Uludağ Üniversitesi Eğitim Fakültesi Dergisi / Journal of Uludag University Faculty of Education
e-ISSN: 2667-6788 https://dergipark.org.tr/tr/pub/uefad
Yıl **2023,** Cilt **36**, Sayı **3**, **866-906** | DOI: https://doi.org/10.19171/uefad.1325587 **Araştırma Makalesi | Research Article**

# Comparison of Test Equating Methods Based on Classical Test Theory and Item Response Theory[1]

# Klasik Test Kuramı'na ve Madde Tepki Kuramı'na Dayalı Test Eşitleme Yöntemlerinin Karşılaştırılması

**Ceren MUTLUER**

Dr. Öğr. Üyesi ◆ Bolu Abant İzzet Baysal Üniversitesi, Eğitim Bilimleri Bölümü ◆ cmutluer@yandex.com ◆ ORCID: 0000-0002-3935-336X

**Mehtap ÇAKAN**

Prof. Dr. ◆ Gazi Üniversitesi, Eğitim Bilimleri Bölümü ◆ cakanmehtap@hotmail.com ◆ ORCID: 0000-0001-6602-6180

**Abstract**

This study aims to choose the equating method with the least equating error by using the equating methods in Classical Test Theory and Item Response Theory. In this study, booklet 1 and booklet 3 data were used for PISA (Programme for International Student Assessment) 2012 Mathematics test. Data from Turkey, Indonesia, Shanghai/China and Finland, countries participating in PISA 2012, were selected for this study. Non-equivalent groups design was used in the test equating process. Linear equating methods [Tucker ($w_1$=1, $w_1$=0.5), Levine observed score ($w_1$=1, $w_1$=0.5), Levine true score, Classical Congeneric and Braun-Holland), equipercentile equating methods (pre smoothing according to C6 polynomial degree, beta4, post smoothing according to S 0.05 cubic function, frequency estimation ($w_1$=1, $w_1$=0.5) ] were used in the study. In Classical Test Theory, the least error is obtained from the frequency estimation method with a synthetic universe weight of $w_1$ = 0.5. For the Item Response Theory, the calibration method was first decided, which is the Stocking-Lord method. After the scale transformation was achieved with the Stocking-Lord calibration method, the equating scores were calculated from the IRT's true and observed equating methods. The least error in IRT was obtained from the true score equating method. For error values, error coefficients were calculated according to Newton-Raphson's delta method and bootstrap methods. When the error coefficients (delta and bootstrap) of the equating methods in both theories were compared, it was found that the equating methods based on IRT had fewer errors than the equating methods in CTT, and the method with the least equating error was the IRT true score equating. The least equating error frequency estimation in CTT ($w_1$=0.5) and the most error Levine true score equating method.

***Keywords***: Test Equating, Classical Test Theory, Item Response Theory, Common Item Non-Equivalent Groups Design

**Özet**

Bu çalışmanın amacı Klasik Test Kuramı (KTK) ve Madde Tepki Kuramı (MTK) bünyesindeki eşitleme yöntemlerini kullanarak en az eşitleme hatasına sahip eşitleme yöntemine karar vermektir. Bu çalışmada PISA 2012 Matematik testi için kitapçık 1 ve kitapçık 3 verileri kullanılmıştır. Bu çalışma için PISA (Uluslararası Öğrenci Değerlendirme Programı) 2012 uygulamasına katılan Türkiye Endonezya, Şangay/ Çin ve Finlandiya ülkelerin verileri seçilmiştir. Test eşitleme sürecinde eşdeğer olmayan gruplar deseni kullanılmıştır. Araştırmada ele alınan KTK'da doğrusal eşitleme yöntemleri [Tucker ($w_1$=1, $w_1$=0.5), Levine gözlenen puan ($w_1$=1, $w_1$=0.5), Levine gerçek puan, klasik konjenerik ve Braun-Holland], eşit yüzdelikli eşitleme yöntemleri [C6 polinomial derecesine göre ön düzgünleştirme, beta4 , S 0.05 kübik fonksiyona göre son düzgünleştirme, frekans kestirim ($w_1$=1, $w_1$=0.5)] kullanılmıştır. Klasik Test Kuramında en az hata $w_1$=0.5 sentetik evren

[1] This article was based on the doctoral thesis that prepared by Ceren Mutluer with the supervision of Prof. Dr. Mehtap Çakan.

