International Journal of Assessment Tools in Education

2021, Vol. 8, No. 4, 729–*743*

https://doi.org/10.21449/ijate.976660

Published at https://ijate.net/ https://dergipark.org.tr/en/pub/ijate **Research Article**

Equality of admission tests using kernel equating under the non-equivalent groups with covariates design

Ozge Altintas[1,*], **Gabriel Wallin**[2]

[1]Ankara University, Faculty of Educational Sciences, Department of Educational Sciences, Educational Measurement and Evaluation, Ankara, Turkey
[2]Université Côte d'Azur, Inria, CNRS, Laboratoire J. A. Dieudonné, team Maasai, Sophia-Antipolis, France

ARTICLE HISTORY

Received: Mar. 03, 2021
Revised: July 03, 2021
Accepted: July 30, 2021

Keywords:
Kernel equating,
Non-equivalent groups design,
NEC design,
Background variables,
Admission tests.

Abstract: Educational assessment tests are designed to measure the same psychological constructs over extended periods. This feature is important considering that test results are often used for admittance to university programs. To ensure fair assessments, especially for those whose results weigh heavily in selection decisions, it is necessary to collect evidence demonstrating that the assessments are not biased and to confirm that the scores obtained from different test forms have statistical equality. Therefore, test equating has important functions as it prevents bias generated by differences in the difficulty levels of different test forms, allows the scores obtained from different test forms to be reported on the same scale, and ensures that the reported scores communicate the same meaning. In this study, these important functions were evaluated using real college admission test data from different test administrations. The kernel equating method under the non-equivalent groups with covariates design was applied to determine whether the scores that were obtained from different periods and measured the same psychological constructs were statistically equivalent. The non-equivalent groups with covariates design was specifically used because the test groups of the admission test are non-equivalent and there are no anchor items. Results from the analyses showed that the test forms had different score distributions and that the relationship was non-linear. Thus, the equating procedure was adjusted to eliminate these differences and thereby allowing the tests to be used interchangeably.

1. INTRODUCTION

Throughout much of human history, tests have been figured prominently as measurement tools in all areas of life. They are used for many purposes including monitoring the development process of individuals, determining the level of readiness for school, identifying the learning achievements of students, issuing diplomas or certificates, and deciding on proper treatment methods for psychological problems. This widespread use reveals the importance of tests in human life. Cronbach (1990) states that tests provide evidence for understanding individuals and gaining knowledge about human behavior. Anastasi (1988) defines psychological tests as

*CONTACT: Özge ALTINTAŞ ✉ oaltintas@ankara.edu.tr Ankara University, Faculty of Education, Department of Educational Sciences, Educational Measurement and Evaluation, Ankara, Turkey

an objective, standardized measure of a psychological variable such as intelligence, ability, aptitude, interest, attitude, and behavior.

One of the most common uses of tests is in schools. In pre-school, primary, and secondary school, basic life skills are taught, while in high school, the focus shifts to developing basic mental skills and orienting students to a future profession. Higher education programs, on the other hand, aim to equip individuals with the requisite set of skills and competencies associated with their profession of choice and at the same time, to enrich their intellectual, factual, and scientific knowledge. With the growing competitiveness in securing admittance to prestigious universities, it is common for students to take multiple admission tests to improve their chances of being accepted (Altıntaş & Kutlu, 2020).

Different forms of tests are used for entrance exams to universities and other educational institutions, for personnel selection, and for exams administered in different years or periods to ensure the security and integrity of the assessment process. In some cases, parallel versions of a test are used to allow the students more than one chance to be evaluated in certain periods. However, the use of different test forms on different dates raises concerns over whether the difficulty level of these forms differs (Kolen & Brennan, 2014). If no adjustment for difficulty differences is made, it is not possible to fairly compare test-takers who have been issued different test forms.

Similar questions asked in different formats, such as graphically, verbally, or symbolically, can be used multiple times in exams that measure the same construct, which is usually the case in exams administered for selection purposes. Although the use of parallel test forms that measure the same characteristics seems to be a reasonable way to ensure fairness (Kan, 2010) and exam security, the issue regarding the comparability of the scores obtained from these different tests is a source of concern.

The construction of parallel forms depends in equal measure on expert judgment and empirical data. The judgment comes into play in determining whether the items on these parallel forms measure the same function, a decision that sometimes is quite difficult to make (Levine, 1955, p.4). Proving that two tests, which are supposed to measure the same construct, are psychometrically equal (equivalent) to one another is essential in terms of preventing possible sources of bias.

Lord (1950) describes “comparability” in the sense that scores from two different tests each represent an equivalent amount of training or promises an equivalent degree of future success in a particular activity or other fields of knowledge. The comparability of scores obtained on different forms of a test depends on the accurate equating of these scores (Holland & Thayer, 1985, p.109). In selection processes, the comparability of the scores acts as an important indicator that the selection procedures are fair. As emphasized by Dorans and Holland (2000, p.281), the comparability of measurements made by different methods and researchers under different conditions is an essential component of the scientific method. Psychological and educational measurements are no exception to this rule.

Equating is a statistical process that is applied to confirm that scores on different test forms are comparable. Equating adjusts for differences in difficulty among forms that are built to be similar in difficulty and content (Kolen & Brennan, 2014). Equality/equivalence of test scores, or test equating was defined by Angoff (1971, 1982) as the conversion of the unit system of one form to the unit system of another form. Test equating is a numerical arrangement made to ensure that scores obtained from forms at different difficulty levels can be used interchangeably (Braun & Holland, 1982). Similarly, Felan (2002) stated that test equating is often used in situations where multiple forms of a test exist, where exams consisting of different forms are compared to each other, or when researchers want to overcome problems of practice effects.

