IJECA
International Journal of
Education & Curriculum Application
http://journal.ummat.ac.id/index.php/IJECA
ISSN 2614-3380 | Vol. 7, No. 3, December 2024

Investigating Test Equating Methods: English Examination Test

Rezkilaturahmi[1], Rahmat Danni[1], Heri Retnawati[1]
[1]Departement of Educational Research and Evaluation, Universitas Negeri Yogyakarta, Indonesia
heri_retnawati@uny.ac.id

Keywords:
Equating;
English Test;
Item Responses Theory.

ABSTRACT

The assessment is carried out in order to open up learning. The instruments used must of course carry out a good development stage process. The purpose of this study is to ensure reliable and fair test items in each educational unit. This research uses a quantitative approach which focuses on the equating method School Examination test equipment especially for English language subjects at the high school level in Muna district which has 2 schools. Data collection was carried out through documentation of student responses on the School Examination for English subjects. Student responses came from 2 question packages from 2 middle schools in Muna District. In addition, there were a total of 96 students involved in the school exam with each SMA 1 having 56 students and SMA 2 having 40 students. There were only 2 high schools of the same level in the Tongkuno area, especially Muna Regency, and took the question packages from the same source. Data were analyzed using an equating technique based on Item Response Theory with the *mean-mean* method. Item parameter estimation and equalization were used with the help of *R Studio* and *Microsoft Excel* programs. Therefore, *R Studio* is used to estimate the validity and reliability of the two question packages while *Microsoft Excel* is used to equalize the 2 different scores. The equating results in an equalization constant that ensures the two test packages now have scores that can be compared more fairly and accurately. Thus, these findings show the importance of the equating process in improving the integrity of educational evaluation in Indonesia especially to ensuring equivalence in two different packages is crucial to produce equal and fair score results.

Article History:
Received: 09-07-2024
Revised : 20-09-2024
Accepted: 31-09-2024
Online : 03-12-2024

Crossref
https://doi.org/10.31764/ijeca.v7i3.25065

———————— ◆ ————————

A. INTRODUCTION

School examinations serve as a means for educational institutions to assess student competencies and reflect the overall learning outcomes achieved by these institutions (Depdiknas, 2003). The instruments used in school exams must adhere to criteria related to content, construction, and language, in line with the educational level and competencies of the students (Prasetyo & Pratomo, 2021). To enhance the quality of education and harmonize assessment standards across Indonesia, the government has implemented the National Standard School Examination (USBN) (Rosidin et al., 2019). Therefore, the objective of USBN is to standardize school examinations at the national level, accommodating the evolving education system in Indonesia.

The USBN initiative aims to evaluate student competencies across various educational units. According to the 2018/2019 Standard Operational Guidelines (POS) issued by the National Education Standards Agency, the instruments used in school exams are developed by teachers at each individual school (Prasetyo & Pratomo, 2021). This diversity results in unique and varied exam instruments across different schools, which can lead to disparities in the difficulty levels of the questions (Kurniawati & Sundawa, 2019). To address this issue and ensure score equality across different instruments, a process of balancing or equating is required (Sainuddin, 2018). This process is crucial to ensure that assessments conducted through USBN are reliable and fair across all educational units, accurately and consistently reflecting the true abilities of the students (Rosidin et al., 2019).

Unfortunately, school exams at the high school level in Muna district were made by teachers who are members of the English teacher MGMP. The question sets developed by teachers ignore several important stages in the development of question items, especially English, one of which is the trial test used. So far, teachers tend to use agreements in determining the question items to be used for school exams. In this case, characteristics such as validity, reliability, level of difficulty, discrimination and pseudo guessing are still ignored. This results in unfair assessments of students, especially those with low abilities because there is no equalization of question items, which results in students with high abilities getting it easier to answer questions with a high level of difficulty. Equating is often underrecognized in the Indonesian educational landscape, despite its critical role in educational assessment (Elvira & Sainuddin, 2021). The significance of equating should be more prominently acknowledged by education experts and practitioners in Indonesia, particularly given the country's regional diversity which leads to variations in instruments used to measure student abilities, even when employing the same framework (Yusron et al., 2020). Therefore, equating is essential when there are multiple tests with the same construct or subject that yield different scores for the same participant.

