EJAES

JOURNAL OF ADVANCED EDUCATION STUDIES

İleri Eğitim Çalışmaları Dergisi

6 (1):92-111, 2024

COMPARING EQUATING ERRORS ON VARIOUS FACTORS FOR SUBTESTS WHICH HAVE ADDED VALUE[1]

Arzu Uçar[2]

Önder Sünbül[3]

Geliş Tarihi/Received: 17.02.2024 Elektronik Yayın / Online Published: 25.06.2024

DOI: 10.48166/ejaes.1438652

ABSTRACT

In this study, 270 different data sets were generated through R 3.1.1. program as dichotomous data for X and Y test forms, each consisting of two sub-tests, according to the two-parameter logistic model (2PLM). Forms include an anchor test, which consists of two sub-tests like test form. The sub-test length of the anchor test is 40% of the total test. For sample sizes of 20, 25, 50, 100, 200, 500, X and Y test forms were created with correlation levels between sub-tests of 0.70, 0.80, 0.90; average difficulty level differences between test forms of 0.0, 0.40, 0.70; sub-test lengths of 10, 15, 30, 50, 80. Sub-tests were equated as a result of 100 replications using identity, linear-chain, Braun/Holland, and circular-arc equating methods. The results obtained from this simulation study were evaluated based on equating error. The findings indicated that in the case when sample size was 100 and more, subtest length was 10, 15 and 30 and the level of average difficulty difference between form 0.0, it was concluded that equating forms would give better results than not equating. Furthermore, the circle-arc method was found to have less equating error than other methods under most of the conditions studied.

Keywords: Equating, added value, augmented subscore, equating error

[1] This study is derived from the second author's master's thesis entitled Comparing Equating Methods For Subtests Which Have Added Value, prepared under the supervision of the first author.

[2]Assist. Prof. Dr., Hakkari University, Faculty of Education, Department of Educational Siences, Türkiye, arzukapcik@gmail.com, ORCID: 0000-0002-0099-1348

[3]Assoc. Dr., Mersin University, Mersin, Türkiye, ondersunbul@gmail.com, ORCID: 0000-0002-1775-1404

ARTI DEĞER ÖZELLİĞİNE SAHİP ALT TESTLERDE ÇEŞİTLİ FAKTÖRLER ALTINDA EŞİTLEME HATALARININ KARŞILAŞTIRILMASI

ÖZET

Bu çalışmada her bir alt testi artı değer özelliğine sahip olan testlerde, alt test ve genişletilmiş alt test puanları kullanılarak ortak maddelere dayalı eşitleme yapılmıştır. Eşitleme yöntemlerinin eşitleme hataları, örneklem büyüklükleri alt testler arasındaki ortalama güçlük düzeyi farkları ve alt test uzunluklarına göre karşılaştırılmıştır. Çalışmada iki parametreli lojistik modele (2PLM) uygun her biri iki alt testten oluşan ve ikili puanlanan X ve Y formu, 540 farklı veri kümesi için, R 3.1.1. programı aracılığıyla üretilmiştir. Üretilen verilerde ankor test, toplam test formu gibi iki alt testten oluşmaktadır. Ankor testin alt test uzunluğu, alt test uzunluğunun %40'ıdır. Örneklem büyüklüğü 20, 25, 50, 100, 200, 500; alt testler arasındaki korelasyon 0.70, 0.80, 0.90; test formları arasındaki ortalama güçlük düzeyi farkı 0.0, 0.40, 0.70 ve alt test uzunluğu 10, 15, 30, 50, 80 olan X ve Y formları oluşturulmuştur. Çalışmada birim, zincirlenmiş lineer, Braun/Holland ve dairesel-yay eşitleme yöntemleri kullanılarak 100 replikasyon sonucunda alt testler eşitlenmiştir. Yöntemlerin eşitleme sonuçları eşitleme hatasına göre değerlendirilmiştir.Çalışmada 100, 200 ve 500 örneklem büyüklüğüne sahip, alt test uzunluğu 10, 15 ve 30 ve test formları arasındaki ortalama güçlük düzeyi farkı 0.0 olduğunda eşitleme yapılması uygun görülürken; dairesel-yay eşitleme yöntemi diğer yöntemlere göre daha az hata değeri göstermiştir.

Anahtar kelimeler: Eşitleme, artı değer, genişletilmiş alt puan, eşitleme hatası

1. INTRODUCTION

Measurement tools used in education to estimate students' interests and abilities and to obtain information about their performance and success are called test (Baykul, 2010). Tests used in education may consist of subfields. For example, a mathematics test consists of algebra and geometry, or a general ability test consists of subfields such as mathematics, writing, and reading. The subtest scores obtained from the subtests consisting of the subfields of the tests are used to analyze the individual's learning deficiencies or the subjects in which he is more successful, as well as to reveal the profiles of the schools. Additionally, subscores are used to draw more detailed conclusions from the structure of a total test. In other words, the items used for the feature that is intended to be measured may be used to measure another feature. In this case, by combining such items, a new sub-feature may emerge and more detailed information about the structure of the test can be obtained (Shinary, Haberman & Score, 2007). With these advantages provided by the subfields that make up the test, interest in subtest scores has increased. Therefore, the individual's score in a subfield consists of the individual's sub-true score and sub-error scores, which express the true value of the feature to be measured by the sub-domain, as in the total test.

With the increasing interest in the use of subtest scores, it has been investigated whether the use of the total test score or the subscore would serve the purpose better. As a result of researches, it has been revealed that it is not appropriate to use every subtest score, but the necessary conditions for the use of the subscore have been partially stated (Haberman, 2008; Sinharay, 2010a; Sinharay, 2010b; Sinharay, Haberman, & Puhan, 2007; Sinharay & Haberman, 2011). It was emphasized that for the subtest score to be usable, the subscore must have a added value. For the added value feature, features

such as high reliability of the sub-scores, relatively low reliability of the total test and being different from other sub-tests (low relationship) are taken into consideration (Sinharay, 2010a).

