ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2020; 11(4); 362-373

The Impact of Covariate Variables on Kernel Equating under the Non-equivalent Groups

Çiğdem AKIN ARIKAN *

Abstract
This study aims to use covariate variables correlated with the test scores instead of common items for non-equivalent groups with covariates (NEC) design in kernel equating. This study used the 2016 Monitoring and Evaluation of Academic Skills Project in Turkey. The study used data from 6,000 students, randomly selected from the Turkish Ministry of National Education's current student data. Three thousand of the students took form A, and 3,000 of them took form B. The data include mathematics test scores and consist of 18 items, nine of which are the first items, and nine of which are anchor items. The equated scores from the NEC design were compared with equated scores from the non-equivalent group (NEAT) design. From the equating results, the root mean squared difference (RMSD) and standard error of equating (SEE) values were calculated. The results showed that NEC design could produce lower standard errors compared with the NEAT design, and the least RMSD was provided by NEAT PSE methods and NEC methods. The general result of this research is that test forms can be equated using covariates when there are no anchor items.

Key Words: NEC design, NEAT design, covariate variables, SEE, RMSD.

INTRODUCTION

In the last century, new equating methods and designs have been developed in test equating. One of these methods, kernel equating, was first defined by Holland and Thayer (1989) and then developed by von Davier, Holland, and Thayer (2004). Kernel equating is an equipercentile score equating technique in which discrete score probabilities are continuized, and score probabilities are equated (von Davier et al., 2006). In this regard, Kernel equating can be considered a developed form of traditional equating techniques. There are two reasons for this view. First, it makes data consistent by using presmoothing, yields smaller errors when compared to other methods by smoothing the transformation of data, and is less dependent on sample variability. Second, kernel equating can be applied to all designs and equating functions (von Davier et al., 2004). Kernel equating consists of five steps, namely, presmoothing, estimation of score probabilities, continuization, equating, and calculating the standard error of equating. Also, in kernel equating, both linear and equipercentile equating functions are used (von Davier et al., 2006).

In test equating, there are various group designs, such as single group design, equivalent group design, and nonequivalent groups with anchor test (NEAT) (Kolen & Brennan, 2014; von Davier et al., 2004). NEAT design is one of the most frequently used designs in the literature. Post-stratification equating (PSE) and chained equating (CE) that were used in this study and Levine observed-score linear equating methods are within the scope of NEAT design in kernel equating (von Davier et al., 2004). Two different test forms, namely, X and Y, in addition to the anchor test A, are taken by two different populations in NEAT design. For a detailed theoretical explanation of all methods, readers are encouraged to look at Chen and Holland (2010), von Davier et al. (2004), and von Davier, Fournier-Zajac, and Holland (2007). To estimate the distribution of X in group I and the distribution of Y in group II, the anchor test A is used by PSE. In this regard, the conditional distribution of X, given A, and the conditional distribution of Y, given A, constitute the population invariant. Afterward, it post-stratifies the distributions of both X and Y in a target population T (a synthetic population of Group I

* Assist. Prof., Ordu University, Faculty of Education, Ordu-Turkey, akincgdm@gmail.com, ORCID ID: 0000-0001-5255-8792

To cite this article:
Akın-Arıkan, Ç. (2020). The impact of covariate variables on kernel equating under the non-equivalent group design. *Journal of Measurement and Evaluation in Education and Psychology, 11*(4), 362-373. doi: 10.21031/epod.706835.
Received: 20.03.2020
Accepted: 15.11.2020

and Group II). In CE, the anchor is used as a part of a chain by linking X to A in group I and then A to Y in group II (von Davier et al., 2004).

In NEAT design, anchor items are used to adjust the differences in ability between the groups. However, anchor items might not appear in the forms of all the test programs or standardized tests. Additionally, test forms might not be equated since it is hard for groups that take different test forms to be equivalent in practice. For instance, if there are no anchor items in non-equivalent groups, significant covariates can be used instead of anchor items and the design is called the non-equivalent groups with covariates (NEC) design (Wiberg & Branberg, 2015). Wiberg and Branberg (2015) also used NEATNEC design, which is a mixture of the NEC design and the NEAT design in their research. NEC design is used in the post-stratification equating method (NEATPSE) of kernel equating and the populations of two groups are weighted to generate a synthetic population to equating the test scores (Andersson & Wiberg, 2017; von Davier et al., 2004). In fact, it is assumed that enhancing the correlation of the covariances used in NEC design with the test will result in similar numbers to that of NEAT CE and NEAT PSE (Wiberg & Branberg, 2015). When the literature was examined, it was seen that different variables (e.g. test scores and gender) were used as covariates (e.g. Branberg & Wiberg, 2011; Wiberg & Branberg, 2015; Yurtçu, 2018). The basic assumption of the NEC design is that these covariates can be used for the ability variability between two groups (Wiberg & Branberg, 2015). In the NEC design, the other critical point is this: for both groups the conditional distribution of the test scores with the covariates is the same (Wiberg & Branberg, 2015). For this assumption, the time, equating between test forms plays a critical role in the results. Therefore, bias can be avoided by adding a variable that affects its change over time to the equating model. (Wiberg & Branberg, 2015).

The literature about covariates in equating revealed that the number of studies are limited. First studies used different variables in test equating, paving the way for future studies to be conducted with covariates (e.g. Cook, Eignor, & Schmitt, 1990; Holland, Dorans, & Petersen 2007; Livingston, Dorans, & Wright, 1990). As for recent studies, Branberg (2010) equated test forms with covariates and claimed that it is possible to use covariates instead of anchor items. Branberg and Wiberg (2011), first of all, conducted a regression analysis between the test scores and variables to determine the covariates, which were education and gender in the real data. The study results showed that by correcting for variations in the test score distributions of covariates, test equating could be improved. Similarly, Wiberg and Branberg (2015) concluded that NEC design is more accurate than the equivalent group design in kernel equating. In line with this conclusion, using both covariates and anchor items resulted in the smallest standard error of equating over a large range of test scores. The research conducted by Gonzalez et al. (2015) revealed that the Bayesian non-parametric model for equating makes many assumptions that used to be vital for test equating unnecessary, demonstrating that even when there is no covariant, equating is possible. Wiberg and von Davier (2017) also stated the effect of covariates in various administrations would aid in the process of ensuring equal testing for test-takers.

