Almufi Journal of Measurement, Assessment, and Evaluation Education (AJMAEE)
Website: http://almufi.com/index.php/AJMAEE
Email: almufi.ajmaee@gmail.com

Comparison Equating Method Based on Sample Size

INFO PENULIS

Deni Iriyadi
Universitas Islam Negeri Sultan Maulana Hasanuddin Banten
deni.iriyadi@uinbanten.ac.id

hevriana hartati
Universitas Islam Negeri Sultan Maulana Hasanuddin Banten

Ahmad Rustam
Universitas Sulawesi Tenggara
ahmad.rustam1988@gmail.com

INFO ARTIKEL

ISSN: 2798-0448
Vol. 3, No. 2 , Desember 2023
http://almufi.com/index.php/AJMAEE

Saran Penilisan Referensi:

Iriyadi, D. (2023). Comparison Equating Method Based on Sample Size. *Almufi Journal of Measurement, Assessment, and Evaluation Education, 3*(2), 18-24

Abstract

The purpose of this research is to compare the linear technique to the equipercentile method in terms of the variance of equalization scores using a sample size of one hundred and three different test lengths of twenty-five, twenty-nine, and thirty minutes respectively. In this particular investigation, the value of the variance of the equalization score serves as the dependent variable, while the equalization technique and the duration of the exam serve as the independent variables. A learning outcome exam for physics was used as the instrument of assessment in this particular research project. The participants were high school students. The replies of all of the people who took part in the study and worked on the several test sets that the researcher had put together constituted the study's population. The research sample consisted of the replies provided by the students for each package, for a total of one hundred responses. These responses were selected using a random selection method that included returns, and the process was repeated thirty times. We put the study hypothesis through its paces by employing comparative hypothesis testing in conjunction with t-test analysis. The findings indicated that the variance values of the equalized results obtained via the application of the linear technique were lower than the variance values obtained through the application of the equipercentile approach for test durations of 20, 25, and 30 when there was a sample size of 100.

Keywords: variance, equating, linear, equipercentile

A. Introduction

Poor learning outcomes are directly tied to the capacity of educators to carry out the learning process and produce learning assessment tools, both of which have an influence on the quality of education provided, which in turn has an impact on poor learning outcomes. This evaluation of learning is inextricably bound up with the assessment and measurement of learning outcomes. During the process of implementation, the possibility of mistakes occurring in the testing of the skills of students owing to the measuring devices that are used as well as the security or confidentiality of the exams, particularly for schools that employ parallel classrooms. Educators will create more than one exam set in order to circumvent this issue, despite the fact that each test set will be organized based on the same grid. However, it is still feasible to have different levels of difficulty, which might cause issues with measuring and evaluation. It is required to equalize scores (equating) across multiple test sets in order to solve this issue. However, the grid that is used to assess the tests must remain the same. This will allow different students who took the various exams to be compared to one another and to get equitable treatment based on the findings of the learning evaluation.

There are two different approaches to equalization that are used in classical score equalization. These approaches are known as the linear equalization technique and the equipercentile equalization method. According to Dali (2013), the relationship between equivalent scores and original scores is linear in the case of linear equalization. On the other hand, in the case of equipercentile equivalence, the percentile rankings on equivalent scores are equated with the percentile rankings on the original score, which results in a relationship that is typically nonlinear. According to the findings of Dvorak's study, the score equalization approach functioned pretty well regardless of the factors tested, including the duration of the exam, the proportion of anchor questions, and the sample size. Using the anchor item group design is one of the scoring equalization designs that may be used to tackle the problem of having several classes that are parallel to one another. According to Kolen and Brennan (1995), the number of anchor points that are used as connections is at least 20% of the overall examination. Crocker and Algina (1986) said that the minimum length of connected things should be equal to 20% of the total length of all items. Angoff (1984) suggests the exact same notion, which is that tiled items should account for 20% of the total number of tests and be included in both sets of tests. Because the test for education has a distribution of content, the number of coupled items needed to achieve equality is greater than 20% of the test set consisting of 40 items or more, according to research carried out by Harris and published in Kolen and Brennan (1995). This information is stated based on the simulation results of the unidimensional model. Heterogeneous. While this is going on, research that was carried out by Yetti Supriyati (2003) indicates that the greatest consistent variance of equalization of scores is discovered in the pair of test kits with the percentage of connected items being 20%. According to the conceptual definition presented earlier, the percentage of connected items that should make up no less than twenty percent of the total grain is the minimal amount that is suggested.