To check whether the sub-score provides more information than the total test score, the sub-score has been expressed with three different calculations (Haberman, 2008). The subscores are based on three different calculations, the regression of the actual subscore on the observed subscore, the observed total test score, the observed subscore and the total test score, the subscore approximation (Sx), the total test score approximation (Sz), the total test score and the observed subscore. sub-score approximation (Sxz) was obtained. The differences between the obtained sub-scores and the actual sub-score values indicate the error of the approach. The variances of these errors give the mean squared error (MSE) value for the true subscore. The root mean square error (RMSE) value is obtained by taking the square root of the mean square error. In order to reach conclusions such as which of the sub-scores obtained with three different approaches would be appropriate to use or whether the total test score should be used, the RMSE values of the sub-scores obtained by using the observed sub-score, the total test score, and both the observed sub-score and the total test score are compared. The subscore with lower RMSE values should be used. In other words, the approach with the least error has the added value for the sub-score. Another coefficient that helps in the research of whether the sub-score has an added value is the PRMSE values that contain the reliability coefficients of the sub-score. It is said that the approach with a higher PRMSE value has an added value for the sub-score.

Tests in education are used in creating profiles of individuals or schools, tracking individuals' education levels, selecting and placing individuals in an institution, and recruiting individuals. Decisions about individuals are made in line with the scores obtained from the tests used. The decisions made play an extremely important role in shaping the lives of individuals. In order for correct decisions to be made, tests must not lead to biased decisions and must be valid and reliable. For this reason, tests related to the feature you want to measure should not be used more than once. In other words, different parallel forms of the tests can be applied on different dates or at the same time for exam security reasons. According to Classical Test Theory, test forms must be parallel in order to compare the scores obtained from different test forms. However, even if a parallel test form is prepared, since the ability levels of the individuals taking the test will be different, there will be different sources of error for each individual. Therefore, the scores obtained from the test will have different amounts of errors. In Classical Test Theory, item difficulty and item discrimination are specific to groups. Therefore, it is difficult to compare people taking different tests with parallel forms.

Test scores need to be compared to ensure that appropriate decisions can be made under appropriate circumstances. This comparison can be made by equating the tests. Equating is a statistical process and allows scores obtained from different test forms measuring the same trait to be used interchangeably. Crocker & Algina (1986) defined test equating as the process of creating equivalent scores from two tests. Angoff defined test equating as converting the unit system of one test form into the unit system of another form (Angoff, 1984).

Obtaining the data set according to certain rules for equating is called equating design (von Davier, 2010). Equating designsdesigns are divided into designs that use common individuals and designs that use common items (Davier, Holland, & Thayer, 2004). Common items are used in the tests to be equated because it is very difficult to prepare the test forms to be equated in parallel and the ability levels of the groups to which the test forms are applied are different from each other. Tests that contain common items are called anchor tests. Equating of tests is achieved through the anchor test.

The purposes of equating are to prevent bias among individuals taking different test forms, to report the scores obtained from different forms on the same scale, and to preserve the meaning of the reported scores (Quoted in: Kan, 2010:3, Barnard, 1996). Achieving the purpose of the equating means that the equating is performed without error. Equating errors occur when equating desings are not used correctly, the appropriate equating method is not applied, there is not a large enough sample, and different test forms are not statistically similar to each other. Kolen and Brennan (2004) divided equating errors into two: random and systematic equating errors. Random error can also be called the difference between the equating relationship estimated from the sample and the population. Systematic error occurs when the assumptions or conditions of the equating method used are not met. Livingston (2004) defined the standard error of equating (SEE) as the standard deviation of the sample distributions of the equated scores. Wang (2006) stated that the standard error of equating is the index of random sampling error and defined it as the standard deviation of equated scores obtained from repeated samples. The sum of the squares of the equating error and the equating bias represents the equating error. A good method should have small bias and small RMSE value (Chu and Kamata, 2003). Therefore, as the average RMSE values decrease, more accurate equating results are obtained.

In order for different test forms measuring the same trait to be comparable and interchangeable, the test scores obtained from the tests must be positioned on the same continuum. This is achieved by equating the tests. In test forms containing subtests, the subtest forms are equated with each other in terms of the measured trait in order to compare the test forms or subtest forms with each other. The information obtained from the subtest forms not only compares the subtest scores with each other, but also provides results that will enable us to make accurate judgments about the profiles of individuals in terms of the trait of interest (Sinharay 2010a, Sinharay, 2010b). It will enable us to observe situations where individuals are weak or strong in terms of the relevant trait. In order to obtain this information, the subtest forms of the test forms must have added value (Sinharay, Haberman, & Puhan, 2007). Just as it is obtained by equating the tests, equating error is also obtained as a result of equating subtests which have added value. Just as the assumptions of the equating method cannot be met or other uncontrollable situations cause systematic error, the sample size also causes error in equating (random error). A decrease in the standard error of equating was observed as a result of studies in which large samples were used in equating or changes were made by increasing the sample size. When working with large samples, the ability of the sample to represent the universe increases and the error in equating decreases, thus the quality of the equating increases.It is not always possible to reach large samples in

order to process the information obtained from different test forms measuring the same feature. It may be necessary to work with small samples in teacher-made exams or exams conducted by some course centers. When the sample size is small, problems arise when the sample does not represent the universe and the assumptions of the methods to be used cannot be met. Kolen and Brennan (2004) have stated that equating should not be done in small samples because these problems will cause too much equating error. Despite this, Livingston and Kim (2009) have expressed that models can be used to overcome the problems that small samples will cause. When tests are equated in small samples, it should be decided which equating design and equating method would be appropriate to use.

Factors such as sample size, equating desing, test length, difficulty level difference between test forms and methods to be used in equating tests affect the equating of the tests. When test levels are parallel, the difference in difficulty level between test forms will be less. However, in practice, it is very difficult to obtain similar test forms. In addition to this situation, obtaining parallel testing will become more difficult when having to work in small samples. Because it will be difficult to make predictions about the item parameters from the data obtained. There is not enough information about equating subtests using different equating methods in small samples and in cases where the difficulty levels of the test forms are different. This study aims to determine which of the various equating methods will give the smallest equating error in the observed subtest score and augmented subtest score of subtest forms with added values in small samples, different sample sizes, different difficulty levels between test forms, subtest lengths, and different lengths of the anchor test. It is thought that it will increase the quality of the equating as it provides a more micro-level examination of the equating studies to be carried out on a subtest basis. Additionally, this study is thought to contribute to the field as it examines the sub-test forms which have an added value in small samples with various equating methods.