A study by von Davier et al. (2004) argued that while there is no unified perspective on test equating, all equating approaches feature at least the five following "requirements": (1) equal construct, (2) equal reliability, (3) symmetry, (4) equity, and (5) population invariance. Here, equality is expressed in terms of the persons taking the exam, equal reliability and population invariance are related to the size of the population, symmetry is a mathematical property, and equal structure is related to the nature and use of the tests.

Since standardized tests are typically given at different times and with different test forms, the test that is administered by the test-takers must not unfairly affect the results capable of being attained (Andersson et al., 2013b). In effect, this means that the comparability of the scores obtained from a test and the interchangeability of the scores obtained in different years are important, insofar as they allow test-takers to compare their current scores with past and future scores. As is the case throughout the world, some tests are used in Turkey regularly (every year, twice a year, etc.) for the same purposes (selection, placement, etc.). The institutions responsible for developing and applying these tests accept that the different forms of the tests make equivalent measurements to realize the same purpose. The Ankara University Examination for Foreign Students (AYOS), which has been applied since 2011 for admission of international students to Turkish universities, especially Ankara University, is considered equivalent to each other. Research on the AYOS Basic Learning Skills Test (BLST) scores, such as measurement invariance and differential item functioning studies (Altıntaş & Kutlu, 2019, 2020), has revealed that AYOS has equivalence in terms of individuals in different groups (i.e., country and gender) who took the test the same year.

Although the psychological constructs measured by the test do not change, AYOS tests are developed for the same purpose and applied once every year. Hence, the groups taking the test are different (Kutlu & Bal, 2011). The gold standard is to use common items, also known as anchor items, to adjust for this kind of imbalance in ability between the test groups. However, AYOS does not include any common items. This study, therefore, follows the suggestion by Wiberg and Bränberg (2015) and uses background information about the test-takers. The idea behind this study is to investigate the equality of test forms that had no anchors, were assumed to measure the same construct and were applied to different groups in different years. This design is known as the non-equivalent groups with covariates (NEC) design (Wiberg & Bränberg, 2015). In the non-equivalent groups with anchor test (NEAT) design, the anchor test score is used as a proxy for the latent variable of ability, while in the NEC design, covariates instead act as proxies of ability. The latter can therefore be viewed as a generalization of the NEAT design since the anchor test score can be seen as a covariate. The NEC design allows for the inclusion of more than one covariate.

Accordingly, the purpose of this research is to identify the statistical equality of the different test forms of the AYOS BLST using the kernel equating method under the NEC design.

2. METHOD

2.1. Research Model

The basic research approach was used as the aim of this research was to equate AYOS tests that were administered in 2017 and 2018, and were assumed to measure the same psychological construct by testing existing techniques on real data. As part of this aim, we utilize covariates gathered at the time of the test administration within the NEC design to equate the test forms. Evaluation of the results is conducted by calculating the standard error of equating (SEE) and the standard error of equating difference (SEED).

In basic research, which is a type of scientific research concerned with clarifying the underlying processes and better understanding the phenomena, the hypothesis is usually expressed as a theory (Fraenkel & Wallen, 2009). Basic research can be exploratory, descriptive, or explanatory. Given that descriptive research is used to describe the characteristics of a population or phenomenon, which was part of the aim of this study, this specific type of basic research was applied.

2.2. The Study Group of the Research

The study group of this research consisted of 5,223 individuals who took the AYOS BLST – 2,460[*] took it in 2017, and 2,763[*] took it in 2018. In the 2017 group, there were slightly more men (52.2%), while there were slightly more women (52.19%) in the 2018 group. Regarding the age groups, about half of the individuals from the 2017 and 2018 groups were below 19 years of age (50.33% and 49.69%, respectively).

2.3. Data Set and the Test Equating Design

The research data included the test-takers' responses to the AYOS tests applied in 2017 and 2018. The AYOS is an assessment to determine international students' qualifications for admission to Ankara University and other universities (those accepting the AYOS score) in Turkey. The test is simultaneously implemented in different countries (exam centers) in a single session once a year. In brief, the AYOS dataset consists of the test-takers' scores (AYOS 2017 and AYOS 2018) and two covariates, gender (with values of 1 if man and 0 if woman), and age (with values of 1 if 18 years of age or younger and 0 if age 19 years of age or older). This means that there are 2 x 2 = 4 possible combinations of covariates and that the frequency vector has a length of 81 x 4 = 324. The data were first sorted by age followed by gender and the test scores on AYOS 2017.

The AYOS BLST is a non-verbal aptitude test with two sections and a total of 100 binary-scored multiple-choice items. The first section tests letter, number, and shape relations as a measure of cognitive skills, such as analytical thinking, reasoning, and abstract and spatial thinking (with 60 items). The second section measures numerical thinking skills that require the use of mathematics and geometry knowledge (with 40 items). The scores obtained from the test are valid for two years. The test is newly developed every year following the psychometric properties of the test applied in the previous year.

Table 1. *Descriptive statistics of AYOS tests.*

AYOS BLST	n	$\bar{X}$	S^2	S	KR-20[**]	Ave. Dif.	Skew.	Kurt.
2017	2.460	54.22	413.10	20.32	0.96	0.54	0.03	-0.86
2018	2.763	57.14	387.15	19.68	0.96	0.57	-0.00	-0.77

[**]The KR-20 formula was applied in cases where the items varied greatly in difficulty (Kuder & Richardson, 1937, p.160).