According to Skaggs (2005), one of the prerequisites for equalization is to possess a sample that is sufficiently big in order to generate an equalization that is both reliable and precise. In spite of this, it is not feasible to produce a sample big enough for the equivalency for a number of different reasons. Few research have been done on equivalency, and even fewer have been done on tiny samples. For example, Livingston (1992) discovered that a refining approach boosted the accuracy of equiperscentile equalization by double the sample size. However, there have been few studies done on small samples. The researchers Parshall, Houghton, and Kromrey (1995) discovered that the amount of standard errors significantly increased as the sample size was reduced. According to the findings of Skaggs (2005), when the sample size is more than 50, mean equating is more accurate than other approaches for determining below-average scores, but it is not as reliable as other methods for determining above-average scores. When the sample size is small and the findings are not representative, Kim, von Davier, and Huberman (2008) discovered that using synthetic functions may be a better option than the chain linear equating approach. This was the conclusion they came to. Using a sample size ranging from 50 to 400, Livingston and Kim (2010) discovered that the circle-arc approach produced the best accurate findings across the board for all sample sizes that were investigated, particularly above the average score distribution. According to these research, there are no definitive regulations addressing tiny sample sizes on the equivalency and the minimum and maximum size restrictions for declaring small samples. However, the number of samples that is most often employed runs anywhere from 25 to 200. According to this explanation, the study

was conducted using a limited sample size, and the total number of participants in the sample was one hundred.

Altering the duration of the exam will cause a shift in the distribution of test questions over its duration. Alterations to the duration of the exam will not only have an effect on the mean and variance of test results; they will also have an impact on the reliability and validity of those scores. According to Azwar (2015), the increase in the number of items on the exam causes a corresponding rise in the reliability coefficient. According to Thorndike's research from 1982, the test's dependability will improve if it is allowed to continue for a longer period of time. For standardized tests, the timing is based on the results of the trial, but for tests in class, the timing is based on the experience of each individual teacher. According to Rasyid and Mansur (2009), the determination of the length of the test is generally based on the coverage of the test material and the fatigue of the test participants. In the meanwhile, according to Mardapi (2012), the duration of the exam is comprised of the total number of items as well as the total amount of time allotted to each question. The amount of time allotted to complete the test is another factor that goes into deciding how many questions will be on the exam. The amount of time spent on this activity varies from 90 to 120 minutes for secondary schools. According to Fitzpatrick and Yen (2001), the number of test items may have an effect on the stability, accuracy, reliability, and validity of the test score. They advise that the test should include at least 8 items or at least 12 test items in order to be considered legitimate. If you want your findings to be reliable, Hambleton and Cook (1979) suggest that you should have at least 200 people take the exam and that you should utilize 20 different questions. According to this explanation, the research used objective multiple-choice examinations of varying durations, including 20, 25, and 30 questions each.

Research on the comparison of the score equalization method has begun to be carried out on a large scale, including by Ongge M. R. M. Samuri Admodipuro (1993) by comparing the linear and equipersentile methods, which resulted in the equiperscentile method being more suitable than the linear method in terms of the magnitude of the standardized error of equalization. Other researchers have also begun to carry out research on the topic of the comparison of the score equalization method. Tri Rijanto (2012) investigated how the variance of the score was affected by both the technique used to equalize the scores and the total number of samples. With a sample size of 800, the equipersentile approach performs better than the linear method. Ariani Arsad's (2014) study uses the results of the National Examination (UN) in mathematics for junior high school students as a sample size of 250 to compare and contrast the linear equalization technique and the equipersentile method with regard to the variance of scores. The study is based on the linear equalization method and the equipersentile approach. While his study demonstrates that at a significance level of 0.5, there are changes in the variance of scores when using the linear technique compared to the equipersentile method, he says that these differences are not significant. Anne R. Fitzpatrick and Wendy M. Yen (2001) investigated the influence of test length, namely 2, 4, 8, 12, and 20, and sample sizes of 200, 500, and 1,000 on reliability and equivalent tests based on the Item Response theory. namely, they looked at the effect of test length on reliability and equivalence tests. According to the findings of the study, a test must include at least 12 questions in order to get a score that is valid and dependable for equalization purposes. According to the findings of Sugeng's (2010) study utilizing the vertical equalization technique on the minimum sample size, the influence of the test length, and the anchor test length, the accuracy of the equalization improves as the test length grows. According to the findings of this research, a minimum anchor test length of five is necessary for a test length of twenty items, whereas a minimum of three is necessary for a test length of ten items. study on the comparative technique of equalization of scores by altering the length of the new tests was carried out for current measures or IRT (item response theory), hence reaffirming the need of study. The research focused on the method of equalization of scores by varying the length of the new tests.