Hambleton, equating as the process of converting scores from score X to score Y, or vice versa, enabling the comparison of scores (Heri Retnawati, 2016). In addition, equating as a statistical process used to adjust scores on test forms so that they can be used interchangeable. In other words, equating as establishing a relationship that allows scores on two different tests to be compared (Nisa & Retnawati, 2018). From these definitions, it is evident that equating involves adjusting scores between two or more tests with the same subject or construct, thus allowing the scores from each test to be comparable or interchangeable. This process aims to produce scores that can be substituted for one another, as articulated by Heri Retnawati in (Muhson et al., 2017). Test equating is categorized into two types: horizontal equalization and vertical equalization, as explained by (Heri Retnawati, 2016). Horizontal equating, according to (Heri Retnawati, 2016) involves equating tests that are administered in different forms or at different times but at the same educational level. Corroborates this, stating that horizontal equating is conducted between two tests at the same level. Conversely, vertical equating adjusts scores between tests taken by participants at different levels but measuring the same traits (Aşiret & Sünbül, 2016). Vertical equating applies to test instruments of varying difficulty levels that measure the same trait, where the score distributions of the participants are not comparable, thus allowing the scores to be interchangeable (Heri Retnawati, 2016). Equating in tests can be further divided based on the level of the test takers. Horizontal equalization occurs when two or more tests are administered to groups at the same level. In contrast, vertical equalization is used for tests given to groups at different educational levels.

Furthermore, while equating test scores, four fundamental designs are often employed. First, single-group designs involve administering two or more tests to the same group. Although straightforward, this method is vulnerable to external factors such as fatigue and practice effects. Second, equivalent-group designs use two equivalent groups of participants who take different test forms, drawn from the same population or assumed to have similar abilities. This design minimizes the impact of external factors but requires a large sample size to account for individual ability variations. Third, the anchor-test design utilizes common items administered to different groups, effectively addressing group equivalence issues in equalization. Lastly, alternate participant designs have the same participants take both tests, which can lead to fatigue and necessitate a significant time gap between tests (Amelia, 2016). Thus, among these designs, the anchor-test design is often prioritized in research due to its efficacy in addressing group differences and yielding the lowest error rates in equating results.

Previous studies were conducted to confirm the equating of the test equipment used in various exams. Mutluer (2017) conducted research which showed that linear and equating methods can be used to equate ALES (Academic Personnel and Postgraduate Education Entrance Exam) scores between different exam periods. The results indicated a positive and linear relationship between the original and equated scores. There was no significant difference in the difficulty level of verbal items on the ALES exam when conducting the equating of the USBN test equipment for elementary schools based on classical test theory.

Additionally AM & Retnawati (2023), conducted research on equating methods and found that the UN IPA questions for junior high schools in Indonesia consisted of 5 test packages, had a good level of difficulty, but some items had a low discrimination index. The equating method used was Item Response Theory 3 PL with R Studio and the Stocking & Lord curve proximity test, which showed the most consistent scores. Similarly Yusron et al. (2020), conducted equating research and found that the five USBN 2018/2019 test packages for compulsory high school mathematics were generally equivalent, with the Haebara method providing the best equating compared to the other three methods. Data were collected from student responses at four schools in Yogyakarta and South Kalimantan and analyzed using equating techniques with the R Program. These findings provide examples of difficult items as references to improve the quality of mathematics education. Based on the studies that have been conducted, there is still no study that directly analyzes the equivalence of 2 school exams conducted in Muna district, so this study takes the idea of conducting an equivalence analysis on the same questions used in school exams in Muna district.