Wallin and Wiberg (2017) investigated the manner in which propensity scores affect equation results when covariates are great in number by comparing the kernel equating methods in NEAT and NEC design. Their research indicated that using propensity scores in kernel post-stratification and kernel chained equating methods increases precision and leads to greater results compared to the equivalent group designs. Moreover, the research showed great similarities with the results of the anchor test design. Yurtçu (2018) used covariates to obtain scores equated by using non-parametric Bayes techniques. According to the research, this model is more informative compared to the traditional methods. Also, covariants can be used instead of anchor items and in some cases, this model has been found to give more accurate results. Likewise, the equated scores obtained through this model were closer to the target. The limited number of studies on the topic indicates a gap in the literature. Likewise, it is necessary to conduct Item Response Theory (IRT) studies, as well as testing kernel equating methods, which are new approaches to the topic. In Turkey, large-scale standardized tests generally are used when making important decisions, such as the selection and placement of students in any kind of educational program. The Monitoring and Evaluation of Academic Skills Study (ABIDE) of the Republic of Turkey Ministry of National Education (MoNE) in Turkey uses large-scale testing to assess the students' mental skills in topics such as math, science, and social studies

(MoNE, 2016). With the exception of the Monitoring and Evaluation of Academic Skills Project, which contains anchor items, all the other exams in Turkey lack anchor items.

The Purpose of the Study

Exams, such as language exams, academic personnel, and postgraduate education entrance exams, have various validity periods so that results from different years can be used to apply for a master's degree, Ph.D. degree, research assistant role, teaching assignments, etc. The fact that the test scores are comparable and interchangeable brings forward the topic of equating test forms. Through the NEC design, equating the test forms will be possible, even in cases where there are no anchor items. Furthermore, designing these sorts of tests will be a guide to equating methods and determining precautions to be taken in case of the existence or non-existence of anchor items. Accordingly, the purpose of the study is to compare the results obtained through the PSE and CE equating methods, which are among the NEAT and NEC kernel equating method designs, and to determine the effect of gender and socioeconomic level as covariates in test equating.

METHOD

In the context of the study, scores obtained from math sub-tests in the 2016 ABIDE were equated by using NEAT and NEC designs with kernel chained equipercentile, kernel post-stratification equipercentile, kernel chained linear, and kernel post-stratification linear methods. Seeing that the existing methods and techniques were verified through real data, it is possible to claim that the research is descriptive.

Sample

The population of the study comprised 38,000 students, from 16,118 schools and 48,091 branches, which were accessed via the ABIDE, with an approximate number of 400 students per city in Turkey (MoNE, 2016). In the scope of the study, data for 6000 students were used, randomly selected from the current student data of MoNE, 3000 of whom took form A, and 3000 of whom took form B. Among the students, 1292 (43.07%) who took form A were female, while 1708 (56.93%) of them were male. Of those completing form B, 1518 (50.6%) were female, whereas 1482 (49.4%) were male.

Process and Data Collection Instruments

Research data consist of math test scores for eighth-graders as a part of the ABIDE. There are 20 items in the math sub-test of the ABIDE. Nine of these items are primary items; nine of them are anchor items, and two of them are pilot items. Furthermore, the project consists of three forms (A, B, and C) and 12 booklets (A1-A4, B1-B4, C1-C4). Each form of the project consists of nine primary items. Form A is connected to form B and C with nine items, while form B is connected to form C with nine items. To put it in a different way, primary items exist in the booklets of forms A, B, and C. Booklet 1 in Form A and Form B were used for this study, and each booklet consists of 18 items, including anchor items. However, the anchor items were determined as external anchors and were not included in the total score. There are partially scored items in the booklet. Among the partially scored items, category 2 was recoded as 1 and was transformed into 1-0 data.

To determine the covariates, the correlation values between math success and student variables in the questionnaire were analyzed in a report prepared for the ABIDE. The socioeconomic index, which had the highest correlation value (r = .37, p < .05), was chosen as the anchor variable (MoNE, 2016). To categorize the socioeconomic level index, which is a continuous variable, three levels: low, middle, and high, were created through a two-step cluster analysis. Another covariant of the study was gender. Socioeconomic status (SES) was coded *low* = 1, *middle* = 2, *high* = 3, whereas gender was coded as *female* = 1 and *male* = 2.

For booklet A, the correlation values of the test and variables were as follows: the correlation of the anchor test and the test is (r = .51, p < .05); the correlation of the test and the socioeconomic variable is (r = .21, p < .05); the correlation of the categorical socioeconomic index and the test is (r = .19, p < .05); and the correlation of gender and the test is (r = .05, p > .05). As for booklet B, the correlation values were as follows: the correlation of the anchor test and the test is (r = .49, p < .05); the correlation of the test and the socioeconomic variable is (r = .16, p < .05); the correlation of the categorical socioeconomic index and the test is (r = .15, p < .05); and the correlation of gender and the test is (r = .02, p > .05). In other words, the relationship between the test and gender is insignificant for the two booklets, but the other relationships are significant. Although the relationship between the test and gender is not significant, it was used in other studies (Branberg & Wiberg, 2011; Gonzalez et al., 2015; Liou et al., 2001, and Yurtçu, 2018). It was also taken as a covariate in this study.

Data Analysis

Firstly, the test scores were equated with a NEAT design in kernel equating. Afterward, gender and socioeconomic level variables were set as covariates, and then the test scores were equated with an NEC design. To equate the tests, the R (R Core Team, 2013) package "kequate" was employed (Andersson et al. 2013). The standard error of equating (SEE) and error of equating (RMSD-root mean squared deviation/error) were used for the evaluation.

RESULTS

Before equating the tests, descriptive statistics for booklets A and B were obtained. The findings are listed in Table 1.