Kabul Tarihi / Received: 10.07.2023 | Onay Tarihi / Accepted: 22.09.2023

ağırlığıyla Frekans kestirim yönteminden elde edilmiştir. MTK için öncelikle kalibrasyon yöntemine karar verilmiş ve bu yöntem Stocking-Lord yöntemidir. Stocking-Lord kalibrasyon yöntemi ile ölçek dönüşümü sağlandıktan sonra MTK'daki gerçek ve gözlenen eşitleme yöntemlerinden eşitlenmiş puanlar hesaplanmıştır. MTK'daki en az hata gerçek puan eşitleme yönteminden elde edilmiştir. Hata değerleri için Newton-Raphson'un delta yöntemi ve boostrap yöntemlerine göre hata katsayıları hesaplanmıştır. Her iki kuramdaki eşitleme yöntemlerinin hata katsayıları (delta ve boostrap) karşılaştırıldığında MTK'ya dayalı eşitleme yöntemlerinin KTK'daki eşitleme yöntemlerinden daha az hataya sahip olduğu ve en az eşitleme hatasına sahip olan yöntemin MTK gerçek puan eşitleme olduğu bulunmuştur. KTK'da en az eşitleme hatası frekans kestirim ($w_1$=0.5) ve en fazla hata Levine gerçek puan eşitleme yöntemidir.

***Anahtar Kelimeler***: Test Eşitleme, Klasik Test Kuramı, Madde Tepki Kuramı, Ortak Maddeli Eşdeğer Olmayan Grup Deseni

## 1. Introduction

In education, tests are applied to students for many purposes. For example, the results of the exams are used in situations such as making decisions about the performance of individuals and the general education level of a country, transition to a higher institution, and placement in a higher institution. In these exams, the participants are expected to measure their abilities with less error, more accurately, and objectively. Therefore, for the stated objective measurement, the measurement process should first be started by converting the application conditions of the tests to standard conditions.

Providing an equating psychological or educational assessment tool is one of the main reasons for standardized testing (Cook & Eignor, 1991). High-stake tests or large-scale exams are tried to be tested for standardization, and the reliability of the exams is tried to be increased. However, Crocker & Algina (1986) stated that errors are always involved in the application measurement results. For this reason, there will undoubtedly be an error in the measurements obtained when the dynamics, such as the application process, the test itself or the raters are considered. This situation prevents the formation of standard conditions. For this reason, asking the same questions over and over in high-stake tests or large-scale exams will cause the person to remember a question they have solved before instead of measuring the expected performance. To prevent this situation, developing different forms of the test that will measure the same feature would be appropriate.

The results obtained in exams such as KPSS (Civil Servant Selection Examination ), YDS (Foreign Language Proficiency Exam), ALES (Academic Personnel and Postgraduate Education Entrance Exam) and YÖKDİL (Foreign Language Exam of Higher Education Institutions) are valid for several years. The results of these exams have an important contribution in situations such as placement in various institutions and promotion. In these processes, different forms are used that are claimed to measure the same features to prevent marking with remembering and to ensure test confidentiality at different sessions. Although the reliability of the test has been ensured by applying different forms, some doubts will undoubtedly arise about the equality of the results obtained. Even though the forms were prepared with the parallel test logic for different sessions, the same score obtained from the forms may not indicate the same skill level for different forms due to the different abilities of the group taking the test . Although the test forms prepared for the same purpose based on the same content were prepared with the claim of parallel, indices such as test reliability, item difficulty, and standard deviation of the test take different values in each application. The scores will vary according to the difficulty and ease of the test among the people who take the different forms. Even if the same individual gets the same score from different forms, making the same interpretation of his performance would not be correct. That is why comparing the scores obtained from the forms is

necessary. The applications performed with different tests at different times, the scores of the same person and different people cannot be directly compared. The test results used have been in use for several years. Since the results of these exams have been used for several years, there is a need to equate them in order to compare the scores.