2. METHOD

In this study, it is aimed to compare the equating error obtained from the sub-test equating performed using equating methods based on common items under different conditions. Thus, by determining the equating errors obtained by equating the sub-test scores of test forms with different sample size, different test length, and different average difficulty, it is aimed to contribute to the theoretical studies on this subject. The research is basic research in this respect.

2.1. Data Simulation

In this study, data generation according to the changing factors and the levels of these factors was carried out using the 2-parameter logistic Item Response Theory model. The data simulated is bicategorical, 1-0. For this purpose, four different individual ability distributions were used. While form X consists of subtests X1 and X2, form Y consists of subtests Y1 and Y2. The ability distribution of individuals for X1 and Y1 was obtained from the standard normal distribution with mean 0 and standard deviation 1. The ability distributions to be used for the X2 and Y2 subtests were produced in a way that would show the desired correlations with the ability distributions previously produced for X1 and Y1.

The discrimination parameters (a) of the items in the subtests and anchor tests of the X and Y test forms were obtained from a log-normal distribution with a mean of -0.15 and a standard deviation of 0.14. The difficulty (b) parameters of the two subtests of the X form and the anchor test were obtained from a uniform distribution with a mean of -2 and a standard deviation of 2. The difficulty (b) parameter of the subtests of the Y form was obtained by adding the values of 0.00, 0.40 and 0.70 to the difficulty (b) parameter in the X form.

In accordance with the varied factors, a total of 270 X forms and 270 Y forms have been generated. This includes the sample size level (20, 25, 50, 100, 200, and 500) (6), sub-test length level (10, 15, 30, 50, and 80) (5), difficulty level between test forms (0.00, 0.40, 0.70) (3), anchor test length (40%) (1), and correlation level between sub-tests (0.70, 0.80, and 0.40) (3).

Four different equating scenario were created and subtests measuring the same feature were equated with each other. In the first equating scenario; Equation was made using the observed total scores of both subtests, which have the positive value feature of the X and Y forms produced at the same factor levels. In the second equating scenario, equating was carried out using the augmented subtest scores of both subtests of the X and Y forms, which were produced at the same factor levels and showed added value. Similarly, in the third equating scenario; equating was made using the observed scores of both subtests of the X and Y forms, which were produced at the same factor levels and showed added value. In the last equating scenario, equating was carried out using the augmented subtest scores of both subtests of the X and Y forms, which were produced at the same factor levels and did not show added value. For equating of forms. The "equate" package of the R.3.1.1 program was used.

The length of the anchor test form, which enables equating, has been changed by 40% of the subtest lengths. Each subtest has been equated using the relevant subtest of the anchor test. Samples were obtained by performing 100 replications for all sample sizes. For each replication, the X form was equated to the Y form scale. Equated scores were obtained at the raw score and augmented score. Equating was carried out with test forms and test scores with the same traits. Equating errors (RMSE) values were calculated for each equating method of the equated subtest forms at all levels of the changing factors. A common effect graph was created by examining the common effects of the factors in the study, such as sample size, subtest length, and difficulty level difference between test forms, on the equating error.

3. FINDINGS

The result of the equating error (RMSE) values obtained from equating methods using the subtest and augmented subtest scores of two subtests with added value under common factors is shown in the common effect graphs below. The colors of the shapes belonging to the error values of the methods in the common effect graphs are located on the right side of the graphs in the color chart, labeled as Braun/Holland (MEAN) equating method, linear chain equating method (LIN), circular arc equating method (CIRC), identity equating (IDEN) method.

Table 1 and Table 2, which contain the common effect graphs, show the RMSE values obtained from the methods as a result of equating the subtest scores and augmented subtest scores of the first subtest, respectively. Table 3 and Table 4 show the RMSE values obtained from the methods as a result of equating the subtest scores and the augmented subtest scores.of the second subtest. Tables consist of 3 cells. The first column contains the findings regarding the subtests with a correlation of 0.70, the second and third columns contain the RMSE value findings obtained from the methods as a result of equating the subtests with a correlation of 0.80 and 0.90.

X and Y tests are tests consisting of two sub-fields. All subtests have added value. Subfields were equated to each other using subtest/augmented subtest scores using identity equating, Braun/Holland equating, linear chained equating and circular arc equating methods. The common effects of changing factors (sample size, subtest length and difficulty level difference) on RMSE values obtained from matching methods are observed on the graph.

When Table 1 shows, it can be seen that as a result of equating subtests with a correlation of 0.70, provided that the average difficulty levels between the subtests remain constant, the RMSE values of the methods decrease as the sample size increases, and while the RMSE values increase with the increase in the subtest length. However, while the subtest length is constant, it is observed that the RMSE values decrease as the sample size increases, and the RMSE values of the methods increase as the difficulty difference between the equated test forms increases. It is seen that when the subtest length is 10, 15 and 30 and the difficulty difference between the test forms decreases, the RMSE values obtained from the equating methods are less and closer to each other. However, when the subtest length is 50 and above, it is observed that the RMSE values of the methods differ from each other as the RMSE values of the equating methods increase. In general, the lowest RMSE value under all factors is seen in the circular-arc equating method, while Braun/Holland and chained linear equating methods are observed to give RMSE values close to each other. When the RMSE graphs of the subtests with a correlation of 0.80 in Table 1 shows; When the difficulty difference between the equated subtest forms remains constant, the RMSE values of the equating methods decrease as the sample size increases, while the RMSE values increase as the subtest length increases. However, it is observed that the RMSE values of the methods increase when the difficulty difference increases while the subtest lengths are constant. However, it is observed that the RMSE values of the methods increase when the difficulty difference increases while the subtest lengths are constant. In addition, when the subtest length is 10, 15 and 30, the average difficulty level difference is 0.0 and it is seen that the equating methods have the least RMSE values. It is observed that the circular-arc equating method generally gives the lowest RMSE value under all conditions. When the RMSE graph of the subtests with a correlation of 0.90 is examined, similar findings are observed when equating the subtests with a correlation of 0.70 and 0.80. While the lowest RMSE value is seen in the circular-arc equating method, the equating methods have low RMSE values when the subtest length is 10, 15 and 30.