The value of the variance score is shown as a description of the variety of values on the data set so that one can evaluate the consistency of the equalization that occurs between the linear and equipersentile approaches. This is done so that one can observe whether or not the equalization occurs. The degree to which the data are distributed may be said to increase proportionally with the score's variance. According to the information presented above, the investigator is curious in contrasting the equalization of scores based on classical test theory using linear and equiperscentile approaches, as well as determining how the duration of the test affects the equalization of results. Therefore, the purpose of this research is to assess whether or not there is a significant difference in the variance of equalization scores when comparing the

linear approach and the equiperscentile method depending on the duration of the test and the sample size that is available.

B. Methodology

This research is a comparative quantitative study that makes use of the experimental method. It gathers its data from the replies that students working on formative physics test kits provide. In this particular investigation, the equivalency technique and the duration of the examination both serve as examples of independent variables. The linear equalization technique and the equipersentile equalization method are the two methods of equalization that are equivalent to one another, and the values of 20, 25, and 30 are utilized for the test length variables. The variance of the scores after they have been equalized is the dependent variable.

The participants in this research were all teenagers enrolled in high school. The people who are working on the pairs of test kits with the number of items each of 20, 25, and 30 as well as the anchor items with a percentage of 20% are separated from the rest of the population into a total of six divisions. The results of the students' test scores were picked at random from as many as 100 responders, with 30 replications for each set of test kits. Using the linear approach, the score acquired from the X test is equalized to the score gained from the Y test kit. As a consequence, a new score of equalization results, denoted by the letter Y*, is generated for each duration of the test.

The normality test using the Kolmogorov-Smirnov analysis approach is used in the testing of the necessary hypothesis, whilst the homogeneity test makes use of the Levene test analysis technique. The comparative hypothesis testing using the t-test is what is utilized to test the hypothesis. Utilizing the SPSS application to do the computation for the required test for the hypothesis and the testing of the hypothesis.

C. Results and Discussion

In this particular research, replication was carried out thirty times for each test set, each of which had a different number of items (20, 25, or 30). In addition, the score of the group that worked on the X instrument was equalized with the score of the group that worked on the Y instrument for each duration of the test. This was done in order to acquire the results of the equalization as well as the score variance.

Test Length 20

The median value of the score variance generated by the equiperscentile approach is higher than that generated by the linear method. Skewness for linear and equipersentile techniques with a test duration of 20 is positive skew, which implies that the data distribution of the value variance scores has a tendency to gather at smaller values. This is because smaller values are more likely to have more variance. There are no outliers in the score variance when using the linear technique; but, when using the equipersentile method, there is a number that is considerably distinct from the other values, to the point where it is considered an outlier; this value is 15.531. It is clear from looking at the boxplot that the scores generated by the linear symmetric equalization approach and the data are centered around the median, which indicates that the variation of the scores is quite low. The equipersentile approach, on the other hand, does not have a symmetrical structure. The distribution of the data on the variance scores of the two approaches is almost identical, as seen by the boxplot that was just presented to you.

Test Length 25

The median variance of the scores obtained using the equipersentile approach has a higher standard deviation than those obtained using the linear method. With a sample size of 25, the linear and equipersentile techniques both produce skewness distributions that have a positive value for the skewness parameter. When calculating the score variance, the linear technique does not produce any outliers; however, the equipersentile method does produce outliers. The boxplot makes it clear that the variation of scores generated by the linear and equipersentile equalization techniques has a tendency to be symmetrical, and that the median serves as the point of focus for the majority of the data. The boxplot shown above also demonstrates that the distribution of data on the linear method's variance of scores tends to be centered. This can be seen in the way that the data are distributed. On the other hand, the

equipersentile technique has a wider dispersion of data on the range of possible score combinations.