As Dorans and Holland (2000) stated, comparing measurements from tests performed with different methods in different situations has been the essential prerequisite of all sciences. In this context, there is a need for a statistical procedure that enables the conversion of the scores obtained from test forms prepared for the same content and the same performance.

For the results obtained from different forms to be used interchangeably, these scores should be formed on a common scale or more specifically. A mutual relationship should be established between the scores of the two tests. This relationship can be realized with test equating (Zhu, 1998).

### 1.1 Test Equating

When the available literature is reviewed, many definitions for the concept of 'test equating' can be found. According to Angoff (1987), test equating is to convert the unit system of one form to the unit system of another form. With another definition, establishing the relationships between the scores in two or more tests with a statistical method or simply placing these test results on a common scale is called “test equating” (Hambleton & Swaminathan, 2013). On the other hand, Kolen & Brennan (2014) defined the test equating as a statistical process that allows the scores obtained from these forms to be used interchangeably by arranging the differences between test forms with similar content and similar difficulty levels.

Before starting the procedure, certain conditions must be met. When the accessible literature is scanned, it has been determined that five conditions must be met (Angoff, 1987; Dorans & Holland, 2000; Kolen & Brennan, 1995; Kolen & Brennan., 2014; Petersen et al., 1989). These five conditions are symmetry, measuring the same features, equal reliability, independence from the group, and equality features.

### 1.2 Test Equating Designs

For test equating, it is necessary to start the data collection process. The data collection process in equating is called the 'test equating design' (Kolen & Brennan, 2014). The selected design is essential for the successful conclusion of the test equating process. For this reason, the design to be chosen for equating is expected to be economical and unbiased (Thorndike, 1982). Therefore, these designs are ‘random, single group, single group design with counterbalancing, common-item nonequivalent group, covariate-design with nonequivalent designs.’

In the current study, a common item non-equivalent group design was chosen. In this design, the same items are included in the forms given to the participants. These are called anchor(common) items. These common contents are the same contents that apply in both forms. The order of these items in the test is also the same. The number of common items should be at least 20% of the total items in the test. Common items should represent the entire item group in the test (Kolen & Brennan, 2014; Petersen et al., 1989). The common items created are included in each test. Therefore, the differences between the two forms can be adjusted depending on the common item statistics, because the two groups that receive the forms do not have to be equivalent.

### 1.3. Test Equating Methods

After the data collection procedure is selected, equating methods should be determined. Test equating methods fall into two general categories based on test theories:

- Equating based on Classical Test Theory (CTT) (Kolen, 1988)
- Equating based on Item Response Theory (IRT) (Cook & Eignor, 1991)

In this Classical Test Theory, there are three equating methods. These are mean, linear(LE) and equipercentile equating(EE) methods. This study focused on LE and EE methods for CTT.

LE is based on the difference of scores from their mean divided by their standard deviations. The difference from the mean equating is the standard deviation value in the following equation (Kolen & Brennan, 2014).

$$\frac{(X-\mu(X))}{\sigma(X)} = \frac{(Y-\mu(Y))}{\sigma(Y)} \quad (1)$$

In this equation, X' defines as the score from X form. "μ(X)" is defined as the mean of X form. "Y" defines as the score from the Y form, "μ(Y)" defined as the mean of Y form and "$\sigma(X)$" define as the standard deviation of X form and the last symbol of the equation, defines as the standard deviation of Y form. The following equation is used to find the percentile rank in the EE.

For common item non-equivalent groups- LE equating methods tests containing common items were applied to two groups of participants from different samples. This pattern is generally used when only one test form is administered at the given test time. Thus, it was emphasized that common items should be prepared in a test, in the same order, with the same content and statistical values (Kolen, 1988). There are two special cases for common-item non-equivalent group patterns. If the first of these is calculated by reflecting the common item on the test scores for all forms, it is indicated as an internal (internal anchor) item. Secondly, if these common items are not considered in the test score, they are called external (external anchor) items. For this research, internal anchor items were chosen.