When Table 2, which shows the RMSE values of the methods as a result of the equating performed using the augmented subscores, is examined; while the difficulty level difference is constant in the subtests with a correlation of 0.70, a decrease in the RMSE values of the methods is observed when the sample size increases. However, as the sub-test length increases, an increase is observed in the RMSE values of the methods. While the minimum RMSE values are seen in cases with 10, 15 and 30 subtest lengths, a decrease in RMSE values is observed as the average difficulty level difference decreases. Under all conditions, the lowest RMSE value is generally observed in the circular arc equating method. As a result of the equating of subtests with a correlation of 0.80, it is seen that the RMSE values of the methods decrease as the sample size increases, the RMSE values increase as the difficulty level difference increases, and the RMSE values of the methods increase as the subtest length increases. While the RMSE values of the methods are close to each other with 10, 15 and 30 subtest lengths and difficulty level of 0.0, the lowest RMSE values are observed under these conditions. When equating subtests with a correlation of 0.90, results similar to those obtained when equating subtests with a correlation of 0.70 and 0.80 are observed. While the lowest RMSE value was observed at 10, 15 and 30 subtest lengths and 0.0 average difficulty level, it was seen that the circular-arc equating method gave the lowest RMSE value under all conditions.

When Table 1 and Table 2, obtained as a result of equating with the augmented subtest scores/subscores of the same subtests showing added value, are examined, it is observed that the RMSE values of the methods are higher in Table 2. It is seen that the RMSE values of the methods are higher by using the augmented subtest scores of the tests.

Table 1. The common effect of sample size, subtest length and difficulty level difference factors on the equating error (RMSE) values of equating methods as a result of equating with subtest scores for the first subtests with 0.70, 0.80 and 0.90 correlations

Table 2. The common effect of sample size, subtest length and difficulty level difference factors on the equating error (RMSE) values of equating methods as a result of equating with augmented subtest scores for the first subtests with 0.70, 0.80 and 0.90 correlations

