Thabiea : Journal of Natural Science Teaching
Vol. 6(1), pp. 51-64, 2023
Available online at
http://journal.iainkudus.ac.id/index.php/Thabiea
p-issn: 2580-8474, e-issn: 2655-898X

Equating of standardized science subjects tests using various methods: which is the most profitable?

Muh. Asriadi AM[1]*, Heri Retnawati[2]
[1]Departement of Early Childhood Education, Universitas Pendidikan Indonesia, Pendidikan Street No.15, Jawa Barat, 40625, Indonesia
[2]Departement of Educational Research and Evaluation, Yogyakarta State University, Colombo Street, Sleman, 55281, Indonesia
**Correspondence: asriadi190197@gmai.com*

Keywords:
Equating;
Standardized Tests;
Science Subjects;
Item Response Theory.

Abstract

A good test set can be reflected in the quality of the items. It can measure the ability of the test takers reasonably even though they are distributed in several question packages. This study uses an exploratory, descriptive method to determine the equivalence of standardized test sets in science subjects for junior high schools in Indonesia. The data were obtained from the database of Junior High School National Examination results in the subject of Natural Sciences, which consisted of 5 question packages with 40 items/package. The equating technique uses the Item Response Theory 3 PL approach with the help of R Studio Software. The research results show that the national exam questions, which consist of 5 question packages, have a good level of item difficulty and all guesses. However, the discrimination index and several items obtained unfavorable results. In addition to the results of equating the graphical method using the closeness of the test characteristic curve, the Stocking & Lord methods produce the most equivalent scores. These findings can be a reference for test developers or researchers in the field of measurement to produce better and more accurate test kits.

To cite this article:
Asriadi AM. M., Retnawati. H. (2023). Equating of standardized science subjects tests using various methods: which is the most profitable? *Thabiea : Journal of Natural Science Teaching*, 6(1), 51-64.

Introduction

The National Examination is one of the standardized tests used in all areas and is still used in Indonesia. This test is used to assess graduates' competency levels across the country. National exam results are used as one of the factors in mapping the quality of academic units, determining student graduation from academic units, and selecting students to enter the next educational level (Herkusumo, 2011). If all test takers in each province work on the same questions, these values can be compared (test set). In practice, however, the National Examination administers more than one test kit in each province and employs different test kits across provinces (Aminah, 2013). The difference in scores between test takers who received different tests cannot be directly concluded that there are differences in ability by administering more than one test set because the difficulty level of the device used will affect these differences.

To address this issue, educational measurement experts created a statistical method known as equating (Baker & Al-Karni, 1991). This method is a scientific method for equating

values from one device's raw score to another device's raw score, resulting in a score conversion table (Peabody, 2020). According to Livingston (2014), score equalization is an empirical procedure because score data is obtained from students' work and then used to transform scores. According to Hambleton et al. (1991), score equalization is the process of comparing scores obtained from one test set (X) and another test set (Y) by equalizing scores on the two test sets. According to Skaggs & Lissitz (1986), the two scores of measurement results obtained with instrument X and instrument Y can be equalized if the two instruments measure the same ability or trait. According to Zhu (1998), score equalization is possible if the test-taker groups are equal because extreme equality will affect the calculation. Equating is a psychometric process that aims to produce a conversion score that can be used to compare the results of multiple parallel test sets (Diao & Keller, 2020). Based on the preceding, it can be concluded that equating is an empirical procedure used to equalize scores from one test set to another to make direct comparisons or conversions based on the results of individuals taking the various test sets.

It is necessary to equalize the test devices used in implementing an evaluation that uses several different test sets and measures the same thing because this equalization ensures justice for test takers. Equalizing scores can be used as a score equalization technique to distinguish bright students from less intelligent students (Akin-Arikan & Gelbal, 2021). Score equalization enables using different test sets for groups based on their ability level. So that the scores obtained can be compared and test takers do not feel disadvantaged or benefited because they received an easier or more difficult test set (Zhang, 2020). The main goal of equating is to ensure fairness for test takers and users of test results (S. Y. Kim, 2022). Because it is assumed that a parallel test in terms of material (derived from the same grid) is incorrect, a process that equalizes the scores of parallel test sets by eliminating the factor of differences in difficulty levels between these devices is required (Furter & Dwyer, 2020). Equating is a procedure for scoring test takers based on their abilities by removing the effect of test device difficulty differences (W. J. van. der Linden, 2022). It follows the demands of justice; do not allow students to receive poor grades. They worked on complicated test sets or students to score well because they took easy tests (Zhang, 2022). Users of test results also demand the validity of the results, lest someone get good results simply because they take more accessible tests despite their low abilities (Hadi et al., 2022). As a result, test developers or test development institutions must equalize the test devices used.

There are several methods for correlating the results of two or more tests. The method of correlating test scores is classified into two methods in terms of calibration technique, namely the particular calibration method and the simultaneous calibration method (Goodman et al., 2020). The two tests are calibrated separately in the different calibration methods, whereas the two are calibrated simultaneously or together in the simultaneous calibration method (Wiberg, 2021). The equalization constant is not calculated during simultaneous calibration. The calibration results of the two tests show that the item parameters and abilities are on the same scale (Lu & Kim, 2021). Two methods are included in the separate calibration method: the moment method and the graphic method (Bramley, 2020). At least three methods can be used for the moment method, namely the Mean - Sigma method and the Mean-Maen method(Uysal et al., 2022). Two methods can be used in the graphical method: the characteristic curve method from Haebara and the characteristic curve method from Stocking

& Lord. These widely used methods produce nearly identical equalization results (Supriyati et al., 2021). The mean-mean, mean-sigma, Haebara, and Stocking and Lord methods will be used in this study for test equalization.