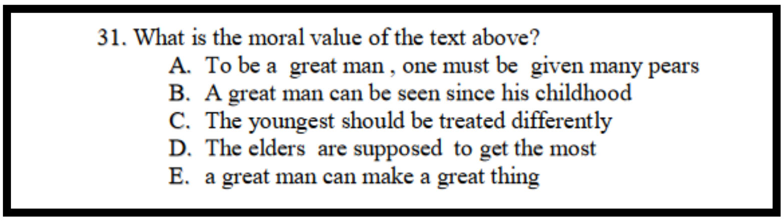
In other words, there has not been much research exploring the equating of English subject test equipment used in school exams, especially at the high school level or equivalent in Muna Regency. Therefore, this research was carried out to fill this gap. Based on this, the research aims to describe the equality of school examination equipment in 2019/2020. This study ensures the equivalence of scores on the same items in two different question packages. This is important because the instruments used are unique and varied, causing differences in difficulty levels. Differences in difficulty levels will benefit students with good abilities. In other words, ensuring equivalence in two different packages is very important to produce equal and fair score results.

B. METHODS

This research is a descriptive study with a quantitative approach. Data collection was carried out through documentation of student responses to the compulsory English subject school exam at Muna Regency in 2019/2020. In addition, the 2 schools exam packages have 10 questions together, where package A is numbered 31, 32, 33, 34, 35, 36, 37, 38, 39 and 40 while package B is numbered 1, 2, 3, 4, 5, 6, 7, 8, 9 and 10. In addition, to support data analysis in this study, *R Studio* and *Microsoft Excel* software are needed. Quantitative analysis with the Item Response Theory approach. The stages of data analysis by ensuring the validity and reliability of the two question packages. Because it has met the requirements of a valid and reliable instrument, it is then continued to ensure the equivalence score of the same item based on different question packages, as shown in Table 1 and Figure 1.

Table 1. Shared Item Distribution

Package	Same item.	Total	Participant
A	31, 32, 33, 34, 35, 36, 37, 38, 39, 40	40	501
B	1, 2, 3, 4, 5, 6, 7, 8, 9, 10	40	506



31. What is the moral value of the text above?
    A. To be a great man , one must be given many pears
    B. A great man can be seen since his childhood
    C. The youngest should be treated differently
    D. The elders are supposed to get the most
    E. a great man can make a great thing

Figure 1. Examples of Joint Question Items in Packages A and B

The table shows one of the same items found in the question items of packages A and B. Item 31 shows the same type of question as package B in item number 8. In addition, other items have similar items between packages A and B. The score results from the same item equation will be equalized to produce a fair score for students, both students with high abilities and students with low abilities.

C. RESULT AND DISCUSSION

1. Validity and Reliability

The validity of the instrument is carried out using expert judgment. Expert judgment consists of 5 people who have expertise in preparing English language question instruments. Then the English questions were reviewed and analyzed using the Aiken index. From the results of the expert judgment study using Aiken calculations, it can be concluded that each question package, both question package A and question package B, has the lowest Aiken index of 0.78 and the highest is 1, so it can be interpreted as a fairly high coefficient. Meanwhile, the package A test instrument has an Aiken index of 0.85 and package B has an Aiken index of 0.87 so that both A and B have relatively high content validity. In addition, the reliability of the instrument has been estimated using the R program. Reliability is carried out to show the consistency of the instrument when tested on different students and at different time (Yılmaz, 2023). The reliability estimate found was 0.96, indicating that the reliability of the test instrument used was very reliable.

2. Analysis of Item Characteristics

English question packages A and B were then analyzed for model suitability first using the Item Response Theory approach. IRT analysis was carried out using the *R Studio* program. The results showed that the two packages were suitable for the 2PL model. After conducting a model suitability analysis, then carry out an analysis of the item characteristics based on the 2PL model. According to the PL 2 model (Logistic Parameters), the criteria for assessing good test items are determined based on two parameters: index discrimination (a) difficulty level (b). Therefore, good discrimination index is indicated by an index (a) which is in the range 0 to 2. Meanwhile, the level of item difficulty is considered to meet the criteria if the index (b) is between -1 and +2. Items that meet these two criteria will be retained for use in further measurement activities. However, if there are items that do not meet the criteria, then these items must be dropped or replaced with backup items that have been prepared previously. Thus, detailed analysis of the items based on student answers, using item response theory with the help of the *R Studio* program, is usually presented in tabular form, as shown in Table 2.