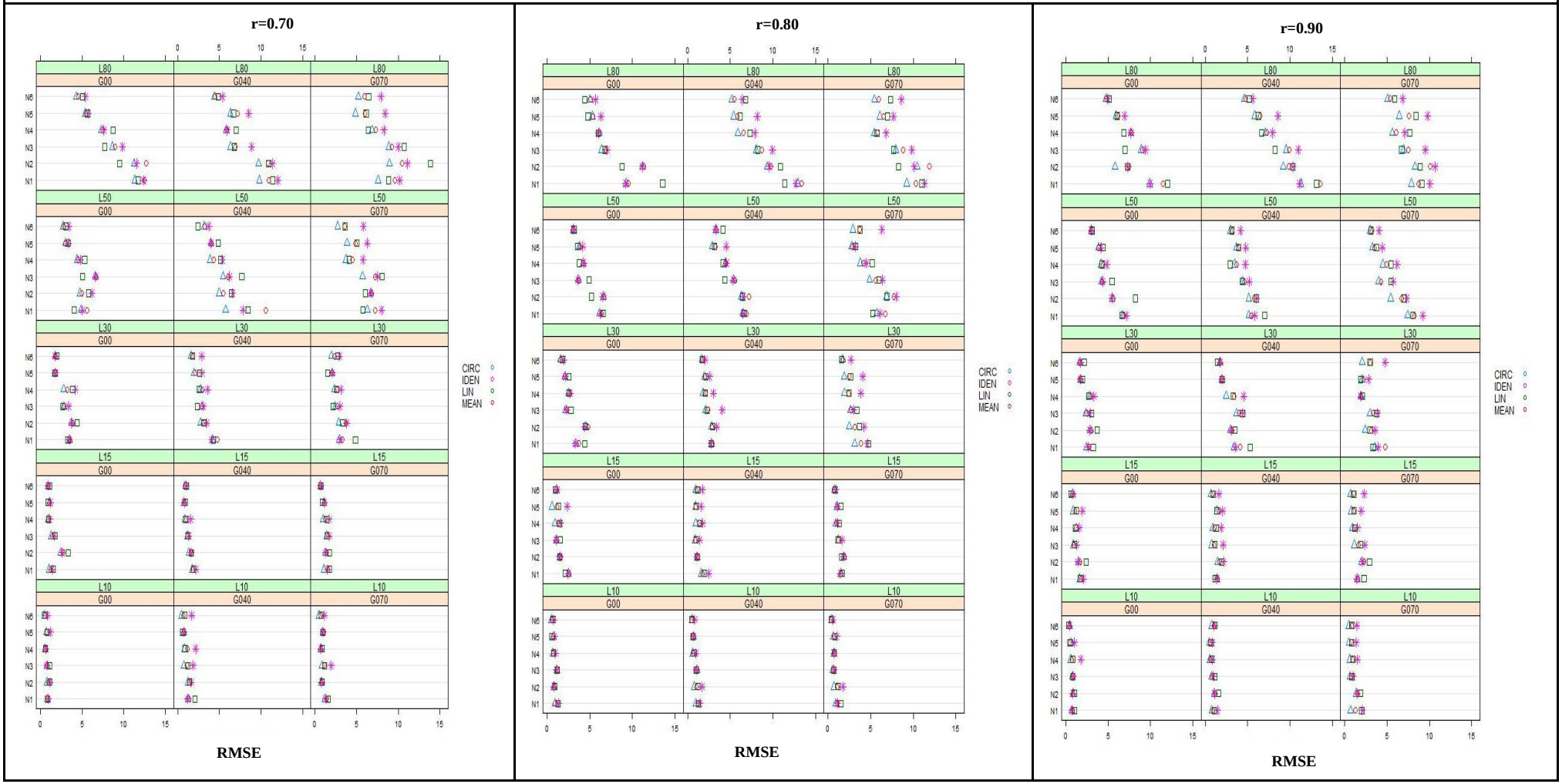


Table 3 shows the equating errors of the methods as a result of the equating performed with the second subtests. The equating errors of subtests with correlation of 0.70, shows, the difficulty levels remain constant between the subtests , the RMSE values of the methods decrease as the sample size increases, while the RMSE values of the equating methods decrease as the subtest length increases. However, while the subtest length is constant, it is observed that the RMSE values decrease as the sample size increases, and the RMSE values of the methods increase as the difficulty difference between the equated test forms increases. It is seen that when the subtest length is 10, 15 and 30 and the average difficulty difference between the test forms decreases, the RMSE values obtained from the equating methods are less and closer to each other. However, when the subtest length is 50 and above, it is observed that the RMSE values of the equating methods increase. In general, the lowest RMSE value under all factors is seen in the circular-arc equating method, while Braun/Holland and linear chained equating methods are observed to give RMSE values close to each other. When the RMSE graphs of the subtests with a correlation of 0.80 in Table 3 are examined; when the average difficulty difference between the equated subtest forms remains constant, the RMSE values of the equating methods decrease as the sample size increases, while the RMSE values increase as the subtest length increases. However, it is observed that the RMSE values of the methods increase when the average difficulty difference increases while the subtest lengths are constant. In addition, when the subtest length is 10, 15 and 30, the average difficulty level difference is 0.0 and it is seen that the equating methods have the least RMSE values. It is observed that the circular-arc equating method generally gives the lowest RMSE value under all conditions. When the RMSE graph of the subtests with a correlation of 0.90 shows, similar findings are observed when equating the subtests with a correlation of 0.70 and 0.80. While the lowest RMSE value is seen in the circular-arc equating method, the equating methods have low RMSE values when the subtest length is 10, 15 and 30.

RMSE values of the methods as a result of the equating performed using the augmented subscores are given in Table 4. The difficulty level difference is constant in the subtests which are a correlation of 0.70, a decrease in the RMSE values of the methods is observed when the sample size increases, while the RMSE values of the methods increase as the subtest length increases. While the minimum RMSE values are seen in cases with 10, 15 and 30 subtest lengths, a decrease in RMSE values is observed as the difficulty level difference decreases. Under all conditions, the lowest RMSE value is generally observed in the circular-arc equating method. As a result of the equating of subtests with a correlation of 0.80, the RMSE values of the methods decrease as the sample size increases, the RMSE values increase as the difficulty level difference increases, and the RMSE values of the methods increase as the subtest length increases. When the RMSE values of the methods are close to each other with 10, 15 and 30 subtest lengths and an average difficulty level of 0.0, the lowest RMSE values are observed under these conditions. When equating subtests with a correlation of 0.90, results similar to those obtained when equating subtests with a correlation of 0.70 and 0.80 are observed. While the lowest

RMSE value was observed at 10, 15 and 30 subtest lengths and 0.0 average difficulty level, it was seen that the circular-arc equating method gave the lowest RMSE value under all conditions.

When Table 3 and Table 4 are examined, it is observed that the RMSE values of the methods are higher in Table 4. It is seen that the RMSE values of the methods are higher by using the augmented subtest scores of the tests.

Table 3. The common effect of sample size, subtest length and difficulty level difference factors on the equating error (RMSE) values of equating methods as a result of equating with subtest scores for the second subtests with 0.70, 0.80 and 0.90 correlations