In Indonesia, several researchers have conducted studies on the process of equating using national examination data. For instance, Retnawati (2016) examined the score equivalence of 20 sets of final exam tests at the junior high school level in Indonesia. The study compared the results of equating designs with and without shared items. While these findings were significant, further studies using different data are needed to validate these results. Another study by Kartowagiran et al. (2018) focused on investigating the equivalence of national mathematics exam test kits for junior high school level in Indonesia. The analysis covered test kits from 2013 to 2016, aiming to determine the equivalence of test packages across different years and how the test kits were comparable between those years. However, this study did not explore the equivalence of scores among test takers on the test packages. In a study by Yusron et al. (2020b) the equalization of the national standardized school exam test package in mathematics at the senior high school level was investigated. This study not only described the equivalence of the test packages but also compared the results of four equalization methods based on the 3PL item response theory. It is worth noting that the test device used in this study was not considered high-stakes. Overall, these studies contribute to understanding the equating procedures and practices in the Indonesian context using national examination data. However, this research will expand the study related to equating by using a more diverse method.

The equating process is necessary in the processing of national exam results in order to obtain an accurate and valid mapping of educational quality without distortion of differences in difficulty levels despite receiving different test kits (Li & Kapoor, 2022). The problem of equating tests in the National Examination in Indonesia needs to be addressed, given the uneven distribution of education in Indonesia's territory and the geographical conditions of Indonesia's territory as an archipelagic country (Sutari, 2017). It is necessary to equalize the score when evaluating the level of the National Examination, which uses several different test sets and measures the same thing. Through the process of standardizing scores on the national examination test device, it becomes possible to estimate minor measurement errors and subsequently compare the academic achievements of students hailing from diverse provinces (Rosidin et al., 2019). Test takers do not feel disadvantaged or benefited based on whether they received an easier or more difficult test package.

Looking at previous literature, the study of equivalence applies to other types of tests, especially tests of natural sciences. Our study attempts to fill a gap in the literature regarding the equivalence procedure applied to natural science tests for junior high school level. This study is also expected to be able to provide guidance for researchers and test developers in developing equivalent test packages. Thus, this study aims to describe and compare the equivalence of natural science test packages at the junior high school level using four methods: mean-mean, mean-sigma, Haebara, and Stocking-Lord.

Method

Study Design

This research is exploratory descriptive (Johnson & Christensen, 2017), focuses on describing the equivalence of the five question packages used in the national exam for science subjects using several methods and comparing the results. Our equalization uses the equivalent group design. Five packages of questions to be equalized were given to five groups of equivalent test takers who were randomly selected from the population. In this case, the five groups of test takers are considered to have the same level of ability. We used an item response theory approach and four methods to relate test scores: mean-mean, mean-sigma, Stocking-Lord, and Haebara. We set the 1st question package as the master package and the test scores of the other question packages will be linked to the master package test scores.

Study Participants

The data for this study were taken from the 2015 Junior High School national exam results database in science subjects. The test takers totaled 42,147 people. The national exam test consists of 5 packages containing 40 questions. The number of test takers who worked on each question package was Package 1 as many as 8,824 people, Package 2 as many as 8430 people, Package 3 as many as 8355, Package 4 as many as 8186, Package 5 as many as 7873 people. Determination of the package of questions that students will work on is done randomly by the computer. All students take the test at the same time and under strict supervision.

Characteristics of the Test Device

We will assess the comparability of five test packages from the national exam in science subjects at the junior high school level. This examination holds significant importance as it determines participants' eligibility to progress to the next educational level and serves as a means for policy makers to evaluate national educational achievements. The test comprises five question packages: Package 1, Package 2, Package 3, Package 4, and Package 5. All packages were developed using a consistent framework. The test construction incorporates two domains: cognitive level and content. The cognitive level domain encompasses three levels: knowledge and understanding, application, and reasoning. The content domain covers topics such as measurements, matter and its properties, mechanics, and the solar system, as well as waves, electricity, and magnetism. Each test consists of 40 multiple-choice items, with each item offering four answer choices and only one correct answer. Participants are provided with 120 minutes to complete the test. The scoring system follows a no penalty approach, meaning that incorrect answers do not result in a score deduction.

Data Collection

Our study uses data from the responses of junior high school students after they take the national exam in science subjects. These response data were documented by the Center for Research and Education (now the Center for Education Assessment), a special institution authorized by the Ministry of Education, Culture and Higher Education of the Republic of Indonesia to process all national exam data. With permission from this institution, we were provided with a data set of student responses to the science subject. The data set we received consisted of participant ID, province code and question packets, and student responses to the 40 test items. We received student responses in the form of a dichotomy: 1 represents the correct answer and 0 represents the wrong answer. Next, we partitioned the data based on the

question packet code, so that five data sets were obtained separately. We label this separate data set Package 1 to Package 5 referring to the available question package codes.