In general, common items are used to correct for sample differences. Although this design includes two populations, an equating function is typically defined for a single population. Therefore, population 1 and population 2 must be combined to define a relationship as if derived from a single population. The "synthetic population" (Braun & Holland, 1982). Considering that the weight of population 1 is $w_1$, the weight of population 2 is $w_2$, $w_1$ and $w_2$ should be ≥0 following the rule of $w_1+w_2=1$ (Kolen & Brennan, 2014).

In the Tucker-LE method, the groups are tried to be equated by considering the different synthetic population weights and the synthetic population weights presented above. According to Kolen & Brennan (2014) and Gulliksen (1950), when V is accepted as a common test, the regression of X on V assumes the same linear function for population1 and population2. Considering this information, the mean and variance values are tried to be estimated using the help of internal anchor V scores and synthetic population weights.

The other equating method is Levine observed score (LevineOS) equating method. This method does not address the concept of a synthetic population. Instead, this is an observed score method that relates the observed scores on X to the observed score scale on Y. The Levine method states that X, Y, and V measure the same things if the correlation between $T_X$ and $T_V$, $T_Y$ and $T_V$, is perfect in population1 and population2 X, Y, and V (Kolen & Brennan, 2014).

The other Levine equating method is Levine True score (LevineTS) equating method. Developed by Levine (1955), it contains the same assumptions as the Levine observed score equating method. The application difference between the observed score and actual score methods is using

actual scores in the equation that converts the observed scores on X to the observed scores on the Y scale. The assumption that the mean score observed in the CTT is equal to the actual mean score is used in this method.

Congeneric test theory is a sub-dimension of CTT (Lucke, 2005). By using this theory. we can equate the scores. In this theory, the observed score equality in this theory is an improved version of the linear model in CTT that includes item characteristics. A classical congeneric model is assumed for X and V and a single population. It extends the results presented here to Y, V, and Population 1 and 2 (Kolen & Brennan, 2014).

Another equating method is Braun-Holland equating. In this method, equating is done using the mean and standard deviations that emerge using the assumptions of the frequency estimation method. With the assumption of frequency estimation. the X form's mean and standard deviation scores can be estimated by the equation below. In this way, the mean and variance values of the X form in the synthetic population are made similar to the Y form. The Braun-Holland method is closely related to the Tucker method.

In the EE methods equating function, if the distribution of the form X scores converted to the form Y scale is equal to that of the form Y scores in the population, this is an EE function. The EE function was developed by defining the scores in form X with the same percentile ranks as those in form Y. In other words, the main thing in EE is to transform the score distributions obtained from different populations into their equivalents in the same percentile order. According to Angoff (1987), the scores obtained in measuring the same feature from the X and Y forms with an equal degree of reliability with equipercentile ranks are accepted as equivalent.

In EE, the following general steps are followed in the graphical and analytical process. First, for a certain X score in Form X, there is the percentage of individuals who achieve this score or below; the percentage found is equal to the score in form Y, which has the same percentage; the Y form score found is the equivalent of the form X score.

$$P(X) = 100\left[F(X-1) + \frac{f(X)}{2}\right] \qquad (2)$$

In this equation “P(X)” defines as the percentile rank function for X, “F(X)” defines as the cumulative distribution for X, “f(X)” defines as the discrete density for X. Although these fluctuations in the score distribution are tried to be avoided by using a very large sample, especially when the sample is small, these curves are usually smoothed by analytical smoothing methods (Kolen, 1988; Livingston, 1993). With smoothing, it is tried to find the relationships in the population and to convert the discrete distributions in the sample into a continuous function.