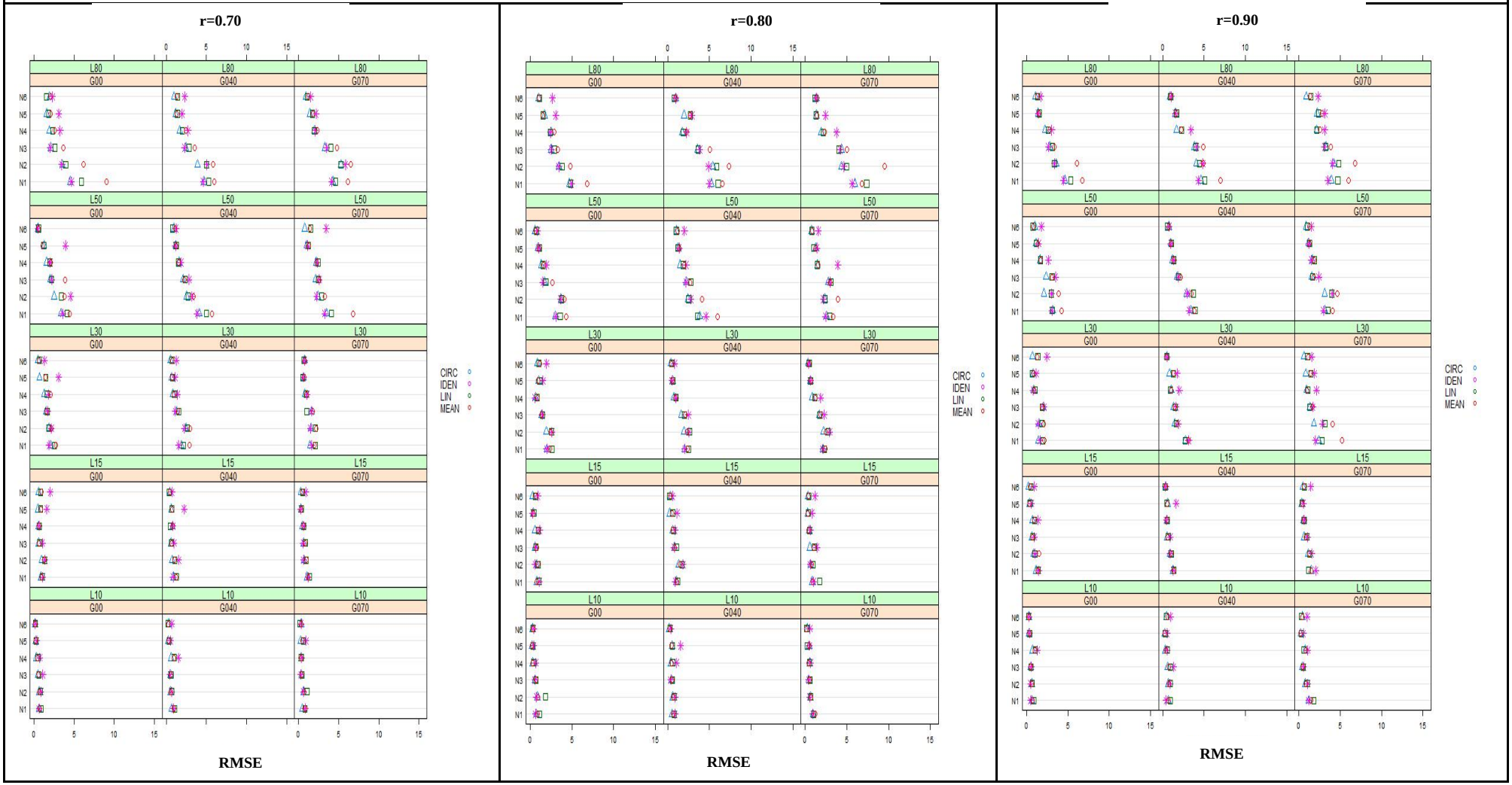
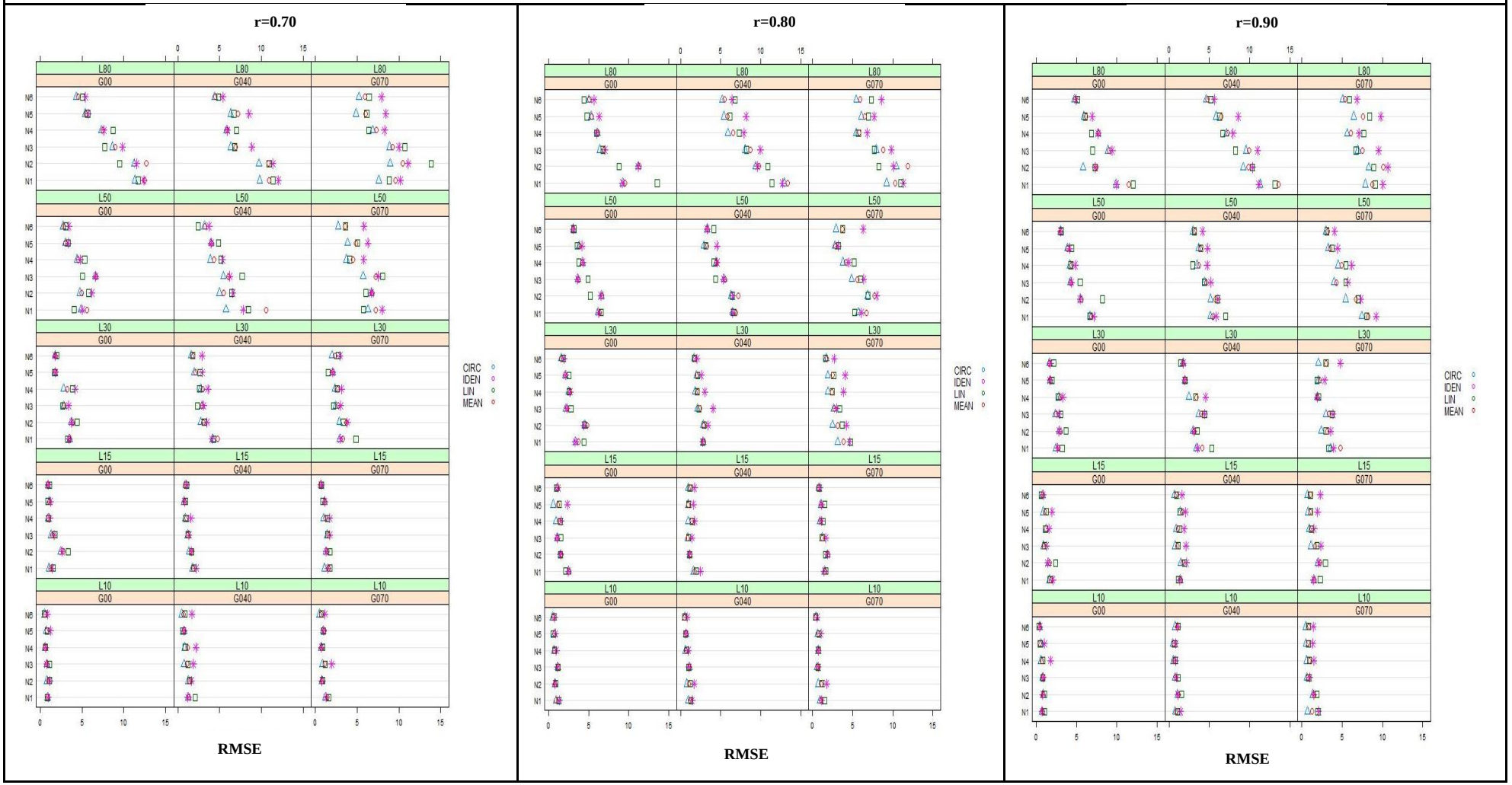


Table 4. The common effect of sample size, subtest length and difficulty level difference factors on the equating error (RMSE) values of equating methods as a result of equating with augmented subtest scores for the second subtests with 0.70, 0.80 and 0.90 correlations



4. DISCUSSION

As a result of the equating made with the first and second subtests, which have added value, a decrease is observed in the RMSE values of the equating methods as the sample size increases. The identity equating method gave the highest RMSE value for sample sizes of 25, 50, 100, 200 and 500. The circular arc equating method gave the lowest RMSE value in all sample sizes. In the second subtest, as the sample size increased, the RMSE value decreased. The identity equating method gave the highest RMSE value for sample sizes of 50, 100, 200 and 500. The lowest RMSE value in all sample sizes was observed in the circular arc equating method. Considering the results obtained as a result of equating the raw subscores and added subscores of the subtests, increasing the sample size causes a decrease in RMSE values. The result obtained is similar to the study of Kim and Livingston (2009).

As a result of the first subtest equating, which has a added value, it is seen that the RMSE values obtained from the equating methods increase with the increase in the length of the test. The circular arc equating method had the lowest RMSE value in all sample sizes. As a result of equating of the second subtests showing added value, the RMSE values of the methods increased as the test length increased. In subtest lengths with 10, 15 and 30 items, the descending order in RMSE values is Braun/Holland, chained linear equating, circular arc equating method. In subtests with 50 and 80 items, the descending order of RMSE values is Braun/Holland, identity equating, linear chained equating and circular arc equating method. As a result of the equating made with the observed scores and augmented scores of the subtests with added value, an increase in the RMSE values of the equating methods was observed as the test length increased.

As a result of equating the first subtests showing added value, no difference occurred as the difficulty level difference in the equating methods increased. As a result of the second subtest equated with added value, no significant change was observed in the RMSE values of other methods except the Braun/Holland method. Considering the equating results with the observed score and the added subscore, as the difficulty level difference between the test forms increased, the RMSE values of the identity equating method increased. This result is parallel to the study of Kim, von Davier and Haberman (2008).