Data Analysis

The equating analysis technique is based on the Item Response Theory 3 PL (Parameter Logistics). The equating design used is the Equivalent Group Design. In the equivalent group design, the two tests to be equalized are given to two equivalent (non-identical) groups randomly selected from the same population, and the two groups are considered to have the same ability level. In item response theory, if the item response model fits a data set, then the distribution of the linear transformation of the measurement scale is also suitable for that data. It means a relationship exists between the measurement scales of the two tests. Thus, if the test scale one is equated with the test scale 2 for the 3PL model, then the relationship between the item parameters and the ability of the participants for the two scales is (Kolen & Brennan, 2014):

$$\theta_{1i}^{*} = \alpha\theta_{1i} + \beta \quad (1)$$

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

$$b_{1j}^{*} = \alpha b_{1j} + \beta \quad (3)$$

$$c_{1j}^{*} = c_{1j} \quad (4)$$

Information:

a1j, b1j, and c1j are item parameters for item j on a test scale of 1.

a1j*, b1j*, and c1j* are item parameters for item j on test scale 1 after being equated with test 2.

θ_{1i} is the ability of participant i on a test scale of 1.

θ_{1i}^{*} is the ability of participant i on test scale 1 after being equated with test 2.

α and β are equalization constants.

The guess parameter c is not transformed because its value does not depend on the metric θ or c is independent of the constant scale transformation. Furthermore, correlating test scores α and β can be calculated using methods connecting test scores. The equating methods used are the Mean & Mean method, the Mean & Sigma method, the Haebara method, and the Stocking & Lord method. The question packages will be equated by making package 1 the default account. The test characteristic curves for all devices are then drawn in one image area to determine equivalence after equalization for each method, both with and without shared items. Students' ability to use equalization constants with shared items and shared items is then calculated. All data were analyzed using R Studio software.

Results and Discussion

Items Parameter Analysis with Item Response Theory

Item parameters were estimated using a 3-PL model. We estimated the discriminant index (a_i), difficulty index (b_i), and pseudo-guessing (c_i) parameters for each test item. Descriptive statistics of the estimated item parameters for the five test packages are presented in Table 1. Overall, the five test packages are relatively equivalent, as shown by the overlapping test characteristic curves (TCCs) in Figure 1. However, statistically, the item parameters for the five test packages are not the same. Table 1 shows that the average difficulty level for item P3 is the highest (0.472) and P5 is the lowest (-0.187), indicating that

P3 is the most difficult compared to the other four test packages, while P5 is the easiest. In terms of item discriminant index, Table 1 shows that P1 is the highest (2.403) and P2 is the lowest (1.919) compared to the other four test packages. Meanwhile, in terms of pseudo-guessing parameter, P2 is the highest (0.289) and P4 is the lowest (0.270) compared to the other four test packages.

Table 1. Descriptive Statistics for Test Item Parameters P1, P2, P3, P4, and P5

	P1			**P2**			**P3**		
	a_i	b_i	c_i	a_i	b_i	c_i	a_i	b_i	c_i
M	2.403	0.148	0.2869	1.919	0.215	0.289	2.051	0.472	0.282
SD	0.950	0.944	0.1114	1.555	0.968	0.158	1.178	1.131	0.195
Min.	-0.152	-2.867	0.0003	-6.200	-2.19	0.004	0.21	-1.133	0.005
Max.	4.743	1.756	0.4697	4.317	3.300	0.794	4.856	6.282	0.722

	P4			**P5**		
	a_i	b_i	c_i	a_i	b_i	c_i
M	2.015	0.062	0.270	2.063	-0.187	0.277
SD	0.887	0.986	0.149	1.611	1.462	0.130
Min.	0.141	-3.500	0.002	-5.929	-7.145	0.003
Max.	3.684	1.369	0.723	4.11	2.411	0.571

Table 1 also shows that the most difficult item is found in P3 (b_i = 6.282) and the easiest item is found in P4 (b_i = -1.369). The item with the highest discriminant index is found in P3 (a_i = 4.856) and the lowest is found in item P4 (a_i = 3.684). In addition, the item with the highest pseudo-guessing parameter is found in P2 (c_i = 0.794) and the lowest is found in item P1 (c_i = 0.0003). Although the TCCs show that the five test packages have relatively equivalent characteristics (see Figure 1), the differences in item parameters have the potential to produce unequal ability scores (θ) across test packages.

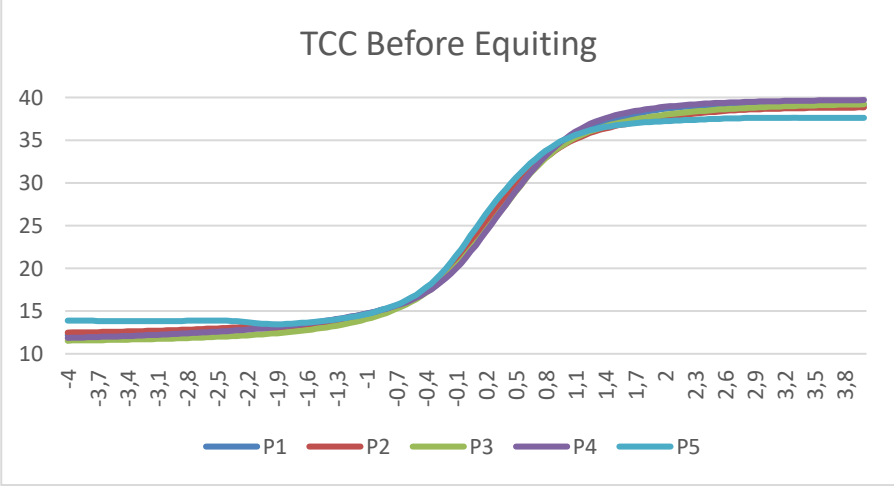

Figure 1. TCC five test packages before equalization

Equating Results with Four Methods

Test P1 is the anchor test, so the four other tests (P2, P3, P4, and P5) were equated to P1. The equating process resulted in four constants α and four constants β for each equating method. These constants are presented in Table 2. Equating P2 to P1 produced the highest α