Table 1. Descriptive Statistics of the Booklets

Statistics	A-main test	A-anchor test	B-main test	B-anchor test
N	3000	3000	3000	3000
Mean	2.59	3.52	2.91	2.47
Standard Deviation	1.60	1.85	1.79	1.68
Skewness	0.48	0.38	0.51	0.54
Kurtosis	-0.14	-0.23	-0.02	-0.11

According to Table 1, the score distribution of booklets A, B, and anchor tests are right-skewed. As the kurtosis coefficients of the score distribution are negative, it can be argued that the distribution is platykurtic (negative kurtosis). The skewness and kurtosis coefficients are between -1.00 and +1.00, so the data indicates normal distribution. Additionally, the mean of the tests is revealed to be low.

Booklet A of the subtest on the Monitoring and Evaluation of Academic Skills Project was equated to Booklet B with PSE-EQ (EQ-equipercentile), PSE-L (L-linear), CE-EQ (equipercentile), CE-L (linear), NEC-EQ SEX (gender), NEC-L SEX, NEC-EQ SES (socioeconomic status), NEC-L SES, NEC-EQ-SEX-SES, and NEC-L-SEX-SES. Equated scores are listed in Table 2.

In kernel equating, the selection of bandwidths is important. If a large bandwidth is used, the equating function gets close to the linear equation, whereas in turn, if a small bandwidth is used, the equating function gets close to the linear equation (von Davier et al., 2004).The results obtained with bandwidths (h) were h_X = 0.51 and h_Y = 0.54 for NEC-EQ (SES), h_X = 1697.17 and h_Y = 1592.48 for NEC-L(SES) h_X = 0.51 and h_Y = 0.54 for NEC-EQ (SEX), h_X = 1681.76 and h_Y = 1601.33 for NEC-L(SEX), h_X = 0.51 h_Y = 01544.18 NEC-EQ (SEX-SES), h_X = 0.51 and h_Y = 0.53 for NEAT PSE-EQ, h_X = 1701.38 and h_Y = 01589.34 for NEAT PSE-L, h_X = 0.51 and h_Y = 0.53 for NEAT-CE EQ, and h_X = 1682.62 and h_Y = 1601.04 for NEAT-CE L.

Table 2. Equated Scores Derived from Different Methods

Booklet A	NEC-SEX-EQ	NEC-SEX-L	NEC-SES-EQ	NEC-SES-L	NEC-SEX-SES-EQ	NEC-SEX-SES-L	NEAT-PSE	NEAT-PSE-L	NEAT-CE	NEAT-CE-L
0	0.18	0.24	0.14	0.20	0.16	0.21	0.10	0.17	0.02	0.04
1	1.23	1.20	1.16	1.14	1.20	1.17	1.12	1.11	0.95	0.93
2	2.18	2.15	2.10	2.08	2.14	2.12	2.07	2.04	1.83	1.83
3	3.11	3.10	3.02	3.01	3.07	3.07	3.00	2.97	2.72	2.73
4	4.04	4.05	3.94	3.95	4.00	4.03	3.91	3.91	3.62	3.63
5	4.97	5.01	4.86	4.89	4.95	4.98	4.81	4.84	4.56	4.52
6	5.92	5.96	5.79	5.83	5.92	5.94	5.71	5.78	5.51	5.42
7	6.89	6.91	6.75	6.77	6.94	6.89	6.61	6.71	6.44	6.32
8	7.89	7.86	7.75	7.71	7.98	7.84	7.53	7.64	7.37	7.22
9	8.92	8.81	8.82	8.64	9.00	8.80	8.55	8.58	8.40	8.11

Raw scores for booklet A were between 0.00-9.00, and the equated scores, based on different equation designs and methods, were in the raw score range. Scores that were equated via NEC-EQ with covariate SEX, NEC-L, and NEC-EQ with two covariates were larger than the raw scores for booklet A in the 0.00-4.00 range and smaller than the raw scores in the 5.00-9.00 range. The scores equated via NEC-L with sex as the covariate were bigger than the raw scores for booklet A in the 0.00-5.00 range and smaller than the raw scores in the 6.00-9.00 range; the scores equated via NEAT PSE, NEC-EQ with covariate SES, and NEC-L with covariate SES, were larger than the raw scores for booklet A in 0.00-3.00 range and smaller than the raw scores in 4.00-9.00 range. The scores equated via NEAT CE were larger than the raw scores for booklet A at 0 and smaller than the raw scores in the 1.00-9.00 range. According to the findings, it is possible to claim that the degree of difficulty is not the same throughout the scale, and it changes depending on the forms. The difference between the equated scores and raw scores are given in Figure 1.

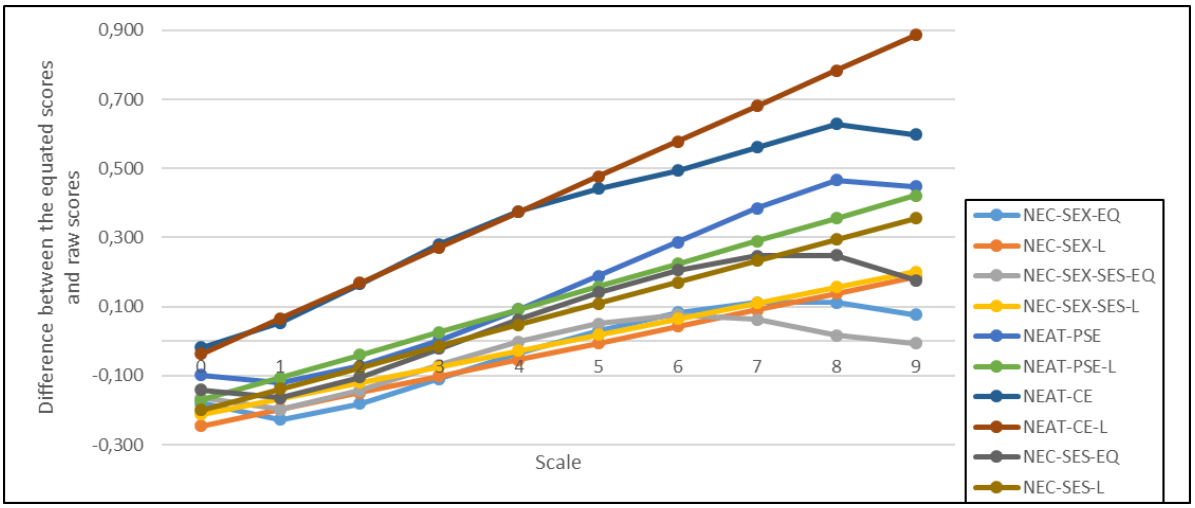

Figure 1. Difference Between Equated Scores and Raw Scores for NEC, NEAT PSE Equipercentile, NEAT PSE Linear, NEAT Chain Equipercentile, NEAT Chain Linear (*NEAT = non-equivalent groups with anchor test; PSE = poststratification; CE= Chain Equating equating; NEC = non-equivalent groups with covariates; EQ= equipercentile and L= linear, SES: socioeconomic status, SEX: Gender*)