5. RESULTS

In line with the findings obtained as a result of the study, the RMSE values of the equating methods are obtained lower when the subtest length is 10, 15 and 30. When the difficulty level difference between test forms is 0.0, the equating error values of the methods are lower than in other cases. Therefore, increasing the subtest length and increasing difficulty level differences cause an increase in the equating error values of equating methods, while an increase in the sample size causes a decrease in equating errors. At sample sizes of 100, 200 and 500, lower equating error values of the methods are obtained under all conditions. In general, the circular arc equating method has the lowest equating error value under all conditions. The RMSE values of the identity equating are observed to be higher than the

RMSE values of other methods. Therefore, the equating error in other equating methods is less. It is seen that the RMSE values of the equating methods obtained by using augmented subtest scores are higher. Since subtests have added value and the equating methods obtained as a result of equating with subtest scores appear to have less equating error values, it is more appropriate to use subtest scores. However, in the study, data was generated to compare the equating methods. A similar study can be done using real data. In the study, equating was performed using linear equating methods. Equating error of equating methods can be compared by using equipercentile equating methods. Similar study can be performed by differentiating the equating design. In this study, different studies which compare the error values of the equating methods can be conducted where the levels of the variables examined are changed or variables are changed.

REFERENCES

Albano, A.D. (2016). equate: An R Package for Observed-Score Linking and Equating. R package version 2.0-3. URL http://CRAN.R-project.org/package=equate.

Angoff, W. H. (1984). *Scales, norms, and equivalent scores*. Princeton, NJ: Educational Testing Service.

Baykul, Yaşar (2010). *Eğitimde ve Psikolojide Ölçme: Klasik Test Teorisi ve Uygulaması*. Ankara: Pegem Yayınları

Crocker, L. and Algina, J. (1986). *Introduction to Classical and Modern Test Theory*. New York: Holt, Rinehart and Winston.

Chu, K. L. & Kamata, A. (2003). *Test equating with the presence of DIF*. Paper presented at the annual meeting of American Educational Research Association, Chicago.

Haberman, S. J. (2008). When can subscores have value? *Journal of Educational and Behavioral Statistics*, 33, 204-229.

Kan, A. (2010). Test Eşitleme: Aynı Davranışları Ölçen, Farklı Madde Formlarına Sahip Testlerin İstatistiksel Eşitliğinin Sınanması. *Eğitim ve Psikolojide Ölçme ve Değerlendirme Dergisi*, 16-21.

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

Kolen, M.J., & Brennan, R.L. (2004). *Test equating: Methods and practices (2nd ed.)*.New York, NY:Springer-Verlag.

Kolen, M. J. & Brennan R. L. (2014). *Test equating, Scaling, and Linking: Method and Practice (Third ed.)*. New York, NY: Springer.

Livingston, S. A. (2004). *Equating test scores (without IRT)*. Princeton, NJ: ETS.

Livingston, S. A. & Kim, S. (2009). The circle-arc method for equating in small samples. *Journal of Educational Measurement*, 46(3), 330–343.

Sinharay, S., & Haberman, S. J. (2011). Equating of augmented sub- scores. *Journal of Educational Measurement*, 48, 122–145.

Sinharay,S., Haberman, S. & Puhan, G. (2007). *Subscores Based on Classical Test Theory: To Report or Not to Report*. Princeton, NJ: Educational Testing Service.

Sinharay, S., Puhan, G., & Haberman, S. J. (2011). An NCME instructional module on subscores. *Educational Measurement: Issues and Practice*, 30(3), 29–40.

Sinharay, S. (2010a). *When Can Subscores Be Expected To Have Added Value? Results FromOperational and Simulated Data*. Princeton, NJ: Educational Testing Service.

Sinharay, S. (2010b). How Often Do Subscores Have Added Value? Results From Operational and Simulated Data. *Journal of Educational Measurement*, 47(2), 150–174.

Sinharay, S. & Holland, P.W. (2007). Is It Necessary to Make Anchor Tests Mini-Versions of the Tests Being Equated or Can Some Restrictions Be Relaxed? *Journal of Educational Measurement*,

44(3), 249–275.

von Davier, A. A., Holland, P. W., & Thayer, D. T. (2004). *The kernel method of test equating.* New York, NY: Springer-Verlag.

von Davier, A.A. (2010). *Statistical Models For Test Equating, Scaling and Linking.* New York: Springer.

Wang, T. (2006). *Standard errors of equating for equipercentile equating with log-linear pre-moothing using the delta method.* Iowa City: Center for Advanced Studies in Measurement and Assessment (CASMA).

GENİŞLETİLMİŞ TÜRKÇE ÖZET

ARTI DEĞER ÖZELLİĞİNE SAHİP ALT TESTLERDE ÇEŞİTLİ FAKTÖRLER ALTINDA EŞİTLEME HATALARININ KARŞILAŞTIRILMASI

GİRİŞ

Puanların geçerliliği, karşılaştırılabilirliği ve güvenilirliği belirlenmedikçe ve standart alt puanlar için de geçerli olmadıkça, bireyler için puanlar raporlanmamalıdır. Bu özelliklerin sağlanması, alt puanların her amaç için kullanılabilir olmasını garanti etmez. Alt puanlardan elde edilen bilgilerin kullanılabilmesi için alt puanların artı değere sahip olması gerekmektedir. Artı değeri olan alt puanlar, alt puanın toplam puandan bağımsız olarak yorumlanabileceğini düşündürmektedir. Alt puanlar artı değere sahip olduğunda, alt puanlar bireylerin veya kuruluşların profilinin çıkarılmasının yanı sıra bireylerin güçlü ve zayıf yönlerini belirlemek için de kullanılabilir.

Karşılaştırılabilirlik, test eşitleme ile mümkündür, böylece alt puanların karşılaştırılabilirliğini sağlamak için alt testi eşitlemek gerekir. Toplam test puanları yerine alt puanların kullanılıp kullanılamayacağının ve eşitlenip eşitlenemeyeceğinin incelenmesi gerekmektedir. Ayrıca alt test puanları eşitleme süreçlerinde katma değer bulunması durumunda, çeşitli eşitleme yöntemlerinin çeşitli koşullar altında hata göstergelerinin incelenmesi gerekmektedir. Bu çalışma, artı değeri olan veri setleri için test eşitleme işleminden elde edilen eşitleme hatası değerlerinin değerlendirilmesini amaçlamaktadır.