constant in the mean-sigma method and the lowest in the mean-mean method. Meanwhile, the highest β constant was produced by the Haebara method and the lowest by the mean-sigma method. Equating P3 to P1 produced the highest α constant in the Haebara method and the lowest in the mean-sigma method. Meanwhile, the highest β constant was produced by the Stocking-Lord method and the lowest by the mean-mean method. Equating P4 to P1 produced the highest α constant in the mean-sigma method and the lowest in the Haebara method. Meanwhile, the highest β constant was produced by the mean-mean method and the lowest by the Stocking-Lord method. Finally, equating P5 to P1 produced the highest α constant in the Haebara method and the lowest in the mean-sigma method. Meanwhile, the highest β constant was produced by the mean-mean method and the lowest by the Stocking-Lord method.

Table 2. Test Equalization Constants

Link	Constant	Mean-Mean	Mean-Sigma	Haebara	Stocking-Lord
P2 to P1	α	0.798	0.976	0.972	0.893
	β	-0.023	-0.061	0.072	0.041
P3 to P1	α	0.853	0.835	0.927	0.894
	β	-0.254	-0.245	-0.038	-0,004
P4 to P1	α	0.839	0.958	0.839	0.938
	β	0.096	0.089	0.014	-0.008
P5 to P1	α	0.858	0.646	1.002	0.945
	β	0.308	0.269	0.093	0.079

Table 2 shows that the alpha constants generated by the mean-mean method vary considerably, ranging from 0.798 to 0.858. The beta constants also show the same pattern, ranging from -0.254 to 0.308. Equating using the mean-sigma method produces alpha constants in the range of 0.646 to 0.976 and beta constants in the range of -0.245 to 0.269. Equating using the Haebara method produces alpha constants in the range of 0.869 to 1.002 and beta constants in the range of -0.038 to 0.093. Equating using the Stocking-Lord method produces alpha constants in the range of 0.893 to 0.945 and beta constants in the range of -0.008 to 0.079.

These equating constants will be used to convert the original scores of participants to new scores after equating the tests. Table 3 presents the score equating equations (θ) for each method. These equations are used to convert the θ of participants who took the initial tests (P2, P3, P4, and P5) to the new θ after equating the tests to P1. We present the results of this conversion in Table 3.

Table 3. Equivalence of Score Equalization (θ)

Link	Mean-Mean	Mean-Sigma	Haebara	Stocking-Lord
P2 to P1	$\theta_{21} = 0.798\theta_2 - 0.023$	$\theta_{21} = 0.976\theta_2 - 0.061$	$\theta_{21} = 0.972\theta_2 + 0.072$	$\theta_{21} = 0.893\theta_2 + 0.041$
P3 to P1	$\theta_{31} = 0.853\theta_3 - 0.254$	$\theta_{31} = 0.835\theta_3 - 0.245$	$\theta_{31} = 0.927\theta_3 - 0.038$	$\theta_{31} = 0.894\theta_3 - 0.004$
P4 to P1	$\theta_{41} = 0.839\theta_4 + 0.096$	$\theta_{41} = 0.958\theta_4 + 0.089$	$\theta_{41} = 0.839\theta_4 + 0.014$	$\theta_{41} = 0.938\theta_4 - 0.008$
P5 to P1	$\theta_{51} = 0.858\theta_5 + 0.308$	$\theta_{51} = 0.646\theta_5 + 0.269$	$\theta_{51} = 1.002\theta_5 + 0.093$	$\theta_{51} = 0.945\theta_5 + 0.079$

Note: θ_{21}, θ_{31}, θ_{41}, and θ_{51} represent the converted ability scores of participants in taking tests P2, P3, P4, and P5, respectively, after they are equated to test P1; θ_2, θ_3, θ_4, dan θ_5 represent the ability scores of participants in taking tests P1, P2, P3, and P4 (before equating).

Table 4 shows the conversion results of several θ after being adjusted to the master test (P1). Using the mean-mean method, when the participant obtains θ = -4 on the P2 test, it is equivalent to θ = 0.292 after P2 is adjusted to P1. When P3, P4, and P5 are adjusted to P1 using the mean-mean method, θ = -4 will be equivalent to 0.287, 0.291, and 0.348, respectively. Thus, at θ = -4, equalizing the four test packages to P1 using the mean-mean method will produce a new θ in the range 0.287 to 0.348. At θ = 0, equalization will produce a new θ in the range 0.453 to 0.647. Whereas at θ = 4, equalization will produce a new θ in the range 0.939 to 0.994. This indicates that the variation in scores after being equalized using the mean-mean method is quite high. So less profitable. This is because the difference in the value of the item parameters has the potential to produce an unequal ability score (θ) between test packages.