In Figure 1, the raw scores and differences between the equated scores for the different equating methods and different equating data designs are displayed. For the NEAT design, PSE and CE with linear and equipercentile types; for the NEC design, equipercentile and linear with two covariates were used separately. The results show that the difference between equated scores and raw scores were smaller in cases where the NEC design was used than those of NEAT design. For linear equating, NEC design with sex and NEC design with sex-SES covariates gave similar results, whereas equipercentile equating yielded somewhat similar results, except for scores between 7-9. The NEC design (linear and

equipercentile) with SES covariate gave similar results, except for scores between 8-9. The greatest gap between the scores equated with the NEC design and raw scores occurred when gender was used as a covariant. Also, NEAT CE methods (EQ and L) had large differences compared to other methods. Figure 2 shows the SEE values for the equating methods.

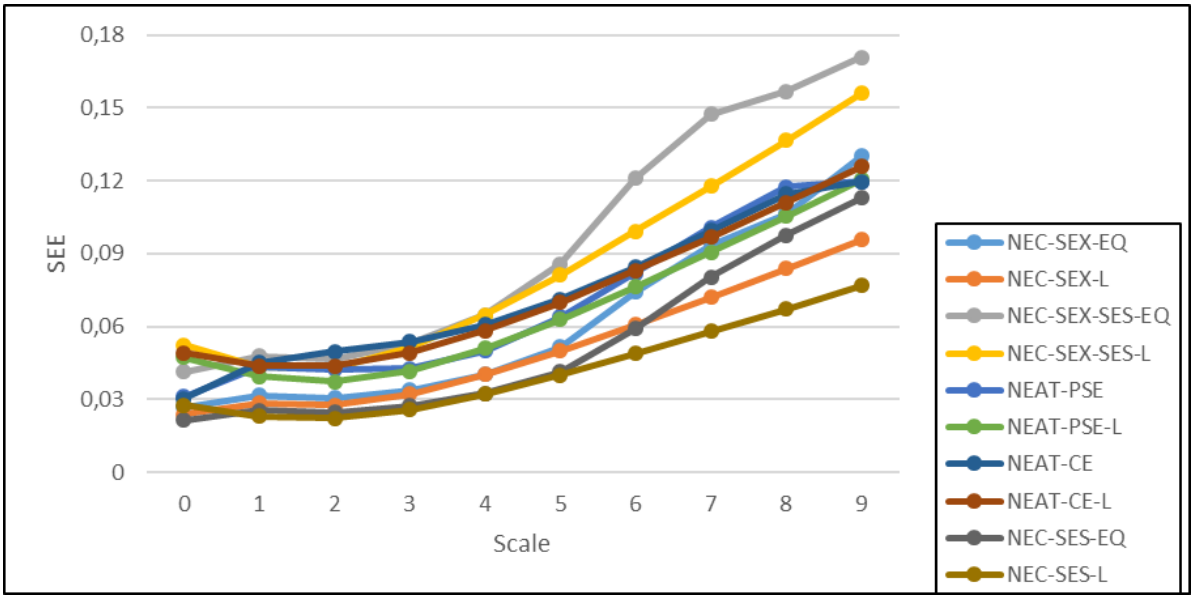

Figure 2. SEE Values for NEC, NEAT PSE Equipercentile, NEAT PSE Linear, NEAT Chain Equipercentile, NEAT Chain Linear (*NEAT = non-equivalent groups with anchor test; PSE = poststratification; CE= Chain Equating; NEC = non-equivalent groups with covariates; EQ= equipercentile and L= linear, SES: socioeconomic status, SEX: Gender*)

When there are few test-takers with very low results and very few with the highest result, the SEE should be larger at the lower end of the scale and at the upper end of the scale. But Figure 2 shows that this is not the case. At the upper end of the scale, SEE is large, while at the lower end of the scale, SEE is quite small. The fact that a lot of test-takers had low results could be the reason for this. Inspection of the SEE results for equating methods, NEAT PSE EQ, NEAT PSE L, NEAT CE L, NEAT CE EQ, NEC EQ with gender covariate gave similar SEE values. Also, all these methods yielded somewhat similar SEE results between 0-4. The SEE values were highest for NEC, with two covariates between 5-9. Another outstanding detail was that NEC with SES covariate (linear) gave the lowest SEE values throughout the score scale, and NEC with SES covariate (equipercentile) gave the lowest SEE values between 0-5. An RMSD coefficient was calculated to evaluate the random error involved in the equating methods. The resulting coefficients are given in Table 3.

Table 3 reveals that equal RMSD coefficients exist in scores equated with the NEAT PSE EQ, NEC SES EQ, and the NEC-SEX-SES-L methods. The smallest RMSD (0.10, 0.11, 0.12, and 0.13) coefficients were obtained from scores equated with the NEAT PSE, NEAT PSE EQ, NEC SES L, NEC SES EQ, and the NEC-SEX-SES-L method, while the largest RMSD coefficients were obtained through KE CE equating methods. It can be inferred that the maximum random error was provided by chained equating methods, whereas the least random error was yielded by NEAT KE PSE methods and NEC SES methods.