Test Legth 30

The median variance of the scores obtained using the equipersentile approach has a higher standard deviation than those obtained using the linear method. When using a test duration of 30, the linear and equipersentile techniques both produce skewness distributions that have a positive value for the skewness parameter. In the value of variance, the score calculated using the linear approach contains two outlier values, namely 10.811 and 27.236; nevertheless, the number 34.623 is the sole outlier value calculated using the equiperscentile method. The boxplot makes it clear that the variation of scores generated by the linear and equipersentile equalization techniques has a tendency to be symmetrical, and that the median serves as the point of focus for the majority of the data. The boxplot that is shown above also demonstrates that the distribution of data on the variance of scores for the linear technique has a tendency to be centered, but the distribution of data on the variance of scores for the equipersentile approach is much higher.

Before putting the hypotheses that were created during the compilation process to the test, the preparatory test is carried out. On the data pertaining to the variance of the equalization score, the precondition tests—namely, the normality test and the homogeneity test—were carried out. These tests were carried out. In order to determine whether or not the data were normally distributed, a Kolmogorov-Smirnov test was carried out with the assistance of the SPSS software using a significance level of 5%. If the significance value is more than 0.05, then the data are considered to be normal. According to the findings of the test to determine whether or not the data were normally distributed, the significant value for the variance of the linear and equipersentile method scores for the duration of the 20 test was found to be 0.200. The linear technique yields a value of 0.161 for the test length of 25, but the equipersentile method yields a value of 0.200. The linear technique gives a value of 0.200 for the test duration of 30, whereas the equipersentile method gives a value of 0.135. It can be deduced from these findings that all of the data on the variance score equalization results are normally distributed since the significance value of the Kolmogorov-Smirnov test for all of the data is higher than 0.05. This finding supports the hypothesis that the data are normally distributed.

The next test that is required as a prerequisite is the homogeneity test. The purpose of this test is to provide an overview of the diversity of the groups that are being compared by determining whether or not they have the same variance. If they do, then the differences that occur in the hypothesis testing that will then be carried out will be the result of differences between groups, and not the result of differences within the group. A significance threshold of 5% was used for the data homogeneity test, which was carried out by using the Levene test with the aid of the computer. If the statistical significance value calculated using the Levene method is more than 0.05, then the data are considered to be homogenous. According to the findings of the homogeneity test, the significance value for the length of test 20 was determined to be 0.792, the significance value for the length of test 25 was 0.287, and the significance value for the length of test 30 was 0.121. Since all of the data has a Levene statistical significance value that is larger than 0.05 and the data itself demonstrates that the value is greater than 0.05, one can draw the conclusion that all of the data on the variance of the equalization score are consistent with one another and that the process may thus be continued.

The first hypothesis proposes that the variance of the equated scores using the equipersentile approach is higher than the variance of the equated scores using the linear equivalency method when the sample size is small and the test duration is 20. According to the findings of the investigation, it has been determined that if tcount (2,812) is more than ttable (2,000), then the null hypothesis H0 must be rejected. This indicates that the variance of the equalized score acquired using the equipersentile technique is higher than that obtained through the linear equalization method when the sample size is small and the test duration is 20.

According to the second hypothesis, the variance of the equated scores when using the equipersentile approach is larger than the variance of the equated scores when using the linear equivalency method for test length 25 when the sample size is small. In light of the findings of the investigation, it has been determined that if tcount (2,513) is more than ttable (2,000), then the null hypothesis H0 must be rejected. This indicates that the standard deviation of the equalized score produced via the application of the equipersentile technique is higher than that

acquired through the application of the linear equivalent method when the sample size is small and the test duration is 25.

According to the third hypothesis, the variance of the equated scores when using the equipersentile approach is larger than the variance of the equated scores when using the linear equivalency method for test lengths of 30 when the sample size is small. If the results of the study show that tcount (2.812) is more than ttable (2,000), then H0 must be discarded. Therefore, it is possible to draw the conclusion that the variance of the equated scores when using the equiperscentile technique is bigger than the variance of the equated scores when using the linear equivalence method for the test duration of 30 when the sample size is small.

It is possible to draw the following conclusion from the findings of the three tests: the variation of the scores obtained using the equating percentile approach is higher than those obtained using the linear method for test durations of 20, 25, and 30 with significant differences. Alterations to the duration of the exam will have an effect not only on the mean and variance of test results, but also on the reliability and validity of the scores. According to Azwar, the bigger the percentage of variance that is shared by the test and by the criteria (that is, the greater the validity), the higher the variance-score fraction that seems to be a pure variance-score (that is, the higher the reliability). This demonstrates that as the length of the exam rises, so does its reliability and validity, as well as the amount of variation in the scores.