Table 2. Analysis Result of 2 PL in A Package

Item	Index Discrimination (a)	Category	Item Difficulty (b)	Category	*p value*	Fit	Category
01	0,867	Good	1,327	Good	0,5867	Fit Model	Accepted
02	0,756	Good	1,971	Good	0,8501	Fit Model	Accepted
03	0,846	Good	-0,032	Good	0,8076	Fit Model	Accepted
04	0,594	Good	0,515	Good	0,7085	Fit Model	Accepted
05	0,922	Good	1,492	Good	0,963	Fit Model	Accepted
06	1,22	Good	-0,577	Good	0,0651	Fit Model	Accepted
07	0,775	Good	0,994	Good	0,7559	Fit Model	Accepted
08	0,766	Good	-0,187	Good	0,5696	Fit Model	Accepted
09	1,074	Good	-0,039	Good	0,2635	Fit Model	Accepted
10	0,826	Good	0,439	Good	0,1516	Fit Model	Accepted
11	0,701	Good	0,091	Good	0,6588	Fit Model	Accepted
12	1,033	Good	1,59	Good	0,468	Fit Model	Accepted
13	1,008	Good	-0,292	Good	0,9543	Fit Model	Accepted
14	1,11	Good	0,847	Good	0,13	Fit Model	Accepted
15	0,929	Good	-0,411	Good	0,8967	Fit Model	Accepted
16	0,914	Good	0,175	Good	0,8993	Fit	Accepted

Item	**Index Discrimination (a)**	**Category**	**Item Difficulty (b)**	**Category**	***p value***	**Fit**	**Category**
						Model	
17	0,882	Good	-0,668	Good	0,4083	Fit Model	Accepted
18	1,068	Good	0,694	Good	0,4302	Fit Model	Accepted
19	0,999	Good	-0,491	Good	0,1631	Fit Model	Accepted
20	0,776	Good	-0,409	Good	0,166	Fit Model	Accepted
21	0,71	Good	1,488	Good	0,812	Fit Model	Accepted
22	1,024	Good	-1,628	Good	0,212	Fit Model	Accepted
23	0,974	Good	-0,813	Good	0,7501	Fit Model	Accepted
24	0,897	Good	0,084	Good	0,2882	Fit Model	Accepted
25	0,946	Good	-0,223	Good	0,3421	Fit Model	Accepted
26	0,895	Good	-1,432	Good	0,2836	Fit Model	Accepted
27	1,157	Good	0,227	Good	0,4775	Fit Model	Accepted
28	0,83	Good	-0,71	Good	0,7294	Fit Model	Accepted
29	0,709	Good	-0,349	Good	0,5009	Fit Model	Accepted
30	0,788	Good	-0,898	Good	0,9964	Fit Model	Accepted
31	0,718	Good	-0,873	Good	0,6976	Fit Model	Accepted
32	0,855	Good	0,814	Good	0,8785	Fit Model	Accepted
33	1,051	Good	-0,03	Good	0,9839	Fit Model	Accepted
34	0,586	Good	-1,469	Good	0,4041	Fit Model	Accepted
35	0,814	Good	0,215	Good	0,6913	Fit Model	Accepted
36	0,787	Good	-0,9	Good	0,9886	Fit Model	Accepted
37	0,823	Good	0,068	Good	0,4498	Fit Model	Accepted
38	0,922	Good	-0,226	Good	0,5662	Fit Model	Accepted
39	0,896	Good	0,38	Good	0,9224	Fit Model	Accepted
40	0,645	Good	-0,459	Good	0,479	Fit Model	Accepted

Based on the results of the analysis of package A items using the 2 PL model, information was obtained regarding the characteristics of different power items (a) and level of difficulty (b). As presented in Table 18, it is known that item number 1 to item 40 have a good distinguishing power index (a) because the differentiating power index of all items is within the criteria for good items, namely 0 to 2. The lowest differentiating power index for the question items is owned by the test items. Number 34 is 0.586 and the highest differential power index is owned by item number 6, namely 1.22. The average differential power index in package A is 0.877. Meanwhile, if we look at the difficulty level of the items in package A, from Table 18 it can be seen that all the items in package A have good criteria because all items have a difficulty level index according to good criteria, namely -2 to 2. In package A, the item difficulty level index (b) was obtained by item number 22, namely -1.628 and the item with the highest index was obtained by item number 2, namely 1.971. The average difficulty level index for package A questions is 0.001.