Table 3. RMSD Coefficient According to Equating Methods

Equating methods	RMSD
NEC-SEX-EQ	0.15
NEC-SEX-L	0.14
NEC-SEX-SES-EQ	0.13
NEC-SEX-SES-L	0.12
NEC-SES-EQ	0.12
NEC-SES-L	0.11
NEAT-PSE	0.12
NEAT-PSE L	0.10
NEAT-CE	0.27
NEAT-CE -L	0.28

DISCUSSION and CONCLUSION

In this research, the test forms were equated with the kernel-equating methods under the NEAT and NEC designs, and the equating results were compared according to SEE and RMSD coefficients. For the NEC design, the gender variable and socioeconomic index were used as covariates. After separately adding the covariates to the design, two covariates were added together, resulting in three different NEC designs. Equated scores obtained with kernel linear and kernel equipercentile equating techniques are in the raw scores range (0-9). The greatest gap between the raw scores and equated scores was seen in the NEAT CE methods, while the results of the other techniques were relatively similar to each other. The gap between the raw scores and equated scores was obtained smaller in NEC design, and scores obtained with NEAT PSE and NEC designs were similar to each other, as the PSE technique was used in the NEC design. This finding is consistent with the claims of Wiberg and Branberg (2015).

An inspection of the standard errors of the equating methods reveals that in the 0-4 range, standard errors of the methods are relatively similar and close; nevertheless, towards the middle and tail, the NEC-SES-L, NEC-SES-EQ, and NEC-SEX L equating methods show less standardized error. Conversely, the greatest standardized error is seen when the NEC design is used with two covariates. In their research, Wiberg and Branberg (2015) stated that NEC design shows greater standardized error compared to the NEAT design in the middle scale score range, while NEC, NEATCE, NEAT PSE, and NEATNEC techniques show similar SEE values throughout the score scale.

In this research, SEE values were relatively similar in the 0-4 score scale, while SEE values differed depending on the techniques for the 5-9 range. It is possible to state that the findings of the research are partially inconsistent with the findings of Wiberg and Branberg (2015). Sansivieri and Wiberg (2016) ascertained that using anchor test with covariates lessens the standard error in IRT-based tests equating with equivalent groups and NEAT design. It is possible to claim that this finding is consistent with the research of Sansivieri and Wiberg (2016). In kernel equating, the SEE values in the lower and upper parts of the score scale are generally higher compared to the middle part (von Davier et al., 2004; Wiberg & Branberg, 2015). However, in this research, the standard error was lesser at the lower tail of the score scale. This contrast is possibly caused by the high number of low scores. Branberg and Wiberg (2011) ascertained the fact that using covariates increases accuracy and decreases the standard error of the equation. However, it was revealed that the difference between covariant equating and using anchor items was small. This research revealed that for the NEAT and NEC designs, standard errors are similar in tail scores, while using only two covariates results in an increase in the standard error. Wiberg and Branberg (2015) stated that using more than one covariant causes increases in the sparse data for some cells. So, it is important to limit the number of categories, especially when using continuous variables as covariates. The cause of the increase in the standard error could be traced to the fact that there was sparsity in some cells, or the socioeconomic index being a continuous variable and not subcategorized meaningfully. The decrease in the sample number for the socioeconomic category within the gender category could be another reason for the increase in the standard error. Wallin and Wiberg (2017) suggested using propensity to avoid the problem of decline in the observation number for each category in NEC design.

Another finding of the research is that standard errors of linear equating are lower than those of equipercentile equating. This finding is consistent with research by Choi (2009), Liou, Cheng, and Johnson (1997), Mao (2006), Akın-Arıkan, and Gelbal (2018). The main reason for this is that the large h parameter value reduces the standard error. Additionally, this research revealed that random errors in NEAT CE methods are higher than other methods. The errors in NEAT PSE and NEC designs are also partially similar to each other. Usage of the PSE techniques in NEC design caused the similarity of the random errors. Also, a comparison of the techniques in NEC design indicates that using the socioeconomic level variable as the covariant leads to the lowest error value, whereas using the gender variable as the covariant causes the highest error value. The reason for this is the relationship between the covariant and the test. The correlation value of the gender variable and the test is statistically insignificant, where the SES variable has a significant low-level correlation. Despite higher correlation values between the anchor test and the test, the SES variable was able to define groups like anchor items. When the gender variable is used as the covariant, the error was high; however, adding the SES covariant to the gender variable lowered the error rate. Yurtçu (2018) argues that using two covariates is more effective than using anchor items in studies where researchers used covariates in equating.

The general result of this research is that test forms can be equated using covariates when there are no anchor items. Additionally, anchor tests might not be sufficient if the ability difference among the groups is high, the difficulty difference of the test forms is excessive, and the anchor tests are weak (Albano & Wiberg, 2019). Covariates could be used in such cases. Branberg (2010) states that covariates could be used instead of anchor items. Gonzalez et al. (2015) and Yurtçu (2018) used the Bayesian non-parametric model of covariates and stated that equating is possible even in cases where there are no anchor items. At this point, the ability of the covariant to explain the differences among the groups is critical. For this reason, inspecting the correlations and test scores is vital for the determination of the covariates (Branberg & Wiberg, 2011; Liou et al., 2001; Wiberg, 2015; Wiberg & Branberg, 2015).

Among the large-scale exams in Turkey, only the ABIDE has anchor items. Thanks to this study, it became obvious that the test forms could be used in cases where there are no anchor items, resulting in similar findings. To equate the scores of exams with more than one year of validity period, such as academic personnel exams, postgraduate education entrance exams, and public personnel selection exams, test forms must be equated with an equivalent group/random group design. However, in cases of equivalent groups, test forms could be equated with a covariant, as it is difficult to provide conditions for the groups to be equal. Similar research could be conducted comparing equivalent group design with NEC design and for different subtests of the ABIDE, such as Turkish language tests, science tests, etc. Moreover, Bayesian non-parametric models and kernel equating technique results can be compared.

Acknowledgements

The permission for the data was granted from the Ministry of National Education (Dated 24.12.2018 and No. 57750415-605.01-E.24790203).