Smoothing methods are designed to produce smoothing functions with less random errors obtained from unsmoothed EE (Hanson et al., 1994). With the use of smoothing methods, the total error and random error are reduced. However, it can increase systematic error (Felan, 2002). There are two smoothing methods: pre-smoothing and post-smoothing. In pre-smoothing, firstly, the score distributions are smoothed. Accurate estimation of score distributions is an important point to be considered in the smoothing process. The EE process is done later, Log-linear smoothing based on the polynomial function is used for pre-smoothing and the Beta4 method is used to reach the true score (Kolen & Brennan, 2014).

In the post-smoothing method, smoothing is done after obtaining equipercentile equivalents. As Tan (2015) stated, the transformed scores are smoothed in the final smoothing method, not the distribution of test scores. In this method, cubic intermediate values are used instead of the polynomial values in the log-linear method. Therefore, the cubic spline method is used as the final straightening method.

With using common item non-equivalent groups–EE methods, paying attention to the distribution of the total scores and the scores obtained from the common items is important. This method requires consideration of the synthetic phase. While trying to equate the total score and common item scores obtained in the common item non-equivalent group design with EE, the equating functions of the frequency estimation method according to different synthetic population weights were used.

For common item non-equivalent groups–EE methods, the frequency estimation method is one of the test equating methods. The frequency estimation EE method described by Angoff (1987). Braun and Holland (1982) provide a mean for the cumulative score distribution estimation on Form X and Form Y for a synthetic population from data collected using the common item non-equivalent group design. The percentiles are obtained from the cumulative distributions and the forms are equated with the EE method.

The equating methods that were shown above are used for the CTT procedure. In IRT, there are different equating methods than CTT. IRT was developed against the weak assumptions of the CTT (Embretson & Reise, 2013; Lord, 1980). To examine research within the body of the IRT, three important assumptions must be met. There are unidimensionality, local independence, and monotonic increase in the item characteristic curve (ICC) (Embretson & Reise, 2013).

The parameter estimations resulting from the IRT parameter estimation operations are usually on different IRT scales (Hambleton & Swaminathan, 2013). For example, parameters for IRT models are estimated for the X form on which the participant sample in sample 1 is based and for the Y form on which the participant sample in sample 2 is based, and these two samples are not equal. Computer programs often define the θ scale as analyzed data with a standard deviation of 1 and a mean of 0. In this case, talent estimations are made for each group with a mean of 0 and a standard deviation of 1. Therefore, conversion to IRT scales is required. In test equating, scale conversion (calibration) according to parameter estimations is divided into two "simultaneous calibration" and "asynchronous calibration". In simultaneous calibration. the parameters of the forms are estimated together. In contrast, the form parameters are estimated separately and located on the same scale with the linear equation in asynchronous calibration. A linear equation is used to convert the a and b parameters of the scores from each form to the same scale (Stocking & Lord, 1982).

Scale conversion methods based on IRT are divided into two main headings. These are moment methods (mean-mean equating, mean-standard deviation) and characteristic curve methods (Haebara, Stocking-Lord). Before starting the test equating process within the body of IRT. the scores obtained from the forms are converted to the same scale with moment methods or characteristic curve conversion methods. The following process is testing the IRT test equating methods. There are two methods. The methods are ‘True Score Equating’ and ‘Observed Score Equating’ (Kolen & Brennan, 2014).

After the item parameters are converted to the same scale. IRT true score equating can correlate with the correct answer scores on the X and Y forms. In this process, care is taken to ensure that the score on a form related to a particular θ is equivalent to the score on another form related to this θ. In the true score equating process, three stages must be followed. In the first step, true score $\tau_x$ in Form X should be determined. Then the $\theta_i$ value corresponding to the determined true score should be found. In the final step, find the true score in form Y corresponding to this $\theta_i$ (Petersen, Cook & Stocking, 1983).

IRT observed score equating is the distribution estimation of the correct number of items observed on each form. The composite binomial distribution for the X form generates the correctly answered item score distribution observed for participants at a given ability level.