Table 4. Results of Equalizing Scores for Certain θ

	Mean-Mean				**Mean-Sigma**			
Ability(θ)	**P2 to P1**	**P3 to P1**	**P4 to P1**	**P5 to P1**	**P2 to P1**	**P3 to P1**	**P4 to P1**	**P5 to P1**
-4	0.292	0.287	0.291	0.348	0.294	0.287	0.295	0.347
-3	0.296	0.293	0.299	0.347	0.301	0.292	0.304	0.348
-2	0.308	0.307	0.315	0.347	0.316	0.306	0.321	0.346
-1	0.344	0.362	0.351	0.343	0.363	0.359	0.36	0.337
0	0.578	0.647	0.495	0.453	0.591	0.646	0.502	0.439
1	0.914	0.921	0.896	0.845	0.891	0.922	0.869	0.894
2	0.975	0.966	0.979	0.930	0.962	0.967	0.972	0.939
3	0.987	0.979	0.991	0.941	0.984	0.979	0.989	0.939
4	0.989	0.985	0.994	0.939	0.988	0.985	0.993	0.938
	Haebara				**Stocking-Lord**			
Ability(θ)	**P2 to P1**	**P3 to P1**	**P4 to P1**	**P5 to P1**	**P2 to P1**	**P3 to P1**	**P4 to P1**	**P5 to P1**
-4	0.311	0.288	0.291	0.347	0.311	0.287	0.295	0.347
-3	0.317	0.293	0.300	0.346	0.315	0.292	0.305	0.346
-2	0.328	0.306	0.317	0.339	0.327	0.304	0.323	0.343
-1	0.351	0.348	0.356	0.361	0.347	0.343	0.364	0.358
0	0.534	0.552	0.526	0.538	0.544	0.539	0.535	0.542
1	0.855	0.885	0.912	0.864	0.875	0.886	0.895	0.875
2	0.941	0.956	0.981	0.929	0.949	0.957	0.976	0.931
3	0.966	0.975	0.991	0.940	0.969	0.976	0.989	0.941
4	0.971	0.982	0.994	0.940	0.972	0.983	0.993	0.939

Table 4 also shows that the variation of θ after being adjusted using the mean-sigma method is quite high. For example, at θ = -4, equalizing the four test packages to P1 using the mean-sigma method yields a new θ in the range 0.287 to 0.347. At θ = 0, equalization will produce a new θ in the range 0.439 to 0.646. Whereas at θ = 4, equalization will produce a new θ in the range 0.938 to 0.993.

Equalization using the Haebara method shows that the variation of θ after being equalized is not too varied (see Table 4). For example, at θ = -4, equalizing the four test packets to P1 using this method yields a new θ in the range 0.288 to -0.347. At θ = 0, equalization will produce a new θ in the range 0.526 to 0.552. Whereas at θ = 4, equalization will produce a new θ in the range 0.940 to 0.994.

Equalization using the Stocking-Lord method shows that the new θ after being equalized also does not vary too much (see Table 4). For example, at θ = -4, equalizing the

four test packages to P1 using the Stocking-Lord method only produces a new θ in the range 0.287 to 0.347. At θ = 0, equalization will only produce a new θ in the range 0.535 to 0.544. Whereas at θ = 4, equalization will produce a new θ in the range 0.939 to 0.993. This interval is much shorter than the previous three methods. This shows that equalizing using the Stocking-Lord method produces a conversion score that is more equal than the previous three methods, so it is more profitable. This is because there is no difference in the values of the parameters that have the potential to produce equivalent participant ability scores (θ) between test packages.

Furthermore, it is also reported the test characteristic curve after equalization to strengthen the previous findings. Figure 2 presents the TCC of the four tests after being adjusted to P1 using the mean-mean, mean-sigma, Haebara, and Stocking-Lord methods. The method that produces the most equivalent score is shown by the most closely matched TCC. In Figure 7 the Stocking-Lord method produces the most closely aligned TCC compared to the other methods. This reinforces previous findings that the Stocking-Lord method produces the most equivalent scores compared to the mean-mean, mean-sigma, and Haebara methods.