Table 1 shows the mean, standard deviation, variance, KR-20 reliability, average difficulty, skewness, and kurtosis coefficient values for the AYOS 2017 and 2018 tests. The first noteworthy finding was that the KR-20 reliabilities of both tests were equal and quite high (0.96). The KR-20 value is an overall measure of internal consistency (Cronbach, 1951, p.300) and provides information about the purity of random errors. Therefore, the fact that the values obtained from both tests were quite high is evidence that the tests involving the identification of number, shape, and letter relationships do measure the cognitive skills they aim to measure as a whole. Although the mean score on the 2018 test was higher than that on the 2017 test, both tests have values close to the average difficulty value of 0.50, which indicates that the

[*] Individuals from whom data on gender and age variables were collected were included in the study group.

students' scores generally hover around 50 points, which is the average score of the tests. This also shows that students can answer about 50% of the items on the test. Moreover, since the skewness coefficient of the score distribution in both tests is positive, the distribution is skewed to the right of what is considered normal. The kurtosis coefficients were negative for both tests, meaning that the score distributions, when compared to the normal distribution, were slightly flattened. The two score distributions are also presented in a histogram, which is given in Figure 1.

Figure 1. *AYOS BLST Score Distributions.*

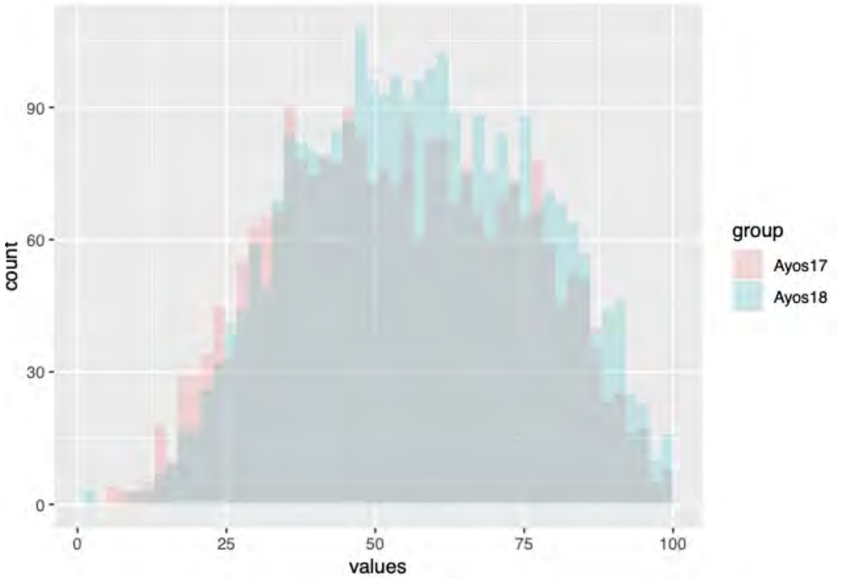


Figure 1 shows that the score distributions are slightly skewed, with relatively few test-takers having low scores and many having high scores. This is reflected in the SEE plot (Figure 3). Considering the test design, a NEAT design is typically preferred in the test score equating, but some tests do not have common items. If the groups are non-equivalent, an equivalent groups (EG) design cannot be applied (Sansivieri & Wiberg, 2017). When the test groups are non-equivalent and no anchor items are available, Bränberg and Wiberg (2011), Andersson et al. (2013a), and Wiberg and Bränberg (2015) recommend that background information about the test-takers be used to adjust for the ability difference, a design referred to as the non-equivalent groups with covariates (Wallin, 2019).

The idea of test linking using variables is not new, as demonstrated by Kolen (1990) and Livingston et al. (1990), who suggested that linking can be used in cases of groups matching on variables other than ability (as cited in Wiberg & Bränberg, 2015). The NEC design is an important alternative to the NEAT design when there is no anchor test available for equating (Wiberg & Bränberg, 2015). In the NEC design, background information on the individuals taking the tests is used instead of using an anchor test to facilitate the equating of two tests when the groups taking the test are not equivalent (Andersson et al., 2013a). As is the case in the NEAT design, two groups are independently sampled from different populations, P and Q, and each is administered either of the test forms, X and Y. In the absence of an anchor test form, the NEC design uses relevant covariates, denoted by C, that can account for differences in the groups of test-takers (González & Wiberg, 2017).

According to Wallin and Wiberg (2019), equating non-equivalent test groups requires adjusting for two sources of bias: differences in the difficulty of the forms and differences in the abilities of the test groups. A proper equating conversion should address both of these, but when the second is observed, some substitutes are required in place of ability. The most common substitute is an anchor test. However, since not all test programs can include an anchor, the background information of test-takers can be used. This is the scenario for the NEC design, where the fundamental assumption is that if the test groups are conditionally equivalent concerning the background information, they will differ only randomly from one another in terms of ability.

The NEC design was applied in this study due to the non-equivalent test groups of AYOS and the absence of anchor items. According to Bränberg and Wiberg (2011), one important consideration when using background information is the choice of variables, which should be correlated with the test scores. On the other hand, the variables should "explain" the differences between the groups in the non-equivalent groups design. Accordingly, the covariates used in this study were age and gender, denoted as A and G, respectively based on the availability of the AYOS data.

2.4. Data Analysis

The scores on the AYOS BLST 2017 and 2018 tests were equated using the kernel equating method under the NEC design in this study (Wiberg & Bränberg, 2015). The R package "kequate" was used for kernel equating analyses (Andersson et al., 2013a, 2013b; R Core Team, 2018).