REFERENCES

Akın-Arıkan, Ç., & Gelbal, S. (2018). A comparison of traditional and kernel equating methods. *International Journal of Assessment Tools in Education*, *5*(3), 417-427. doi: 10.21449/ijate.409826

Albano, A. D., & Wiberg, M. (2019). Linking with external covariates: examining accuracy by anchor type, test length, ability difference, and sample size. *Applied psychological measurement*, *43*(8), 597-610. doi: 10.1177/0146621618824855

Andersson, B., & Wiberg, M. (2017). Item response theory observed-score kernel equating. *Psychometrica*, *82*(1), 48-66. doi: 10.1007/s11336-016-9528-7

Andersson, B., Branberg, K., & Wiberg, M. (2013). Performing the kernel method of test equating with the package kequate. *Journal of Statistical Software*, *55*(6), 1-25. Retrieved from https://www.jstatsoft.org/article/view/v055i06

Branberg, B. (2010). *Observed score equating with covariates* (Statistical Studies No. 41). Umea: Umea Unıversity, Department of Statistics. Retrieved from https://www.diva-portal.org/smash/record.jsf?pid=diva2%3A306427&dswid=8645

Branberg, K., & Wiberg, M. (2011). Observed score linear equating with covariates. *Journal of Educational Measurement, 48(*4), 419-440. doi: 10.1111/j.1745-3984.2011.00153.x

Chen, H., & Holland, P. (2010). New equating methods and their relationships with Levine observed score linear equating under the kernel equating framework. *Psychometrika*, *75*(3), 542-557. doi: 10.1007/S11336-010-9171-7

Choi, S. I. (2009). *A comparison of kernel equating and traditional equipercentile equating methods and the parametric bootstrap methods for estimating standard errors in equipercentile equating* (Unpublished doctoral thesis). University of Illinois at Urbana-Champaign.

Cook, L. L., Eignor, D. R., & Schmitt, A. P. (1990). *Equating achievement tests using samples matched on ability* (College Board Report No. 90-2). New York: College Entrance Examination Board.

Gonzalez, J., Barrientos, A. F., & Quintana, F. A. (2015). Bayesian non-parametric estimation of test equating functions with covariates. *Computational Statistics and Data Analysis, 89,* 222-244. doi: 10.1016/j.csda.2015.03.012

Holland, P. W., Dorans, N. J., & Petersen, N. S. (2007). Equating test scores. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics* (Vol. 26, pp. 169-203). Oxford, UK: Elsevier.

Holland, P. W., & Thayer, D. T. (1989). *The kernel method of equating score distributions* (ETS RR-89-07). Princeton NJ: ETS.

Kolen, M. J. (1990). Does matching in equating work? A discussion. *Applied Measurement in Education, 3*(1), 97-104.

Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking: Methods and practices* (3nd. ed.). New York: Springer.

Liou, M., Cheng, P. E., & Johnson, E. G. (1997). Standard errors of the kernel equating methods under the common-item design. *Applied Psychological Measurement, 21*(4), 349-369.

Liou, M., Cheng, P. E., & Li, M. (2001). Estimating comparable scores using surrogate variables. *Applied Psychological Measurement, 25*(2), 197-207. doi: 10.1177/01466210122032000

Livingston, S. A., Dorans, N. J., & Wright, N. K. (1990). What combination of sampling and equating methods works best? *Applied Measurement in Education, 3*(1), 73-95.

Mao, X. (2006). *An investigation of the accuracy of the estimates of standard errors for the Kernel equating functions.* (Unpublished doctoral thesis). University of Iowa, Iowa City.

MoNE, Measurement Assessment and Examination Services General Directorate [Milli Eğitim Bakanlığı Ölçme Değerlendirme ve Sınav Hizmetleri Genel Müdürlüğü] (2016). Monitoring and evaluating of academic skills study, 8th students *report (ABIDE) [Akademik becerilerin izlenmesi ve değerlendirilmesi, 8. sınıflar raporu]*. Ankara: Republic of Turkey Ministry of National Education.

R Core Team. (2013). *R: A language and environment for statistical /computing*. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/**.**

Sansivieri, V., & Wiberg, M. (2016). IRT observed-score equating with the nonequivalent groups with covariates design. In L. A. van der Ark, M. Wiberg, S. A. Culpepper, J. A. Douglas, & W.C. Wang (Eds.), *Quantitative psychology- 81st annual meeting of the psychometric society* (pp. 275-285). Asheville, NC: Springer.

von Davier, A. A., Fournier-Zajac, S., & Holland, P. W. (2007). *An equipercentile version of the Levine linear observed-score equating function using the methods of kernel equating.* (Research Report No: RR-87-31). Princeton, NJ: Educational Testing Service.

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

von Davier, A. A., Holland, P. W., Livingston, S. A., Casabianca, J., Grant, M. C., & Martin, K. (2006). *An evaluation of the Kernel equating method. A special study with pseudotests constructed from real test data* (Research Report No: RR-06-02). Princeton, NJ: Educational Testing Service.

Wallin, G., & Wiberg, M. (2017). Non-equivalent groups with covariates design using propensity scores for kernel equating. In L. A. van der Ark, M. Wiberg, S. A. Culpepper, J. A. Douglas, & W. C. Wang (Eds.), *Quantitative psychology – 81st annual meeting of the psychometric society* (pp. 309-319). Asheville, NC: Springer.

Wiberg, M. (2015). Anote on equating test scores with covariates. In E. Frackle-Fornius (Ed.), *Festschrift in honor of Hans Nyquist on the occasion of his 65th birthday* (pp. 96-99). Stockholm, Sweden: Department of Statistics, Stockholm University.

Wiberg, M., & von Davier, A. A. (2017). Examining the impact of covariates on anchor tests to ascertain quality over time in a college admissions test, *International Journal of Testing, 17*(2), 105-126. doi: 10.1080/15305058.2016.1277357

Wiberg. M., & Branberg, K. (2015). Kernel equating under the non-equivalent groups with covariates design. *Applied Psychological Measurement, 39*(5), 349-361. doi: 10.1177/0146621614567939

Yurtçu, M. (2018). *Parametrik olmayan bayes yöntemiyle ortak değişkenlere göre yapilan test eşitlemelerinin karşilaştirilmasi* (Yayınlanmamış doktora tezi). Hacettepe Üniversitesi, Eğitim Bilimleri Enstitüsü, Ankara.