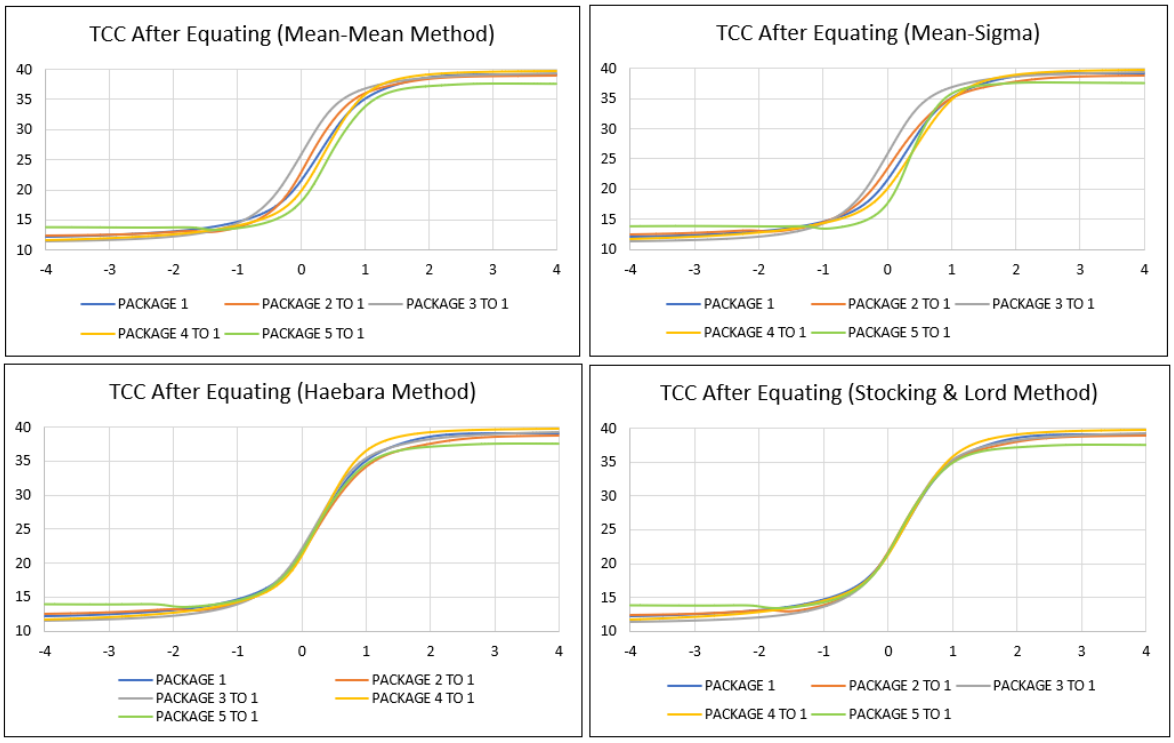

Figure 2. Test Characteristic Curve Equating for All Equating Methods

The natural science test used in the national junior high school exam is classified as a standardized test. This test has significant consequences for students because it determines their graduation status to be able to continue to the next level of education. The tests we used in this study consisted of five test packages. This study reveals that the five test packages in a standardized test are relatively equivalent. This finding is not surprising given that the test was developed following a strict procedure. Several other studies (e.g.,Kartowagiran et al., 2018; Retnawati, 2016) have succeeded in revealing that the test packages used in the national standardized national exams and school exams at various levels in Indonesia are proven to be

equivalent. Thus, the findings of our study are consistent with the findings of previous studies while at the same time confirming that the standardized natural science tests used in Indonesia are of fairly good quality.

Hambleton et al., (1991) have affirmed that when a test consists of more than one test booklet, the test booklets are not truly equivalent, even though they are arranged based on the same blueprint. This suggests that an equating procedure is necessary so that measurement scores from different test booklets can be compared and meet the fairness standards for test takers. Following this advice, this study used equating procedures using four popular methods: mean-mean, mean-sigma, Haebara, and Stocking-Lord. These four methods have been widely applied by other researchers (Retnawati et al., 2017; Yurtcu & Güzeller, 2017). However, not all findings from these studies are consistent with our study's findings.

Of the four equalization methods used, it was found that equalization using the Stocking-Lord method produced the most equivalent scores between packages. These findings are the same as those of previous studies (Uysal & Kilmen, 2016) but different from the findings of other studies (Yusron et al., 2020a). This indicates that the literature has not been able to provide strong evidence that one method is better than the other equalization methods. The difference in findings between studies is very likely due to differences in several things, such as the characteristics of the test takers and the types of test items. For example, the study by (Yusron et al., 2020a) used a test kit for high school level students. This study is clearly different from our study, in which we used a test for junior high school students. Nisa and Retnawati's study (2018) used a test kit consisting of multiple-choice items and descriptions, while our study only used multiple choice type items. Differences in findings due to differences in test taker characteristics and item types provide opportunities for other researchers to explore the topic of test equity by considering this issue.

Based on findings (see Figure 2), the characteristic curve methods (Haebara and Stocking-Lord) produce more equated scores compared to methods that involve item parameter estimation (mean-mean and mean-sigma methods). The mean and sigma methods tend to involve estimating item parameters, such as the level of difficulty and item differentiability. When there are significant differences in the degree of difficulty or discriminatory power between the item packages, the averaging and sigma methods will not cope well with these differences. If item parameter estimates are inaccurate or inconsistent between the question packages to be equalized, this method may produce scores that are less equal or less accurate. These findings are consistent with the findings of (Cohen, 1998; S. Kim & Kolen, 2006). However, other studies have shown different findings, where the characteristic curve method does not always produce more equated scores, as reported in the studies of (Nisa & Retnawati, 2018). Therefore, this issue is also interesting to be investigated by other researchers. This is necessary to enrich the literature on the topic of equating, which focuses on this issue.

The findings of this study are important for filling gaps in the literature related to equalization procedures in high-stakes testing, especially in natural science subjects for junior high school level. The findings of our study are important for policy makers regarding educational assessment to guide them in applying equalization procedures in implementing large-scale assessments at various levels of education. Our findings also guide researchers and practitioners to choose the equivalence method that is most beneficial when they apply it to

testing processes in various contexts. Our study has also uncovered important new issues on the topic of test equivalence and is of interest to be investigated by other researchers in the future. Finally, our study also provides insight to education stakeholders about the procedure for measuring learning outcomes that are fair for all test takers so that no one feels disadvantaged, especially when measurements are carried out using standardized tests.