The analysis of the data was carried out in two stages. In the first stage, pre-smoothing, continuization, equating, and evaluation of the equating function (computing the SEE) processes were carried out. In the second stage, a linear equating function was used to determine the degree of difference in the results of the 2017 and 2018 tests, and the SEED was calculated.

2.4.1. *The Kernel Equating Framework*

The kernel method of test equating includes the following five steps (von Davier et al., 2004; Andersson et al., 2013a, 2013b; Wiberg & Bränberg, 2015; González & Wiberg, 2017; González & von Davier, 2017; Wallin & Wiberg, 2017, 2019): Pre-smoothing, Estimation of the Score Probabilities, Continuization, Equating, and Evaluation of the Equating Function (Calculating the SEE and SEED).

The goal of test equating – if we let X and Y denote the test score from test form X and the test score from test form Y respectively – is to equate X to Y (or vice versa). The test group that was administered the test form X is a sample from population P, while the group that was administered the test form Y is a sample from population Q. To define the kernel equating estimator used in this study, let $r_j = P\left(X = x_j\right)$ and $s_k = P(Y = y_k)$ denote the score probabilities for scores $x_j, j = 1, \dots, J$ and $y_k, k = 1, \dots, K$. Furthermore, let μ_X and σ_X^2 denote the mean and variance of the X scores, respectively, let V denote a continuous random variable with mean 0 and variance σ_V^2, and let $a_X^2 = \sigma_X^2/(\sigma_X^2 + \sigma_V^2 h_X^2)$, where h_X is a smoothing parameter called the bandwidth. Using these defined quantities, a continuous version of the random variable X was introduced:

$$X(h_X) = a_X(X + h_X V) + (1 - a_X)\mu_X.$$

The random variable $X(h_X)$ is defined as such that its mean and variance are the same as for X, and its cumulative distribution function (CDF) is given by

$$F_{h_X}(x) = P(X(h_X) \leq x) = \sum_j r_j K\left(R_{jX}(x)\right),$$

where $K(\cdot)$ is the kernel function following from the distribution of V (which is commonly set to the Gaussian distribution) and $R_{jX} = \left(x - a_X x_j - (1 - a_X)\mu_X\right)/a_X h_X$. Corresponding quantities can be defined to introduce the continuized CDF G_{h_Y}. Replacing the terms in F_{h_X} and G_{h_Y} with estimated quantities, the kernel equating estimator used in this study was defined as

$$\hat{\varphi}_Y(x) = \hat{G}_{h_Y}^{-1}(\hat{F}_{h_X}(x)).$$

The SEE, which was used as part of the evaluation of $\hat{\varphi}_Y(x)$ in this study, equals

$$\mathrm{SEE}_Y(x) = \left\|\hat{\mathbf{J}}_{\varphi_Y}\hat{\mathbf{J}}_{\mathrm{DF}}\mathbf{C}\right\|,$$

where $\hat{\mathbf{J}}_{\varphi_Y}$ equals the Jacobian of the equating function, $\hat{\mathbf{J}}_{\mathrm{DF}}$ equals the Jacobian of the design function that is set according to the data collection design, and $\mathbf{C}$ is defined such that

$$\mathrm{Cov}\begin{pmatrix}\hat{\boldsymbol{R}}\\ \hat{\boldsymbol{S}}\end{pmatrix} = \mathbf{C}\mathbf{C}^{\top},$$

with $\hat{\boldsymbol{R}}$ and $\hat{\boldsymbol{S}}$ denoting vectors of pre-smoothed score distributions. Lastly, we defined the SEED as

$$\mathrm{SEED}_Y(x) = \left\|\hat{\mathbf{J}}_{\varphi_Y}\hat{\mathbf{J}}_{\mathrm{DF}}\mathbf{C} - \hat{\mathbf{J}}_{\varphi_L}\hat{\mathbf{J}}_{\mathrm{DF}}\mathbf{C}\right\|,$$

where φ_L equals the linear equating function

$$\varphi_L = \mu_Y + \left(\frac{\sigma_Y}{\sigma_X}\right)(x - \mu_X).$$

3. RESULTS

In the first stage of the equating process, pre-smoothing of the observed score distributions using the log-linear pre-smoothing was performed. A statistical model was fitted to the empirical distribution obtained from the sampled data in the pre-smoothing step. It is assumed that many of the irregularities observed in the empirical distributions are due to sampling error; thus, the pre-smoothing aims to reduce this error (Wiberg & Bränberg, 2015). Several log-linear models should be fitted and compared in the pre-smoothing step to decide which model fits the data the best (González & Wiberg, 2017).

3.1. Log-linear Pre-smoothing

González and Wiberg (2017) emphasize that several log-linear models should be fitted and compared in the pre-smoothing step regardless of the chosen data collection design. Here, the R function glm( ) was used to obtain a log-linear model in the pre-smoothing step to be used in the conjunction. The models were evaluated using the Bayesian Information Criterion (BIC; Schwarz, 1978), as it was shown to be an appropriate choice for bivariate smoothing (Moses & Holland, 2010). This led to log-linear models that preserved the first four moments of the *X/Y* score, the first two moments of the covariates, and the first cross-moment of the score variable and each covariate.

3.2. Estimation of the Score Probabilities

In the second step, the estimated score probabilities were generated by mapping the pre-smoothed score distributions into the score probability vectors for X and Y using a design function. This function, known as the design function, depends on the data collection design (see Wallin and Wiberg (2019) for the explicit expression of the design function for the NEC design).