Eşdeğer Olmayan Gruplarda Ortak Değişkenlerin Kernel Eşitlemeye Etkisi

Giriş

Yaklaşık yüz yıldır, test eşitlemede yeni yöntemler ve desenler geliştirilmiştir. Bu yöntemlerden biri olan Kernel eşitleme yöntemi, ilk olarak Holland ve Thayer (1989) tarafından tanımlanmış ve daha sonra von Davier, Holland ve Thayer (2004) tarafından geliştirilmiştir. Kernel eşitleme, kesikli puan dağılımlarının sürekli dağılımlara dönüştürerek puan dağılımlarının eşitlendiği bir eşit yüzdelikli gözlenen puan eşitleme yöntemidir (von Davier ve diğerleri, 2006). Kernel eşitleme beş basamaktan oluşur: ön düzgünleştirme, puan dağılımlarının kestirilmesi, süreklileştirme, eşitleme ve eşitlemenin standart hatasının hesaplanmasıdır (von Davier ve diğerleri, 2004). Kernel eşitleme doğrusal ve eşit yüzdelikli eşitleme fonksiyonlarını içerir (von Davier ve diğerleri, 2006).

Test eşitlemede; tek grup deseni, eşdeğer grup deseni, dengelenmiş grup deseni ve denk olmayan gruplarda ortak madde deseni (NEAT) gibi birçok farklı grup deseni bulunmaktadır (Kolen ve Brennan, 2014; von Davier ve diğerleri, 2004). Alanyazında en sık kullanılan desenlerden biri NEAT desendir. Kernel eşitlemede NEAT deseninde; son tabakalama (PSE), Levine gözlenen puan doğrusal, zincirleme eşitleme (CE) yöntemleri kullanılmaktadır (von Davier ve diğerleri, 2004). NEAT deseninde iki farklı grup vardır ve bu gruplar, iki farklı test formu X ve Y ve ortak test olan A testini alır. PSE, grup I'deki X dağılımını ve grup II'deki Y dağılımını tahmin etmek için ortak test olan A'yı kullanır. A verilen X'in koşullu dağılımının ve A verilen Y'nin koşullu dağılımının popülasyonun değişmez olduğunu varsayar. CE'de ise ortak test zincirin bir parçası olarak kullanılır: ilk önce grup I üzerinden X testini A'ya, sonra da grup II üzerinden A ortak testini Y testine bağlar.

NEAT deseninde, gruplar arasındaki yetenek farkını ayarlamak için ortak maddeler kullanılmaktadır. Ancak bütün test programları veya standartlaştırılmış testlerde ortak maddeler test formlarında yer almayabilir. Ayrıca farklı test formlarını alan grupların eşdeğer olması da uygulamada çok zor olduğundan, eşdeğer grup desenine göre test formları eşitlenmeyebilir. Bu durumda, eğer denk olmayan gruplarda ortak madde yer almıyorsa anlamlı ortak değişkenler ortak maddeler yerine kullanılabilir (Wiberg & Branberg, 2015). Eğer ortak maddeler yerine ortak değişkenler kullanılıyorsa bu desen denk olmayan gruplarda ortak değişken deseni (NEC) adını alır.

Eşitlemede ortak değişkenler ile ilgili alan yazın incelendiğinde, sınırlı sayıda çalışma yapıldığı görülmüştür. Yurtdışında yapılan ilk çalışmalarda test eşitlemede farklı değişkenler kullanılarak ilerde ortak değişkenlerle yapılacak araştırmalara ışık tutulmuştur (Cook, Eignor & Schmitt, 1990; Holland, Dorans & Petersen, 2007; Kolen, 1990; Livingston, Dorans, & Wright, 1990). Son yıllarda yapılan çalışmalara baktığımızda, Branberg (2010) çalışmasında ortak değişkenleri kullanarak test formlarını eşitlemiştir ve ortak değişkenlerin ortak maddeler yerine kullanılabileceğini ifade etmiştir.

Bu bağlamda bu çalışmanın amacı, NEAT ve NEC desenlerinde Kernel eşitleme yöntemlerinden PSE ve CE doğrusal ve eşit yüzdelikli eşitleme yöntemleri ile elde edilen sonuçların karşılaştırılması ve cinsiyet ve sosyoekonomik düzey ortak değişkenlerin test eşitlemeye etkisini belirlemektir.

Yöntem

Bu çalışma kapsamında, 2016 ABİDE projesi kapsamında uygulanan matematik alt testlerinden elde edilen puanlar NEAT ve NEC desenlerine göre Kernel zincirleme eşit yüzdelikli, Kernel son tabakalama eşit yüzdelikli, Kernel zincirleme doğrusal ve Kernel son tabakalama doğrusal eşitleme yöntemleri kullanılarak eşitlenmiştir. Bu araştırmada, var olan yöntem ve tekniklerin gerçek veri üzerinden sınanması yapıldığından araştırma betimsel araştırmadır.

Bu çalışma kapsamında, Millî Eğitim Bakanlığı'ndan var olan öğrenci verilerinden rastgele olarak seçilen A formunu alan 3000 ve B formunu alan 3000 öğrenci olmak üzere toplamda 6000 öğrenciye ait veriler kullanılmıştır. A formunu alanların 1292'si (%43.07) kız ve 1708'i (%56.93) erkek; B formunu alanların 1518'i (%50.6) kız ve 1482'si (%49.4) erkek öğrencilerden oluşmaktadır.

NEC desende ortak değişkenleri belirleyebilmek amacıyla ABİDE projesi kapsamında hazırlanan raporda matematik başarısı ile öğrenci anketinde yer alan değişkenler arasındaki korelasyon değerleri incelenmiş ve bu değişkenlerden korelasyon değeri en yüksek olan sosyoekonomik indisi ($r = .37$, $p < .05$) ortak değişken olarak seçilmiştir (MEB, 2016). Sürekli bir değişken olan sosyoekonomik düzey indisini kategorilere ayırmak için iki aşamalı kümeleme analizi kullanılarak düşük, orta ve yüksek olmak üzere üç düzey oluşturulmuştur. Çalışmadaki bir diğer ortak değişken ise cinsiyet (sex) değişkeni olarak belirlenmiştir (Branberg & Wiberg, 2011; Gonzalez, Barrientos, & Quintana, 2015; Liou, Cheng, & Li, 2001, Yurtçu, 2018). Sosyoekonomik statü (ses); *düşük* = 1, *orta* = 2 ve *yüksek* = 3 ve cinsiyet değişkeni ise *kadın* = 1 ve *erkek* = 2 olarak kodlanmıştır. Testlerin eşitlenmesi için R programında (R Core Team, 2013) yer alan "kequate" paketi (Andersson, Branberg, & Wiberg, 2013) kullanılmıştır. Değerlendirme kriterleri olarak, eşitleme yöntemleri için eşitlemenin standart hatası (SEE) ve eşitleme hatası (RMSD) kullanılmıştır.