### 1.4. Importance and Aim of The Research

Comparing scores from different test forms and using them interchangeably justifies performing test equating studies. In applications where different forms are used simultaneously in large-scale exams such as PISA, test equating studies are emphasized to determine the success situations and correctly make the success order. The PISA 2012 data selected as the research data are equated by making scale point conversions. The scores obtained for 13 booklets are tried to be equating by common items. In the equating process for PISA 2012 data, ability levels are estimated at the IRT and the Rasch model is used for the same year data (OECD, 2014). A linking scale was prepared to compare scores with PISA 2012 data, PISA 2003, 2006 and 2009 data. It is critical to decide on the equating methods in applications where the country success status of PISA data is compared and to determine the equating method with the least errors. The most appropriate equating method should be determined according to the data structure used in the process and the equating process should be completed with the least error. When the literature is examined, there is no common opinion about the most appropriate equating method. The equating method with the least error varies in conditions such as the pattern used, the ratio of common items, whether the data are simulation or true data, sample size and distribution of this sample. It has been seen in the literature that studies using different sample sizes and true data generally include equating studies (Özdemir, 2017; Sezer Başaran, 2023; Skaggs, 2005; Tan, 2015; Von Davier & Kong, 2005; Wang, et al. 2008) within the body of CTT. Equating studies using simulation data or working with larger samples (Brossman & Lee, 2013; Gündüz, 2015; Kilmen, 2010; Yurtçu & Güzeller, 2018) were carried out within the scope of IRT, since it is more difficult to meet the assumptions and equating conditions in theory. Equating studies based on CTT and IRT are quite limited. This research is thought to contribute to the field as a study in which the Rasch model specified in the PISA 2012 report is not used. However, the analysis under 3PLM, the equating methods within the scope of CTT and IRT are discussed in detail, the most appropriate one with equating scores is determined and real data is used.

Many studies have been designed regarding the intended use of these methods and the structure discussed. In this study, it was aimed to equate the mathematics scores in booklet 1 and booklet 3 of PISA 2012 with the equating methods based on CTT and IRT by using the pattern of unequivocal groups with common items, and to determine the most appropriate equating method used.

Answers to the following questions were sought considering the problem statement created for the research.

1-Which equating method contains the least equating error for the equating scores obtained from different booklets of PISA 2012 using Tucker equating method, LevinTS, LevineOS, congeneric and Braun-Holland LE methods in CTT?

2- Which equating method contains the least equating errors for the equating scores obtained from the different booklets of PISA 2012 using the frequency estimation EE methods in the CTT?

3-Which equating method contains the least equating error for the equating scores obtained from the different booklets of PISA 2012 using the actual and observed score equating methods, which are equating methods based on IRT?

4- When the equating method with the least equating error in the CTT and the equating method with the least equating error from the IRT are compared, which theory's equating method contains the least equating errors?

## 2. Method

### 2.1. Research Method

In this study, it was aimed to select the most appropriate equating methods based on CTT and IRT-based equating methods using the common-item non-equivalent groups design and PISA 2012's mathematics test scores. Since it is aimed to find the one that gives the least error among the different equating methods used in this research, this study is descriptive research. Descriptive research is suitable for research that aims to reveal the existing situation as it is (Karasar, 2005).

### 2.2. Sample

The population of the study consists of 15-year-old students who participated in PISA 2012. When the mathematical literacy scores of the 65 countries participating in the PISA application were examined for the study group, four countries in total were selected as the most successful, the most unsuccessful, below the mean and above the mean. Since they represent the countries participating in PISA 2012, it was deemed appropriate to select these countries. The selection of the countries in the working group was carried out as follows.

*Shanghai/China was chosen as the country with the best performance in PISA 2012 science, mathematics and reading skills.

*Although Peru was the most unsuccessful country for PISA 2012, the country above Peru in the order of success. Indonesia, was determined because the booklets in Peru and other countries could not match.

*Turkey, which participated in the PISA 2012 application, was deemed appropriate because it was below the mean.

* Finland, whose overall level of success in PISA 2012 has decreased compared to the previous PISA application, is above the mean. Compared to Turkey, which is below the mean. Finland, which is above the mean, was chosen because it represents successful countries.