3.3. Continuization

The Gaussian kernel was used in kernel equating to continuize the two estimated discrete cumulative distribution functions. The Gaussian kernel function is used to smooth the discrete score distributions, and the full penalty function is applied to select the smoothing parameter (von Davier et al., 2004). The estimated distributions $\hat{r}_j$ and $\hat{s}_k$, the bandwidths h_X and h_Y, and estimates of the means and variances of X and Y in population T were used in the application of the Gaussian kernel smoothing.

According to von Davier et al. (2004, pp.61-64), there is a variety of ways to select the bandwidth (h_X), which refers to controlling the degree of smoothness in the continuization, but the most common way was used in this study to minimize the penalty function.

The bandwidth for each continuized score distribution was selected by minimizing the sum of the squared distances between the observed score probabilities and the estimated density. To ensure smoothness in the estimated, continuized score distributions, the minimization operation included a term that penalized a density that had more than a few modes along with an added penalty term that penalized large fluctuations in the estimated density. Specifically, the bandwidth that minimized the following function was selected:

$$\sum_j \left(\hat{r}_j - F'_{h_X}(x_j)\right)^2 + \sum_j A_j,$$

where $F'_{h_X}(x_j)$ denotes the derivative of $F_{h_X}(x_j)$, $A_j = 1$ if $f'_{h_X}(x_j - v) > 0$ and $f'_{h_X}(x_j + v) < 0$, or $f'_{h_X}(x_j - v) < 0$ and $f'_{h_X}(x_j + v) > 0$, and $A_j = 0$ otherwise.

3.4. Equating

In the last step, the results were graphically examined by plotting the equated scores (Figure 2) and SEE (Figure 3). The table presenting the equated scores can also be found in the appendix (Annex 1).

Figure 2. *Equating results.*

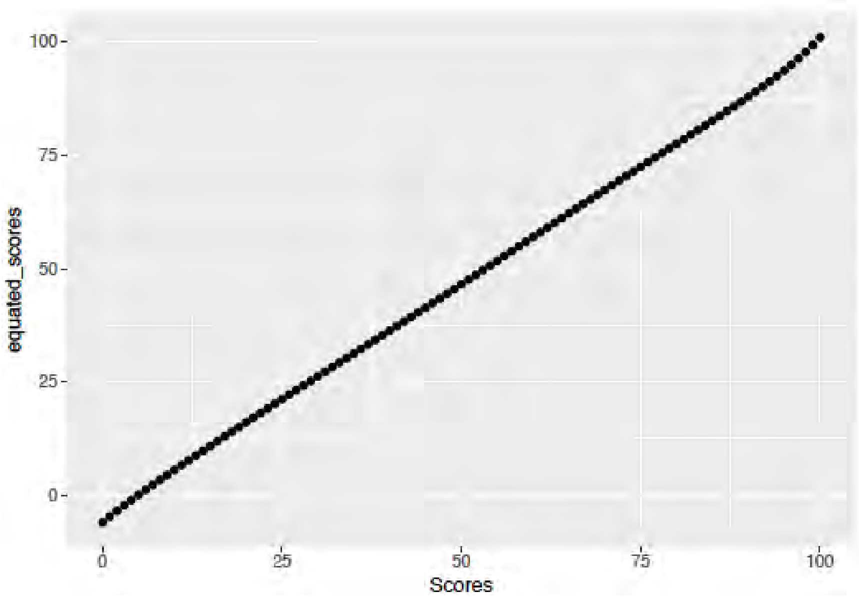
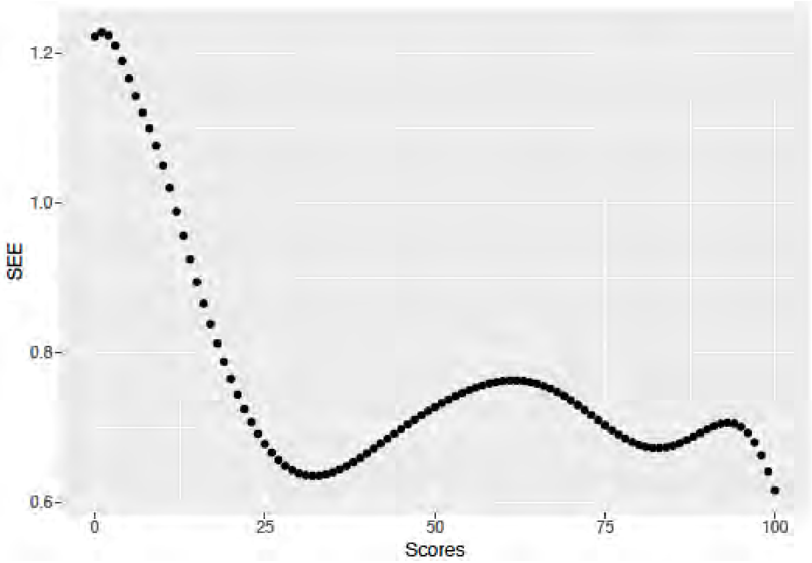


Figure 2 shows that there was a linear relationship between the raw scores and equated scores. Although the equating function was linear, there were non-linearities in the tails of the score distribution. It is also clear that the Y test form (AYOS 2018 test) was easier than the X test form (AYOS 2017 test), a difficulty difference that the equating function helped to adjust for.

3.5. Standard Error of Equating

Figure 3 shows the values of the SEE obtained for raw scores from the equating function.

Figure 3. *Standard error of equating.*



As illustrated in Figure 3, the SEE was larger at the lower end of the score scale. This is quite natural though, as there were very few test-takers with a total score below 10 (See Figure 1). Moreover, the SEE was relatively lower in the range of 25 - 100 scores.