Conclusion

This study reveals that the five test packages in the 2015 National Science Examination for Junior High School level have relatively the same grain characteristics. However, the item parameters (discriminant index, difficulty index, and pseudo-guessing) of the test packages are not exactly the same, so an equalization procedure is necessary. Equalization results using four methods (mean-mean, mean-sigma, Haebara, and Stocking-Lord) produce various equalization constants. Our study concludes that equalization using the Stocking-Lord method produces the most equivalent scores compared to other equalization models. This indicates that the Stocking-Lord method produces the most comparable scores when the test package is equalized to the main package. The four equalization methods provide information that the increase in students' abilities from one level to another is not significant, so it is likely only caused by age factors and learning experience.

This study has a weakness, namely the equalization procedure does not involve anchor items. The authorities did not provide information to researchers regarding which items were determined as anchors in the natural science test packages used in the national exam. It is important to note that the absence of anchor items in a study can be considered a limitation. Anchor items are specific test items that are included in multiple forms or versions of a test. They serve as common reference points or links between different forms, allowing for the equating of test scores across forms. Anchor items are carefully selected to ensure that they measure the same construct and have consistent difficulty and discrimination parameters across different forms. Without anchor items, the equating procedure relies solely on statistical methods and assumptions, which may introduce additional uncertainty in the equating process and potentially affect the accuracy of the equated scores. Therefore, the inclusion of anchor items is generally recommended to enhance the validity and reliability of the equating results.

However, these limitations do not diminish the importance of our study findings. In the future we recommend that researchers investigate the use of anchor items to equalize scores on standardized science tests. This is important for enriching the literature on equivalence procedures on standardized natural science tests. Related to this, there are two interesting topics to be investigated in future studies. First, how are the results of standardized natural science tests equivalent when using and without using anchor items. Second, what is the effect of the number of anchor items on the equivalent results of standardized natural science tests. These two issues are important to be explored in more depth to improve the quality of measuring and testing mathematics learning outcomes in the future.

Credit Authorship Contribution Statement

Muh Asriadi AM: Conceptualization, Methodology, Software, Visualization, Formal analysis, Writing – original draft, Writing – review & editing. **Heri Retnawati:**

Conceptualization, Review - original draft, and Supervision.

References

Akin-Arikan, Ç., & Gelbal, S. (2021). A comparison of kernel equating and item response theory equating methods. *Eurasian Journal of Educational Research*, *21*(93), 179–198. https://doi.org/10.14689/ejer.2021.93.9

Aminah, N. S. (2013). Karakteristik metode penyetaraan skor tes untuk data dikotomos. *Jurnal Penelitian Dan Evaluasi Pendidikan*, *16*, 88–101. https://doi.org/10.21831/pep.v16i0.1107

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

Bramley, T. (2020). Comparing small-sample equating with angoff judgement for linking cut-scores on two tests. *Research Matters, 2017*, 23–27.

Cohen, A. S. (1998). An investigation of linking methods under the graded response model. *Applied Psychological Measurement*, *22*(2), 116–130. https://doi.org/10.1177/01466216980222002

der Linden, W. J. va. (2022). What is actually equated in "test equating"? A didactic note. *Journal of Educational and Behavioral Statistics*, *47*(3), 353–362. https://doi.org/10.3102/10769986211072308

Diao, H., & Keller, L. (2020). Investigating repeater effects on small sample equating: Include or exclude? *Applied Measurement in Education*, *33*(1), 54–66. https://doi.org/10.1080/08957347.2019.1674302

Furter, R. T., & Dwyer, A. C. (2020). Investigating the classification accuracy of rasch and nominal weights mean equating with very small samples. *Applied Measurement in Education*, *33*(1), 44–53. https://doi.org/10.1080/08957347.2019.1674307

Goodman, J. T., Dallas, A. D., & Fan, F. (2020). Equating with small and unbalanced samples. *Applied Measurement in Education*, *33*(1), 34–43. https://doi.org/10.1080/08957347.2019.1674311

Hadi, S., Haryanto, H., AM, M. A., Marlina, M., & Rahim, A. (2022). Developing classroom assessment tool using learning management system-based computerized adaptive test in vocational high schools. *Journal of Education Research and Evaluation*, *6*(1), 143–155. https://doi.org/10.23887/jere.v6i1.35630

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). Fundamentals of item response theory. In *SAGE Publications, Inc.* (Vol. 29, Issue 07). https://doi.org/10.5860/choice.29-4185

Herkusumo, A. P. (2011). Penyetaraan (Equating) Ujian Akhir Sekolah Berstandar Nasional (UASBN) Dengan Teori Tes Klasik. *Jurnal Pendidikan Dan Kebudayaan*, *17*(4), 455–471. https://doi.org/10.24832/jpnk.v17i4.41

Johnson, R. B., & Christensen, L. (2017). Educational research: Quantitative, qualitative, and mixed approaches. In *SAGE Publications, Inc.*

Kartowagiran, B., Munadi, S., Retnawati, H., & Apino, E. (2018). The equating of battery test packages of mathematics national examination 2013-2016. *SHS Web of Conferences*, *42*(January), 00022. https://doi.org/10.1051/shsconf/20184200022

Kim, S., & Kolen, M. J. (2006). Robustness to format effects of IRT linking methods for mixed-format tests. *Applied Measurement in Education*, *19*(4), 357–381. https://doi.org/10.1207/s15324818ame1904

Kim, S. Y. (2022). Digital module 29: Multidimensional item response theory equating. *Educational Measurement: Issues and Practice*, *41*(3), 85–86. https://doi.org/10.1111/emip.12525

Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking: Methods and practices (statistics for social science and public policy)* (3rd ed.). Springer Science+Business Media, LLC.