3. Equating Test

In this research, the equating design used is the common items model. Both Package A and Package B contain 10 shared items that make up 25% of the total items, with the shared items in Package A located at positions 31 to 40, and in Package B at positions 1 to 10. The equating between Packages A and B is carried out with reference to on two main parameters, namely distinguishing power (a) and level of item difficulty (b), which guides the use of the average method in equating. This method calculates the equating constants, α and β, based on the average differential power and difficulty level of the shared items in both packages. Detailed information about the average power difference and level of difficulty before equalization can be found in Table 3.

Table 3. Analysis Result of 2 PL in A Package

Nu.	A		Nu.	B	
	Discrimination Index (a)	**Item Difficulty (b)**		**Discrimination Index (a)**	**Item Difficulty (b)**
1	0,867	1,327	31	0,497	0,8737
2	0,756	1,971	32	0,583	0,8196
3	0,846	-0,032	33	0,423	0,2024
4	0,594	0,515	34	0,535	0,3426
5	0,922	1,492	35	0,473	0,6942
6	1,22	-0,577	36	0,54	0,7312
7	0,775	0,994	37	0,523	0,8171
8	0,766	-0,187	38	0,565	0,4158
9	1,074	-0,039	39	0,576	0,9874
10	0,826	0,439	40	0,501	0,9702
Average	0,8646	0,5903	Average	0,5216	0,68542

Therefore, once the average index discrimination and item difficulty of the anchor items in each test package are determined, the equating process is conducted using the *mean-mean* method. This approach calculates the equalization constant required to adjust scores from package A to package B. The process is manually performed using *Microsoft Excel*. The results of these calculations reveal the adjusted power values and the new item difficulty post-equating. In other words, this ensures that the scores from the two test packages are now comparable in a fair and accurate manner, as shown in Table 4.

Table 4. Average discrimination index and item difficulty after equating

Equating	Average	
	Discrimination Index (a)	Item Difficulty (b)
A Package to B Package	0,5216	0,6854

After the equating process using the mean/mean method, the differential power (a) and difficulty level (b) of items in package A have been adjusted. The differential power decreased from 0.8646 to 0.5216, and the difficulty level increased from 0.5903 to 0.6854. This process also generated constants α and β, which are used to adjust the student ability parameters (θ). As a result, the abilities of students taking package A can now be considered equivalent to those taking package B. The derived equation for adjusting student abilities between the two packages is $\theta_i = 1.65576\theta_i - 0.2930$. ensuring consistent and fair measurements across both groups of students. Furthermore, after obtaining the equalization equation $\theta_i = 1.65576\theta_i - 0.2930$, this equating is then applied to the ability scores of students who work on package A questions. This process allows the scores obtained from package A to be adjusted so that they are equivalent to the scores that might be obtained if this participant worked on package B. In this way, we can compare effectively and fairly between participants who worked on two different question packages. Therefore, the results from applying this equation shows changes in participants' ability scores, reflecting adjustments to achieve equality between the two question packages. This ensures that the evaluation of student abilities is carried out using consistent and standardized criteria, minimizing bias that may be caused by differences in the item difficulty or index discrimination of the items in each package, as shown in Table 5.