YÖNTEM

Bu araştırmada ortak madde deseni kullanılmıştır. Ankor test maddeleri alt test uzunluğunun %40'ı uzunluğundadır. Eşitleme için birim, zincirleme doğrusal eşitleme, dairesel-yay eşitleme ve Braun/Holland eşitleme yöntemleri kullanılmıştır. R 3.1.1 programında kullanılarak, artı değeri özelliğine sahip olan veriler üretilmiştir. Bu çalışmada X formu ve Y formu için iki parametreli lojistik modele (2PLM) göre ikili veriler üretilmiştir. Her test formunda iki alt test bulunmaktadır. Ankor testinin de iki alt testi vardı. Her iki form için de alt testler arasındaki korelasyon (0,70, 0,80 ve 0,90) değişimlenmiştir. Ayrıca X ve Y formlarının alt testleri arasındaki ortalama güçlük farkı üç düzeyde (0,0, 0,4 ve 0,7) değişimlenmiştir. Simüle edilen formlar birim, zincirleme doğrusal, Braun/Holland ve dairesel-yay yöntemleri kullanılarak altı farklı örneklem büyüklüğü (20, 25, 50, 100, 200 ve 500) için 100 tekrarlamayla eşitlenmiştir. Alt test puanları (ham alt puanlar ve ham alt puanlar, genişletilmiş alt puanlar ve genişletilmiş alt puanlar) kullanılarak eşitleme yöntemlerinden elde edilen eşitleme hatası değerleri olan RMSE (eşitleme hatası) elde etmek için 100 tekrara dayalı yineleme işlemi gerçekleştirilmiştir. R 3.1.1 programındaki "equate" paketi kullanılarak alt testler eşitlenmiştir. RMSE (eşitleme hatası) değerlerine ait ortak etki grafiği R 3.1.1 programı kullanılarak çizilmiştir.

BULGULAR

Korelasyonu 0,70 alt testlerin eşitlenmesi sonucunda genel olarak tüm faktörler altında en düşük RMSE değeri dairesel yay eşitleme yönteminde görülürken Braun/Holland ve zincirlenmiş lineer eşitleme yöntemlerinin birbirine yakın RMSE (eşitleme hatası) değerleri verdiği görülmüştür. Korelasyonu 0,80 olan alt testlerin RMSE (eşitleme hatası) grafikleri incelendiğinde; genel olarak dairesel yay eşitleme yönteminin tüm koşullar altında en düşük RMSE (eşitleme hatası) değerini verdiği gözlenmiştir. Korelasyonu 0,90 olan alt testlerin RMSE (eşitleme hatası) grafiği incelendiğinde en düşük RMSE (eşitleme hatası) değeri dairesel yay eşitleme yönteminde elde edilmiştir. Genişletilmiş alt puanlar kullanılarak yapılan eşitleme sonucunda alt testlerde tüm koşullar altında 0,70 korelasyon ile en düşük RMSE değeri genel olarak dairesel yay eşitleme yönteminde görülmüştür. Alt testlerin 0,80 korelasyonla eşitlenmesi sonucunda örneklem büyüklüğü arttıkça yöntemlerin RMSE (eşitleme hatası) değerlerinin düştüğü elde edilmiştir. Alt testler 0,90 korelasyonla eşitlendiğinde dairesel yay eşitleme yönteminin tüm koşullar altında en düşük RMSE (eşitleme hatası) değerini verdiği gözlenmiştir.

İkinci alt testlerle yapılan eşitlemede; 0,70 korelasyona sahip alt testlerin eşitlenmesi sonucunda tüm faktörler altında en düşük RMSE (eşitleme hatası) değerinin dairesel yay eşitleme yönteminde, Braun/Holland ve zincirlenmiş lineer eşitleme yöntemlerinin ise birbirine yakın RMSE (eşitleme hatası) değerleri verdiği görülmüştür. Genel olarak dairesel yay eşitleme yönteminin tüm koşullar altında en düşük RMSE (eşitleme hatası) değerini verdiği gözlenmiştir. Genişletilmiş alt puanlar kullanılarak yapılan eşitlemede; korelasyonu 0,70 olan alt testlerde tüm koşullar altında en düşük RMSE (eşitleme hatası) değeri genel olarak dairesel yay eşitleme yönteminde elde edilmiştir. Korelasyonun 0.80 olduğu alt testlerin eşitlenmesi sonucunda, örneklem büyüklüğü arttıkça yöntemlerin RMSE (eşitleme hatası) değerlerinin düştüğü, ortalama zorluk düzeyi farkı arttıkça RMSE (eşitleme hatası) değerlerinin arttığı, alt test uzunluğu arttıkça yöntemlerin RMSE (eşitleme hatası) değerlerinin ise arttığı görülmüştür. Korelasyonun 0,90 olduğu alt testlerde tüm koşullar altında en düşük RMSE (eşitleme hatası) değeri genel olarak dairesel yay eşitleme yönteminde elde edilmiştir.

SONUÇ

Genel olarak dairesel yay eşitleme yöntemi tüm koşullar altında en düşük eşitleme hatası değerine sahip sonuçlar üretmiştir. Birim eşitleme yönteminin RMSE değerleri zincirlenmiş lineer eşitleme, dairesel yay eşitleme ve Braun/Holland eşitleme yöntemlerinin RMSE değerlerinden daha yüksek olduğu görülmüştür. Bir diğer ifade ile zincirlenmiş lineer eşitleme, dairesel yay eşitleme ve Braun/Holland eşitleme yöntemlerinin eşitleme hatası birim eşitleme yönteminden elde edilen eşitleme hatasından daha düşük değere sahiptir. Genişletilmiş alt test puanları kullanılarak gerçekleştirilen eşitlemelerde elde edilen RMSE değerlerinin diğer puanlar kullanılarak gerçekleştirilen eşitlemelerde eşitleme yöntemlerinin daha yüksek RMSE değerleri üretmiştir. Sonuç olarak, alt testlerin artı değer özelliğine sahip olması ve alt test puanlarıyla eşitleme sonucunda elde edilen eşitleme yöntemlerinin eşitleme hata değerlerinin daha az olması nedeniyle alt test puanlarının kullanılması daha uygundur.