Furthermore, an examination was performed to determine how the results obtained from a linear equating function would differ from the results already obtained. Therefore, the test forms were equated using a linear equation function, and then the difference between the previous equation function and the linear equating function was calculated. The results of this calculation indicated that the relationship between AYOS 2017 and 2018 tests was non-linear. In addition, the SEED was also calculated and added to Figure 4.

Figure 4. *Standard error of equating difference between current and linear equating function.*

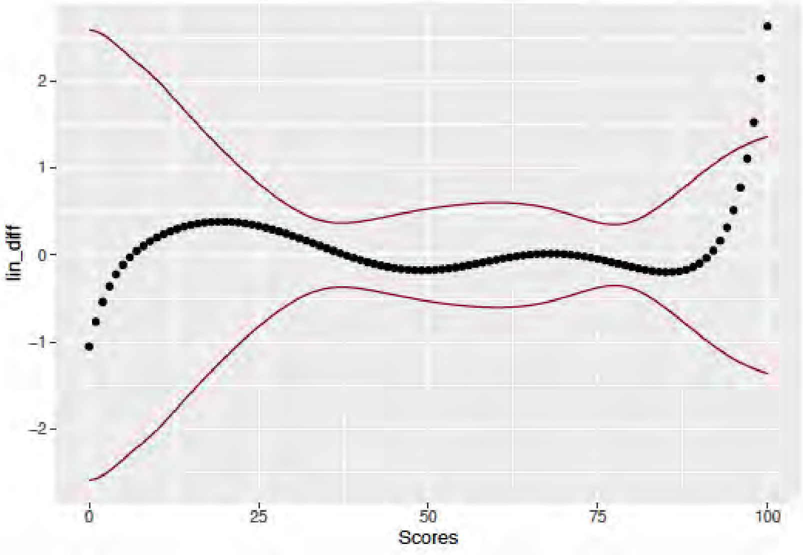


Figure 4 shows that the linear equating function deviated from the non-linear equating function. The line indicated by black dots shows the difference, while the red lines (±2SEED) represent twice the standard error of the difference between equating functions. The black line, however, only breaks through the SEED barrier once. This indicates that a non-linear equating method should be used instead of linear equating. Moreover, the SEED is relatively higher at the lower and higher ends of the score scale, which means that the linear equation methods give higher standard errors in extreme scores than in middle scores.

4. DISCUSSION and CONCLUSION

The aim of this research was to investigate the equality of test forms that had no anchors, were assumed to measure the same construct and were applied to different groups in different years. To fulfill this aim, variables correlated with AYOS BLST scores were used as a substitute for common items in non-equivalent groups with covariates design. This method introduced in studies by Bränberg and Wiberg (2011), Andersson et al. (2013a), and Wiberg and Bränberg (2015).

The NEC design was specifically used because the test groups of the admission test were non-equivalent and there were no anchor items. Results from the analyses showed that the test forms had different score distributions and that the relationship was non-linear. The equating procedure was thus adjusted to eliminate these differences and thereby allow the tests to be used interchangeably. Real data from a non-verbal aptitude college admissions test were used.

In a similar study, Akın-Arıkan (2020) used real data from the Monitoring and Evaluation of Academic Skills Project in Turkey to examine the NEAT and NEC designs comparatively. In this context, she equated the scores obtained from Mathematics subtests according to the kernel chained equipercentile, kernel post-stratification equipercentile, kernel chained linear, and kernel post-stratification linear methods. Furthermore, she sought to determine the affection status of the covariates (gender variable and socioeconomic index) used in the NEC design. From her research, it was determined that test forms can be equated using covariates when there are no anchor items. This is a noteworthy finding in terms of contributing valuable information for future studies to be carried out using the NEC design. When the findings obtained using the methods under the NEC design were specifically examined, the lowest error value was found in the design involving the socioeconomic index as a covariate, while the highest error value was found in the design involving the gender variable as a covariate. Akın-Arıkan reported that the reason for this was the relationship between the covariates and the test.

In this research, the point-biserial correlations were very low, and for the values between the covariates, none of the correlations were statistically significant (p>0.05). However weak correlation values between the covariates and the test scores do not mean that they are not good proxies of the latent ability. As we controlled for covariates that were confounders of the relationship between the test form assignment and the test score, we argue that as a rule the subject-matter knowledge of such covariates could be included to achieve a strong correlation. Similarly, Bränberg et al. (1990), in their research, found that there was a correlation between gender, education, and age in the test scores obtained from the Swedish Scholastic Aptitude Test (SweSAT).

In her research on real data, Yurtçu (2018) used gender, mathematics self-efficacy scores, and common item scores as covariates to obtain equated scores with the Bayesian nonparametric model. She concluded that covariates can be used in place of common items, and in some cases, perform even better, and that equated scores obtained with the said model can generate results closer to the target test.

The use of real-life data is important insofar as it reveals the psychometric properties of the tests used in real life. However, Wiberg and Bränberg (2015) warned that using real data is limiting because the true equating is not known. Therefore, simulation studies are recommended as they allow defining the true value of the equating (parameter) function, and they should be conducted using an NEC design within the kernel equating framework.

The evidence from the simulation study performed by Bränberg and Wiberg (2011) indicates that using covariates in the equating process can increase the accuracy of equating. In the present study, gender and age variables were used as covariates in the equating model. A review of the literature showed that background variables, such as gender, age, educational status, socioeconomic index, mathematics self-efficacy scores, etc., are being used as covariates (Bränberg & Wiberg, 2011; González et al., 2015; Wiberg & Bränberg, 2015; Wiberg & von Davier, 2017; Yurtçu, 2018; Akın-Arıkan, 2020). There are additional factors that may affect the student's success. These include student background variables, as used in PISA, such as the number of books at home, time allocated to studying, etc., or high school grades and performance test scores of the students. These variables can be taken as covariates in test equation studies using the NEC design. González et al. (2015) stated that an additional advantage of including covariates in the modeling of the equating function is the possibility of a customized transformation between any pair of subpopulations as long as they are characterized by covariates.