Table 5. Summary of equating the abilities of package A students to package B

Indicator	A Package (X)	A Package (Y*)	B Package
Number of students	501	501	506
Number of Items	40	40	40
Average θ	0	-0,293	0,00
θ High	2,611	4,036	2,997
θ Low	-2,657	-4,698	-2,490

Through the equating process of two School Examination English test packages from two schools in Muna District, which involved comparing different power indices and levels of difficulty, it was discovered that the two test packages were almost equivalent. This finding aligns with the explanation by Retnawati (2016) who stated that the questions used in the National Standard School Examination met the criteria for equality of characteristics. Furthermore, the almost equal characteristics of the two English test packages indicate that there are no significant differences that could harm or benefit students in facing the exam. There are no significant differences, making it easier for students not to think about the differences in items, which makes students less confident in solving questions with friends who have different ability levels. This provides many concrete examples where students with low abilities will be very unsure about question items that are different but not equivalent. In fact, the questions developed in the national exam that have been tested provide equality results that are not very significant. As explained by Yusron et al. (2020) equaling in exam instruments is expected to maintain student motivation and avoid the perception that differences in exam results are caused by variations in the difficulty level of question packages. In addition, Variations in difficulty levels are often debated in the differences in exam question packages. This difference in difficulty level reduces the motivation of students who have low ability to answer the questions. Thus, the equality that has been tested at least encourages students' enthusiasm and motivation to be able to increase their self-confidence in answering question items, especially in English subject exams.

In the context of learning assessment, it is crucial to have accurate instruments to assess student learning outcomes effectively (Argianti & Retnawati, 2020). Therefore, accurate instruments can help facilitate effective assessment. In addition, the assessment process must include tracking students' learning activities and achievements, as well as evaluating their understanding of the material taught. This context emphasizes that understanding of material in English language learning in the classroom must be evaluated effectively and efficiently. This assessment will have an impact on good learning outcomes. However, assessments that do not truly evaluate students' understanding of learning will provide ineffective assessment results. Tshere are two important aspects in the test equalization process: parameter estimation and equalization estimation. Both aspects must be involved in the equalization process in order to produce significant results. Thus, respondents will need the number of items and estimation methods to achieve significant equalization results. Aspects that need to be considered in parameter estimation include the number of respondents, number of items, and the estimation method used (Wasidi & Widiyati, 2022). Meanwhile, in equating estimation, it is necessary to pay attention to the distribution of item parameters, distribution of ability parameters, the method used, number of items, and the software used in the analysis (Elvira & Sainuddin, 2021). All of this aims to produce question packages that have equivalent characteristics, supporting validity and reliability in learning assessment. It was agreed that this could help achieve more significant distribution of results so that it could have an impact on students' motivation and enthusiasm for working on the questions. However, if one aspect is not involved in the equalization analysis it will certainly have different equalization results. Thus, item estimation, distribution of ability parameters, methods used, number of items must be involved in the equalization analysis

In educational assessments, equating is a crucial statistical process used to ensure that different test forms or versions are comparable (Fong & Chuen, 2023). Therefore, statistical analysis plays an important role in ensuring that the test used is very good. The test used, in this case the English exam questions, must carry out good and correct statistical tests to provide significant results, making it easier for students to answer the questions with confidence. In addition, have extensively studied various equating methods, including linear equating, equating, and Item Response Theory (IRT) based equating. This equalization analysis is very useful for providing equalization results that are not significantly significant for the two different packages. Equalization analysis is carried out to see the equalization results on two different test packages. It's the same with the English exam questions which have two different packages. In addition, this different test package reduces students' motivation in completing questions, especially students with low abilities. IRT analysis in equating is really needed to prove that differences in question packages are not something that is debated by students. Their work has provided a foundation for understanding how to maintain fairness and accuracy in assessments across different test administrations (UYSAL & DOĞAN, 2021). In this way, the questions worked on are fair and have proper accuracy. Working on questions with different packages should not be a debate in assessing marks because IRT statistical analysis has been carried out with equating.

Additionally, the importance of using robust equating methodologies to account for differences in test forms, ensuring that scores are interpretable and comparable. Therefore, equalization is crucial in ensuring differences in compared tests and scores. These two things will prove to provide fairness to students, especially the English exam. Thus, English language tests administered as school tests or national standardized tests must have evidence of significant equity to ensure scores can be interpreted and compared. In addition, this emphasized the need for comprehensive equating studies that consider various factors, such as test length, item difficulty, and test-taker ability distributions (Tanjungpura, 2015). Their research supports the idea that proper equating practices can mitigate the effects of different test conditions, thereby maintaining the integrity of the assessment process (Yusron et al., 2020). Thus, testing using statistical methods is very crucial for the questions to be tested in order to provide significant equal results by looking at various factors, for example the duration of the test, the difficulty of the questions and the distribution of test takers' abilities.