Li, D., & Kapoor, S. (2022). Evaluating population invariance of test equating during the COVID-19 pandemic. *Education Measurement*, *41*(1), 33–41.

Livingston, S. A. (2014). Equating test scores (without IRT). In *ETS Report* (Second). Educational Testing Service. papers3://publication/uuid/753FF7E7-6A9F-4F37-99FA-5FC927542973

Lu, R., & Kim, S. (2021). Effect of statistically matching equating samples for common-item equating. *ETS Research Report Series*, *2021*(1), 1–14. https://doi.org/10.1002/ets2.12313

Nisa, C., & Retnawati, H. (2018). Comparing the methods of vertical equating for the math learning achievement tests for junior high school students. *Research and Evaluation in Education*, *4*(2), 164–174. https://doi.org/10.21831/reid.v4i2.19291

Peabody, M. R. (2020). Some methods and evaluation for linking and equating with small samples. *Applied Measurement in Education*, *33*(1), 3–9. https://doi.org/10.1080/08957347.2019.1674304

Retnawati, H. (2016). Perbandingan metode penyetaraan skor tes menggunakan butir bersama dan tanpa butir bersama. *Jurnal Kependidikan: Penelitian Inovasi Pembelajaran*, *46*(2), 164–179. https://doi.org/10.21831/jk.v46i2.10383

Retnawati, H., Kartowagiran, B., Arlinwibowo, J., & Sulistyaningsih, E. (2017). Why are the mathematics national examination items difficult and what is teachers' strategy to overcome it? *International Journal of Instruction*, *10*(3), 257–276. https://doi.org/10.12973/iji.2017.10317a

Rosidin, U., Herpratiwi, Suana, W., & Firdaos, R. (2019). Evaluation of national examination (UN) and national-based school examination (USBN) in Indonesia. *European Journal of Educational Research*, *8*(3), 827–837. https://doi.org/10.12973/eu-jer.8.3.827

Skaggs, G., & Lissitz, R. W. (1986). IRT test equating: Relevant issues and a review of recent research. *Review of Educational Research*, *56*(4), 495–529. https://doi.org/10.3102/00346543056004495

Supriyati, Y., Iriyadi, D., & Falani, I. (2021). The development of equating application for computer based test in physics hots category. *Journal of Technology and Science Education*, *11*(1), 117–128. https://doi.org/10.3926/jotse.1135

Sutari, V. R. (2017). National examination in Indonesia and its backwash effects: Teachers' perspectives. *Ninth International Conference on Applied Linguistics (CONAPLIN 9)*, *82*(Conaplin 9), 331–333. https://doi.org/10.2991/conaplin-16.2017.76

Uysal, İ., & Kilmen, S. (2016). Comparison of item response theory test equating methods for mixed format tests. *International Online Journal of Educational Sciences*, *8*(2). https://doi.org/10.15345/iojes.2016.02.001

Uysal, İ., Şahin-Kürşad, M., & Kılıç, A. F. (2022). Effect of item parameter drift in mixed format common items on test equating. *Participatory Educational Research*, *9*(5), 143–160. https://doi.org/10.17275/per.22.108.9.5

Wiberg, M. (2021). Practical assessment, research, and evaluation on the use of different linkage plans with different observed-score equipercentile equating methods. *Practical Assessment, Research & Evaluation*, *26*(23), 1–18.

Yurtcu, M., & Güzeller, C. O. (2017). Investigation of equating error in tests with differential item functioning. *International Journal of Assessment Tools in Education*, *January*, 50–57. https://doi.org/10.21449/ijate.316420

Yusron, E., Retnawati, H., & Rafi, I. (2020a). Bagaimana hasil penyetaraan paket tes USBN pada mata pelajaran matematika dengan teori respon butir? *Jurnal Riset Pendidikan Matematika*, *7*(1), 1–12. https://doi.org/10.21831/jrpm.v7i1.31221

Yusron, E., Retnawati, H., & Rafi, I. (2020b). Bagaimana hasil penyetaraan paket tes USBN pada mata pelajaran matematika dengan teori respon butir? [What are the results of equating the USBN test package in mathematics with item response theory?]. *Jurnal Riset Pendidikan Matematika*, *7*(1), 1–12. https://doi.org/10.21831/jrpm.v7i1.31221

Zhang, Z. (2020). Asymptotic standard errors of equating coefficients using the characteristic curve methods for the graded response model. *Applied Measurement in Education*, *33*(4), 309–330. https://doi.org/10.1080/08957347.2020.1789142

Zhang, Z. (2022). Estimating standard errors of IRT true score equating coefficients using imputed item parameters. *Journal of Experimental Education*, *90*(3), 760–782. https://doi.org/10.1080/00220973.2020.1751579

Zhu, W. (1998). Test equating: What, why, how? *Research Quarterly for Exercise and Sport*, *69*(1), 11–23. https://doi.org/10.1080/02701367.1998.10607662