Since the test groups were non-equivalent and the AYOS tests do not contain any common items, this analysis used background information about the test-takers to equate the test forms. Although common items are the gold standard for adjusting for ability imbalance between test groups, previous studies have shown that equating under the NEC design produces smaller standard errors (Sansivieri & Wiberg, 2017) and lower MSE (Bränberg & Wiberg, 2011). While the model specification and the kernel equating framework are somewhat more complicated (Andersson et al., 2013a), they have advantages in terms of modeling flexibility.

In this research, since there were no anchor items, the two covariates, gender and age, were used to equate the different test forms of AYOS. Using covariates to obtain equated scores in the Bayesian nonparametric model, Yurtçu (2018) emphasized that the use of two covariates was more effective than the use of anchor items. Similarly, in another study, it was stated that a large number of covariates would cause a decrease in the number of individuals who fall into common categories and thereby result in errors in score estimation (Wallin & Wiberg, 2017).

Declaration of Conflicting Interests and Ethics

The authors declare no conflict of interest. This research study complies with the research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the author(s).

Authorship Contribution Statement

Ozge ALTINTAS: Investigation, Methodology, Resources, Visualization, Software, Formal Analysis, and Writing of the original draft. **Gabriel WALLIN**: Investigation, Methodology, Resources, Visualization, Software, Formal Analysis, and Writing of the original draft.

ORCID

Ozge Altintas https://orcid.org/0000-0001-5779-855X
Gabriel Wallin https://orcid.org/0000-0002-7930-6701

5. REFERENCES

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. http://dx.doi.org/10.21031/epod.706835

Altıntaş, Ö., & Kutlu, Ö. (2019). Investigating differential item functioning of Ankara University Examination for Foreign Students by Rasch model. *International Journal of Assessment Tools in Education*, *6*(4), 602-616. http://dx.doi.org/10.21449/ijate.554212

Altıntaş, Ö., & Kutlu, Ö. (2020). Investigating the measurement invariance of Ankara University Foreign Student Selection Test by latent class and Rasch model. *Education & Science, 45*(203), 287-308. http://dx.doi.org/10.15390/EB.2020.8685

Anastasi, A. (1988). *Psychological testing* (6th ed.). Macmillan.

Andersson B., Bränberg, K., & Wiberg, M. (2013a). kequate: The Kernel Method of Test Equating. *R package version 1.6.3.* https://CRAN.R-project.org/package=kequate

Andersson, B., Bränberg, K., & Wiberg, M. (2013b). Performing the Kernel Method of Test Equating with the Package kequate. *Journal of Statistical Software*, *55*(6), 1-25. https://www.jstatsoft.org/v55/i06/

Angoff, W. H. (1971). Scale, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 509-600). American Council of Education.

Angoff, W. H. (1982). Summary and derivation of equating methods used at ETS. In P. W. Holland & D. B. Rubin (Eds.), *Test equating* (pp. 55-69). Academic.

Bränberg, K., Henriksson, W., Nyquist, H., & Wedman, I. (1990). The influence of sex, education, and age on test scores on the Swedish Scholastic Aptitude Test. *Scandinavian Journal of Educational Research, 34*(3), 189-203. https://www.tandfonline.com/doi/abs/10.1080/0031383900340302

Bränberg, K., & Wiberg, M. (2011). Observed score linear equating with covariates. *Journal of Educational Measurement, 48*(4), 419-440. https://www.jstor.org/stable/41427533

Braun, H. I., & Holland, P. W. (1982). Observed-score test equating: A mathematical analysis of some ETS equating procedures. In P. W. Holland & D. B. Rubin (Eds.), *Test equating* (pp. 9-49). Academic.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*, 297-334. https://doi.org/10.1007/BF02310555

Cronbach, L. J. (1990). *Essentials of psychological testing* (5th ed.). Harper Collins.

Dorans, N. J., & Holland, P. W. (2000). Population invariance and the equitability of tests: Basic theory and the linear case. *Journal of Educational Measurement*, *37*(4), 281-306. https://doi.org/10.1111/j.1745-3984.2000.tb01088.x

Felan, G. D. (2002, February, 14-16). *Test equating: Mean, linear, equipercentile, and item response theory*. [Paper presentation]. The Annual Meeting of the Southwest Educational Research Associations, Austin, TX, United States. https://files.eric.ed.gov/fulltext/ED462436.pdf

Fraenkel, J. R., & Wallen, N. E. (2009). *How to design and evaluate research in education* (7th ed.). McGraw-Hill.

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

González, J., & von Davier, A. A. (2017). *An illustration of the Epanechnikov and adaptive continuization methods in kernel equating.* In L. A. van der Ark, M. Wiberg, S. A. Culpepper, J. A. Douglas, & W. C. Wang (Eds.), *Quantitative psychology* (pp. 253-262). IMPS 2016. Springer Proceedings in Mathematics & Statistics, vol 196. Springer. https://doi.org/10.1007/978-3-319-56294-0_23

González, J., & Wiberg, M. (2017). *Applying test equating methods using R*. Springer.