In addition, (Nisa & Retnawati, 2018) contributed significantly to the field by developing IRT-based equating methods, which offer more sophisticated techniques for dealing with complex data structures and ensuring the accuracy of test score interpretations. These methods consider item characteristics and test-taker abilities, providing a more detailed and nuanced approach to equating compared to traditional methods. IRT statistical methods in equalization must be involved in types of questions that have different packages. This is in line with the existence of different packages for English exam questions which offer two different question packages. Two different question packages will be debated by students, especially students with different abilities. In addition, differences in different question packages will provide a perspective of unfairness in giving questions. Thus, statistical methods in IRT are really needed, especially for question item developers.

In summary, the findings from the equating process of the two School Examination English test packages in Muna District highlight the importance of maintaining equivalent test characteristics to ensure fairness and accuracy in student assessments. The contributions to improve the quality and equity of educational assessments. By adhering to rigorous equating practices, educators and assessment developers can produce reliable and valid test instruments that accurately reflect student learning outcomes and support effective educational decision-making. In addition, reliability validity in the IRT statistical analysis process is highly emphasized to see the accuracy and consistency of each question item. The two types of items tested for equality must meet valid and reliable standards. The results of the items that have been tested then ensure that the two question packages have significant equality. This will have an impact on student motivation in working on questions. Thus, it cannot be influenced by the motivation of students who have low ability to answer questions that have different question packages because they have been tested using the equating method.

D. CONCLUSION AND SUGGESTIONS

Based on the results of the research carried out, it can be concluded that the instrument for the School Examination for English language subjects in Muna District academic year 2019/2020, which was prepared by a group of English teachers, has been tested for very high validity and reliability. Apart from that, the results of the logistic parameter test carried out using the *R Studio* program show that the English School Examination test equipment instruments, both packages A and B, fit the 2 PL model. Apart from that, 2PL analysis shows that the average index discrimination (a) in package B questions is 0.542, where the lowest index discrimination (a) is owned by item number 33, namely 0.423, while the highest index is owned by item number 9, namely 0.721. Apart from that, it is known that the level of item difficulty (b) in question package

B, the lowest index is owned by item number 29, namely -1.799 and the highest index is owned by item number 21, namely 1.357 with an average index of difficulty level for item (b) is -0.3078.

Furthermore, the results of equating the abilities of students who worked on package A and package A and package B questions obtained the following equation $\theta_i = 1.65776\theta_i - 0.2930$. Applying this equation, the average ability (θ) of students who worked on package A questions was obtained has an average ability of -0.293 and those who work on package B questions have an average ability (θ) of 0.00. Thus, with this equating method, students do not need to be concerned about facing different test packages, as the difficulty levels of the items have been standardized. Therefore, this findings from the process of equalizing two school exam English test packages in Muna Regency highlight the importance of maintaining equivalent test characteristics to ensure fairness and accuracy in student assessment. Ensuring that test packages are equivalent is critical to maintaining the integrity of the assessment process and upholding student trust in the system. In addition, the results of this study indicate that rigorous equity practices can significantly improve the quality and equity of educational assessments. By following these practices, educators and assessment developers can create reliable and valid test instruments that accurately reflect student learning outcomes. This, in turn, supports effective educational decision making, allowing educators to identify areas where students may need additional support and adjust instruction accordingly.

REFERENCES

AM, M. A., & 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. https://doi.org/10.21043/thabiea.v6i1.19503

Argianti, A., & Retnawati, H. (2020). Characteristics of Math national-standardized school exam test items in junior high school: What must be considered? *Jurnal Penelitian Dan Evaluasi Pendidikan*, *24*(2), 156–165. https://doi.org/10.21831/pep.v24i2.32547

Aşiret, S., & Sünbül, S. Ö. (2016). Investigating test equating methods in small samples through various factors. *Kuram ve Uygulamada Egitim Bilimleri*, *16*(2), 647–668. https://doi.org/10.12738/estp.2016.2.2762

Depdiknas. (2003). *Sistem Penilaian Kelas SD, SMP,SMA dan SMK*. 93–115.

