Her research areas of interest include chemistry education, item analysis, and learning assessment. She can be contacted at: rizkinoramelia@mail.unnes.ac.id.