The linear approach has the same level of sensitivity all along its linear lines, but the equipersentile method generates a non-linear line that has a lower level of sensitivity at the ends and bases of the line and a higher level of sensitivity in the center part of the line. Therefore, the variability of the equalization score in the equipersentile approach increases proportionally with the length of the test questions. When it comes to test lengths of 20, 25, and 30, using the equipersentile approach yields results that are less consistent and reliable than those obtained using the linear method.

D. Conclusion

The findings of the study that was carried out led to the conclusion that the variance of the equating percentile method scores was higher than that of the linear method for test durations of 20, 25, and 30 that had significant disparities. This was determined by comparing the two methods' respective scores. Alterations to the duration of the exam will have an effect not only on the mean and variance of test results, but also on the reliability and validity of the scores. As a component of the process of evaluating a learner's progress, research on the equivalency technique and the duration of this exam may be seen as an attempt to enhance the quality of measurement.

Some ideas that need to be taken into consideration include the need of doing research using alternative methodologies, namely the methodology known as Item Response Theory (IRT), as well as the requirement of carrying out more study by contrasting various equalization techniques. If comparative equivalents are going to be determined using equiperscentile and linear approaches, then the test length need to be more varied. It is possible to do more research on individuals that have the same features as the physics subjects who were used in this study.

E. Reference

Gregory, R. J. (2002). *Psycological Testing*. USA: Pearson Education Company.

Hambleton, R.K dan H. Swaminathan. (1985). *Item Respon Theory Principle and Aplications.* Boston: Kluwer Nijhoff Publishing.

Hambleton, *R.K.,* Swaminathan, H., & Roger, H. J. (1991). *Fundamentals of Items Respon Theory*. Newsbury Park; Sage Publication.

Harris, D. J., & Kolen, M. J. (1990). A Comparason of Two Equipercentil Equating Methods for Common Item Equating, *Educational and Psychological Measurement. 50*.

Herkusumo, A. P. (2011). Penyetaraan (equating) ujian akhir sekolah berstandar nasional (UASBN) dengan teori tes klasik. *Jurnal Pendidikan Dan Kebudayaan*, *17*(4), 455-471.

Hippel, P. V. (2010). Skewness, *International Encyclopedia of Statistical Science*.

Hsiao, Y. Y., Shih, C. L., Yu, W. H., Hsieh, C. H., & Hsieh, C. L. (2015). Examining unidimensionality and improving reliability for the eight subscales of the SF-36 in opioid-dependent patients using Rasch analysis. *Quality of Life Research*, *24*, 279-285.

Huda, N., & Mardapi, D. (2015). Komparasi Model Penskoran Berdasarkan Teori Respons Butir pas Soal Ujian Nasional Mata Pelajaran Matematika. *Junal Evaluasi Pendidikan, 3*(1).

Kadir. (2016). *Statistika Terapan*. Jakarta: Rajawali Pers.

Kartono. (2008). Equating The Combined Dichotomous and Politomous item test Model in an Achiement Test. *Jurnal Penelitian dan Evaluasi Pendidikan, 7*(2), h. 315.

Kilmen, S., & Demirtasli, N. (2012). Comparison of test equating methods based on item response theory according to the sample size and ability distribution. *Procedia-Social and Behavioral Sciences*, *46*, 130-134.

Kim, J. S., & Hanson, B. A. (2002). Test equating under the multiple-choice model. *Applied Psychological Measurement*, *26*(3), 255-270.

Kim, S., & Livingston, S. A. (2010). Comparisons among small sample equating methods in a common-item design. *Journal of Educational Measurement*, *47*(3), 286-298.

Kim, S., Livingston, S. A., & Lewis, C. (2011). Collateral information for equating in small samples: A preliminary investigation. *Applied Measurement in Education*, *24*(4), 302-323.

Kim, S., von Davier, A. A., & Haberman, S. (2006). An alternative to equating with small samples in the non-equivalent groups anchor test design. *ETS Research Report Series*, *2006*(2), i-40.

Kim, S., Von Davier, A. A., & Haberman, S. (2008). Small-sample equating using a synthetic linking function. *Journal of Educational Measurement*, *45*(4), 325-342.

Kim, S., von Davier, A. A., & Haberman, S. (2011). Practical application of a synthetic linking function on small-sample equating. *Applied Measurement in Education*, *24*(2), 95-114.

Kurtz, A. M., & Dwyer, A. C. (2013). *Small sample equating: Best practices using a Sas Macro*. Retrieved from http://analytics.ncsu.edu/sesug/2013/BtB-11.pdf