Elvira, M., & Sainuddin, S. (2021). Equating Test Instruments Using Anchor to Map Student Abilities Through the R Program Analysis. *Proceedings of the International Conference on Engineering, Technology and Social Science (ICONETOS 2020)*, *529*(Iconetos 2020), 651–657. https://doi.org/10.2991/assehr.k.210421.095

Fong, C. Z., & Chuen, T. Y. (2023). Test Score Equating and Item Anchoring for High Stakes Examination Test Score Equating and Item Anchoring for High Stakes Examination. *University of Malaya*, *18*(March), 1–11. https://doi.org/10.15625/2615-8957/22210401

Heri Retnawati. (2016). www.nuhamedika.gu.ma

Kurniawati, S., & Sundawa, D. (2019). An Analysis of National Standard School Examination Items Based on the Characteristics of Hots (Higher Order Thinking Skills) Questions for the Main Items of K13-071 Academic Year 2016/2017 in Karawang Regency. *International Journal Pedagogy of Social Studies*, *3*(2), 100–112. https://doi.org/10.17509/ijposs.v3i2.15793

Muhson, A., Lestari, B., Supriyanto, S., & Baroroh, K. (2017). The development of practical item analysis program for indonesian teachers. *International Journal of Instruction*, *10*(2), 199–210. https://doi.org/10.12973/iji.2017.10213a

Mutluer, C. (2017). *European Journal of Education Studies (Academic Personnel And Postgraduate Education Entrance Exam) Scores Obtained At Different Times In A Year i*. 96–120. https://doi.org/10.5281/zenodo.1101229

Nisa, C., & Retnawati, H. (2018). Comparing the methods of vertical equating for the math learning achievement tests for junior high school students. *REID (Research and Evaluation in*

Education), *4*(2), 164–174. https://doi.org/10.21831/reid.v4i2.19291

Prasetyo, O., & Pratomo, A. R. (2021). Evaluasi Penghapusan Ujian Sekolah Berstandar Nasional (USBN). *Edukatif : Jurnal Ilmu Pendidikan*, *3*(6), 4102–4107. https://doi.org/10.31004/edukatif.v3i6.1281

Rizki Nor Amelia, F. A. S. (2016). Aplikasi Model Penskoran Equal Weighting Dan Differential Weighting Untuk Mengestimasi Skor Kimia Siswa. *Jurnal Evaluasi Pendidikan*, *4*(1), 80–89. http://journal.student.uny.ac.id/ojs/index.php/jep%0AAPLIKASI

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

Sainuddin, S. (2018). Analisis Karakteristik Butir Tes Matematika Berdasarkan Teori Modern (Teori Respon Butir). *Jurnal Penelitian Matematika Dan Pendidikan Matematika*, *1*(1), 1–12.

Tanjungpura, U. (2015). *Prosiding SEMIRATA 2015 bidang MIPA BKS-PTN Barat Universitas Tanjungpura, Pontianak Hal. 171 - 179*. 171–179.

Uysal, İ., & Doğan, N. (2021). Automated Essay Scoring Effect on Test Equating Errors in Mixed-format Test. *International Journal of Assessment Tools in Education*, *8*(2), 222–238. https://doi.org/10.21449/ijate.815961

Wasidi, W., & Widiyati, E. (2022). Pelatihan Penyetaraan Skor Hasil Ujian di SMP IT Khairunnas Bengkulu. *Dharma Raflesia : Jurnal Ilmiah Pengembangan Dan Penerapan IPTEKS*, *20*(2), 262–270. https://doi.org/10.33369/dr.v20i2.21105

Yılmaz, Ö. (2023). Adaptation of the Job Stress Scale into Turkish. *International Journal of Contemporary Educational Research*, *10*(2), 535–543. https://doi.org/10.52380/ijcer.2023.10.2.444

Yusron, E., Retnawati, H., & Rafi, I. (2020). 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