Holland, P. W., & Thayer, D. T. (1985). Section pre-equating in the presence of practice effects. *Journal of Educational Statistics, 10*(2), 109-120. https://www.jstor.org/stable/1164838

Kan, A. (2010). Test eşitleme: Aynı davranışları ölçen, farklı madde formlarına sahip testlerin istatistiksel eşitliğinin sınanması [Test equating: Testing the statistical equality of tests that measure the same behavior, and have different item forms]. *Journal of Measurement and Evaluation in Education and Psychology*, *1*(1), 16-21. https://dergipark.org.tr/en/download/article-file/65994

Kolen, M. J. (1990). Does matching in equating work? A discussion. *Applied Measurement in Education*, *3*(1), 97-104. https://doi.org/10.1207/s15324818ame0301_7

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

Kuder, G. F., & Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika, 2*(3), 151-160. https://doi.org/10.1007/BF02288391

Kutlu, Ö., & Bal, Ö. (2011). *Ankara Üniversitesi Yabancı Uyruklu Öğrenci Seçme ve Yerleştirme Sınavı (AYÖS) projesi kesin raporu* [Ankara University Student Selection and Placement Exam for Foreign Students (AYOS) project final report]. (Project No. 11Y5250001). Ankara University Scientific Research Project Office.

Levine, R. (1955). Equating the score scales of alternate forms administered to samples of different ability. *ETS Research Bulletin Series, 55*(2), i-118. Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1955.tb00266.x

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. https://doi.org/10.1207/s15324818ame0301_6

Lord, F. M. (1950). Notes on comparable scales for test scores. *ETS Research Bulletin Series, 50*(48), 1-20. Educational Testing Service. https://onlinelibrary.wiley.com/doi/pdf/10.1002/j.2333-8504.1950.tb00673.x

Moses, T., & Holland, P. W. (2010). A comparison of statistical selection strategies for univariate and bivariate log-linear models. *British Journal of Mathematical and Statistical Psychology*, *63*(3), 557-574. https://doi.org/10.1348/000711009X478580

R Core Team (2018). *R: A language and environment for statistical computing*. [Computer software]. R Foundation for Statistical Computing. *http://www.R-project.org/*

Sansivieri, V., & Wiberg, M. (2017). IRT observed-score equating with the nonequivalent groups with covariates design. In L. A. van der Ark, M. Wiberg, S. S. Culpepper, J. A. Douglas, & W. C. Wang (Eds.), *Quantitative psychology* (pp. 275-285). IMPS 2016. Springer Proceedings in Mathematics & Statistics, vol. 196. Springer. https://doi.org/10.1007/978-3-319-56294-0_25

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*(2), 461-464. https://projecteuclid.org/download/pdf_1/euclid.aos/1176344136

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

Wallin, G. (2019). *Extensions of the kernel method of test score equating*. [Doctoral dissertation, Umeå University]. Umeå University Libraries. http://umu.diva-portal.org/smash/get/diva2:1378833/FULLTEXT01.pdf

Wallin G., & Wiberg, M. (2017) Nonequivalent groups with covariates design using propensity scores for kernel equating. In L. A. van der Ark, M. Wiberg, S. S. Culpepper, J. A. Douglas, & W. C. Wang (Eds.), *Quantitative psychology* (pp. 309-319). IMPS 2016. Springer Proceedings in Mathematics & Statistics, vol. 196. Springer. https://doi.org/10.1007/978-3-319-56294-0_27

Wallin, G., & Wiberg, M. (2019). Kernel equating using propensity scores for nonequivalent groups. *Journal of Educational and Behavioral Statistics*, *44*(4), 390-414. https://doi.org/10.3102/1076998619838226

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

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. https://doi.org/10.1080/15305058.2016.1277357

Yurtçu, M. (2018). *Parametrik olmayan Bayes yöntemiyle ortak değişkenlere göre yapılan test eşitlemelerinin karşılaştırılması* [The comparison of test equating with covariates using Bayesian nonparametric method]. [Doctoral dissertation, Hacettepe University]. Hacettepe University Libraries. http://hdl.handle.net/11655/5295

6. APPENDIX

Annex 1. *Equating results.*

Scores	Equated Scores	Scores	Equated Scores
0	-6.04	51	47.55
1	-4.72	52	48.59
2	-3.46	53	49.63
3	-2.25	54	50.67
4	-1.08	55	51.72
5	0.07	56	52.77
6	1.19	57	53.82
7	2.29	58	54.87
8	3.39	59	55.91
9	4.47	60	56.96
10	5.55	61	58.01
11	6.62	62	59.06
12	7.69	63	60.11
13	8.75	64	61.15
14	9.81	65	62.19
15	10.86	66	63.23
16	11.91	67	64.27
17	12.95	68	65.30
18	13.99	69	66.33
19	15.03	70	67.36
20	16.07	71	68.39
21	17.10	72	69.41
22	18.12	73	70.44
23	19.15	74	71.46
24	20.17	75	72.47
25	21.18	76	73.49
26	22.20	77	74.50
27	23.21	78	75.52
28	24.22	79	76.53
29	25.23	80	77.55
30	26.24	81	78.57
31	27.24	82	79.58
32	28.25	83	80.60
33	29.25	84	81.63
34	30.26	85	82.66
35	31.26	86	83.69
36	32.27	87	84.73
37	33.27	88	85.79
38	34.28	89	86.85
39	35.28	90	87.93
40	36.29	91	89.02
41	37.30	92	90.14
42	38.32	93	91.29
43	39.33	94	92.47
44	40.35	95	93.70
45	41.37	96	95.00
46	42.39	97	96.37
47	43.42	98	97.82
48	44.44	99	99.36
49	45.48	100	100.99
50	46.51		