Purposeful sampling was used as it allows in-depth research in the equating process by selecting information-rich situations depending on the purpose of the research (Büyüköztürk et al., 2008).

In this study, within the scope of purposive sampling, booklet 1 received 1921 people, and booklet 3 received 1900 people. The distribution of the people who took the booklets by country is given in Appendix 1. In line with Appendix 1, the scores obtained from a total of 3821 people were used in the equating.

### 2.3. Data Collection Tool

Program for International Student Assessment-PISA is one of the most comprehensive educational studies in the world organized by the Organization for Economic Co-Operation and Development (OECD) (MEB, 2013). With this research, which has been carried out every three years since 2000, it is evaluated to what extent 15-year-old students in OECD member countries and other participating countries (approximately 90% of the world economy) have the basic knowledge and skills necessary to take their place in modern society (MEB, 2013). In the PISA 2012 application, the weighted

area is mathematical literacy. Therefore, this study's analyses were carried out completely according to the mathematical results. In the PISA application, there are 13 booklets and the answers such as true-complete and true-invalid-blank are converted into scores by the mean 500 and standard deviation 100 rule. For some of the common items in these 13 booklets, comparing the scores and using them interchangeably was possible. In this context, in this study, the process of equating the scores obtained from the two booklets (booklet 1 and booklet 3) selected for PISA 2012 is explained. When the common item syntax and the overlap of other items are examined, the booklets to be equating were determined as booklet 1 and booklet 3. The items in the booklets and their coding are presented in Appendix 2. The items in Appendix 2 were created by sorting according to countries and booklets. When Appendix 2 is examined, there are 25 items in each booklet, 13 of these items are common and 12 are non-common items. In the cognitive test, correct answers were coded as “1”, partially correct, incorrect, and “0” for other answers. Since coding based on the dual scoring model was preferred in this way, partially correct answers were also evaluated as incorrect answers.

The equating process is based on equating the new form to the old form. In this research, booklet 1 is the new form (form X); booklet 3 is designated as the old form (form Y). In this case, it was desired to equate the scores obtained from Booklet 1 to the scores obtained from Booklet 3.

In the research, PISA-2012 booklet 1 and booklet 3 scores of China, Finland, Turkey, and Indonesia were equating. In the booklets used in equating, 12 items were prepared differently for both groups. The 13 common items in the booklets were applied to both groups in the same order. For this reason, the scores of booklet 1 and booklet 3 can be equating by using common items (13 items). The research data was download from http://www.oecd.org/pisa/data/ website. Since the data were collected in this way during the research, ethical permission is not required.

### 2.4. Data Analysis

In the study, first, test equating assumptions were tested. The analysis of the assumptions that need to be tested in the equating process before proceeding to the analysis of the data (Appendix 3). While examining the equating methods within CTT, besides analytical and graphical solutions, error values were examined with the 'Equate-Error' program. While the LE methods in CTT, one of the equating methods, were compared among themselves, the equating methods in EE were also compared within themselves. Since the equating scores in the equating methods were compared with the raw scores. The equating error in all methods was calculated with the WMSE (Weigted Mean Square Error) coefficient. In IRT, on the other hand, firstly, the most appropriate calibration method was selected. And the scores were converted into a single scale and then the actual and observed score methods were compared. To find the appropriate scale conversion method during the calibration process. RMSE (Root Mean Square Root Mean Square Error Squares) was used. In the process, RMSE statistics were used to select the most suitable equating method from the calibration and equating methods in IRT. Error values were also calculated for each method with the Delta method.

## 3. Findings

### 3.1 The Results of First Research Question

In Table 1 below, the mean and standard deviation values of the X and Y form are given before starting the LE process.