Sonuç ve Tartışma

Bu araştırmada, Kernel eşitleme yöntemleriyle NEAT ve NEC desenlerinde test formları eşitlenerek SEE ve RMSD katsayılarına göre eşitleme yöntemleri karşılaştırılmıştır. NEC desende ortak değişkenler olarak cinsiyet değişkeni ve sosyoekonomik indisi kullanılmıştır. Ortak değişkenler ayrı ayrı desene eklendikten sonra, iki ortak değişken birlikte eklenerek üç farklı NEC deseni oluşturulmuştur. NEAT ve NEC desenlerinde on ayrı eşitleme puanı elde edilerek, sonuçlar karşılaştırılmıştır.

Kernel doğrusal ve Kernel eşit yüzdelikli eşitleme yöntemleriyle elde edilen eşitlenmiş puanların, ham puan ranjında olduğu (0-9 aralığında) görülmüştür. Ham puan ile eşitlenmiş puanlar arasındaki en büyük farklılığın ise NEAT CE yöntemlerinde olduğu, diğer yöntemlerin kısmen birbirine daha yakın olduğu görülmüştür. Ayrıca NEC desende ham puan ile eşitlenmiş puanlar arasındaki farkın az olduğu ve NEC desende PSE yöntemi kullanıldığından NEAT PSE ile NEC desenlerinden elde edilen eşitlenmiş puanların benzer olduğu sonucu elde edilmiştir. Elde edilen bu bulgu Wiberg ve Branberg (2015) tarafından ulaşılan bulgularla tutarlıdır.

Eşitleme yöntemlerine ilişkin eşitlemenin standart hataları incelendiğinde; 0-4 puan aralığında yöntemlerin kısmen birbirine yakın veya benzer standart hatalara sahip olduğu; ancak orta puan ve uç puanlara doğru gidildikçe NEC-SES-L, NEC-SES-EQ ve NEC-SEX L eşitleme yöntemlerinin daha düşük standart hata verdiği görülmektedir. Ancak NEC desenin iki ortak değişken ile birlikte olduğu eşitleme deseninde en yüksek standart hata elde edilmiştir. Wiberg ve Branberg (2015) çalışmasında orta ölçek puan aralığında NEC desenin NEAT desenden daha büyük standart hataya sahip olduğu, ancak bütün puan ölçeği boyunca NEC, NEATCE, NEAT PSE ve NEATNEC yöntemlerinin benzer SEE değerlerine sahip olduğu sonucuna ulaşılmıştır. Bu çalışmada ise 0-4 puan ölçeğinde eşitleme yöntemlerinden elde edilen SEE değerleri kısmen benzer iken, 5-9 puan aralığında SEE değerlerinin yöntemlere göre farklılaştığı sonucuna ulaşılmıştır. Elde edilen bu bulgunun Wiberg ve Branberg (2015) tarafından ulaşılan bulgularla kısmen tutarlı olmadığı söylenebilir. Elde edilen bu bulgu Sansivieri ve Wiberg'in (2016) MTK'ya dayalı test eşitleme yöntemlerinde eşdeğer grup ve NEAT

desenlerinde ortak test ile birlikte ortak değişkenler kullanıldığında standart hatanın azaldığı bulgusuyla tutarlı olduğu söylenebilir.

Ayrıca NEC deseninden elde edilen yöntemleri karşılaştırırsak, sosyoekonomik düzey değişkeninin ortak değişken olarak kullanıldığı desende en düşük hata değerinin olduğu, en yüksek ise cinsiyet değişkeninin ortak değişken olarak kullanıldığı desende olduğu bulunmuştur. Bunun nedeni ise ortak değişkenler ile test arasındaki ilişkidir. Cinsiyet değişkeni ile test arasındaki korelasyon değeri istatistiksel olarak anlamsız iken, SES değişkeni ile düşük düzeyde anlamlı bir ilişki bulunmaktadır. Ortak test ile test arasındaki korelasyon değeri daha yüksek olmasına rağmen, ses değişkeni ortak maddeler gibi gruplar arasındaki farkları açıklayabilmiştir. Cinsiyet değişkeninin ortak değişken alındığı durumda hata yüksek iken, cinsiyet değişkenine ek olarak SES ortak değişkeninin eklenmesi de hatayı azalmıştır. Yurtçu'nun (2018) Bayes modelde ortak değişkenler ile eşitleme yaptığı çalışmada, iki ortak değişken kullanılmasının ortak maddelerden daha etkili olduğu sonucunu elde etmiştir.

Türkiye'de yapılan geniş ölçekli sınavlar göz önüne alındığında sadece ABİDE projesinde ortak maddeler yer almaktadır. Bu proje kapsamında ortak maddeler olmadığında ortak değişkenler de kullanılarak test formlarının eşitlenebileceği ve elde edilen sonuçların birbirine yakın olduğu elde edilmiştir. Bir yılda fazla geçerliği olan geniş ölçekli sınavların (KPSS, ALES gibi) test puanları eşitlenmek istendiğinde ise bu test formlarında ortak maddeler olmadığından eşdeğer grup/random grup desenine göre test formları eşitlenmelidir. Ancak bu durumda da grupların eşdeğer olma şartının sağlanması çok zor olduğundan, test ile ortak değişkenler arasındaki ilişki göz önüne alınarak, test formları eşitlenebilir. Benzer bir çalışma ABİDE projesinde yer alan Türkçe, Fen bilgisi gibi farklı alt testler için ve eşdeğer grup deseni ile NEC desen karşılaştırılarak yapılabilir. İleride yapılacak bir araştırmada NEC desende Parametrik olmayan Bayes modelleri ile Kernel eşitleme yöntemlerinden elde edilen sonuçlar karşılaştırılabilir.