***Table 1***. *Directly Observed Statistics for Data Used to Equate Form X to Y*

| Groups | Forms | N | $\hat{\mu}$ | $\hat{\sigma}$ | Covariate | Correlation |
|---|---|---|---|---|---|---|
| 1 | X | 1921 | 11.314 | 6.076 | 19.820 | 0.957 |
| 1 | V | | 5.965 | 3.412 | | |
| 2 | Y | 1900 | 11.426 | 6.350 | 20.900 | 0.949 |
| 2 | V | | 6.022 | 3.469 | | |

Table 1 shows that the mean score of the X form ($\hat{\mu}$) was found to be 11.31, and the mean of the Y form score ($\hat{\mu}$) was found to be 11.43. The means for common items are 5.97 and 6.02, respectively. While the covariance between the X form scores (total) and the common items in the X form was 19.82, the covariance between the Y form scores (total) and the common items in the Y form was calculated as 20.90.

In this study, equating scores were calculated for Tucker internal joint scores $w_1$=1, $w_1$=0.50 using the LE method. Appendix 4 presents the findings according to the Tucker internal partner method. In Appendix 4, firstly, the findings for $w_1$=1 weight are presented. In this case, it is accepted as $w_2$=0. Raw scores are given in the first column, equating scores are given in the second column, and difference scores are given in the third column. When the equating scores were lower than 0, they were converted to 0, and when they were higher than the highest 25 values to be taken from the booklets, they were converted to scale scores. This process is called cutting (Kolen and Brennan, 2014). Cut-off scores are also used for other equating scores along with this table. In all equating score tables after this table, the difference scores were obtained by subtracting the equating scores from the raw score. For $w_1$=1, the equating scores for the 0-8 score range are lower than the raw scores, and the equating scores in the 9-25 score range are higher than the raw scores. The difference scores calculated for the value of 1 chosen as the weight of the synthetic population ranged from -0.571 to 0.193. For Tucker $w_1$=0.50, the equating scores for the 0-7 score range are lower than the raw scores, and the equating scores in the 8-25 score range are higher than the raw scores. Equating as difference scores for $w_1$=0.5. it was seen that the scores ranged between -0.486 and 0.193. In all equating score tables after this table, the difference scores were obtained by subtracting the equating scores from the raw score.

The equating scores obtained from the LevineOS equating method, which is another LE method that considers the synthetic population weights after the Tucker method, are presented in Appendix 5. In Appendix 5, the equating scores are presented in the table against the raw scores ranging from 0-25. Firstly, $w_1$=1 and $w_2$=0 are accepted. For the LevineOS equating $w_1$=1 weight, the equating scores corresponding to the 0-10 score range are lower than the raw scores, and the equating scores for the 11-25 scores are higher than the raw scores. Difference scores for $w_1$=1 vary between the lowest -0.522 and the highest 0.347. For the Levine observed weight of $w_1$=0.50, the equating scores corresponding to the 0-10 score range are lower than the raw scores, and the equating scores for the 11-25 scores are higher than the raw scores. Difference scores range from -0.5431 to 0.3541.

After LevineOS equating, the equating scores obtained from the LevineTS equating method, in which the actual scores are used, are presented in the Appendix 6. In the LevineTS method, equating scores were obtained regardless of the weights in the synthetic population (Appendix 6). For this equating method, raw scores ranging from 0-25 points were transformed into equating scores. For the LevinTS, the equating scores for the 0-11 score range were lower than the raw scores, and for the 12-25 score range, the equating scores were higher than the raw scores. When the difference scores were

examined, the minimum -2.501 and the highest 1.867 values were calculated. Equating scores and difference scores obtained with this method differed considerably from the equating scores and raw scores obtained from other Tucker and LevineOS methods.

The results of the equating method using the equating function using the classical congeneric model are given in Appendix 7. Braun-Holland equating method was used by making use of the statistical relationship of each item with the common item. The equating scores obtained are given in Table 2.

In Appendix 8, the equating scores obtained by the Braun-Holland equating method of the 0-12 raw score range are given. When the equating score distributions are examined, it is seen that the equating scores in the 0-9 point range are lower than the raw scores, and between 10-12 points, the raw scores get lower values than the equating scores. The difference scores for Braun-Holland were calculated as 1.0665 at the highest and -0.283 at the lowest.

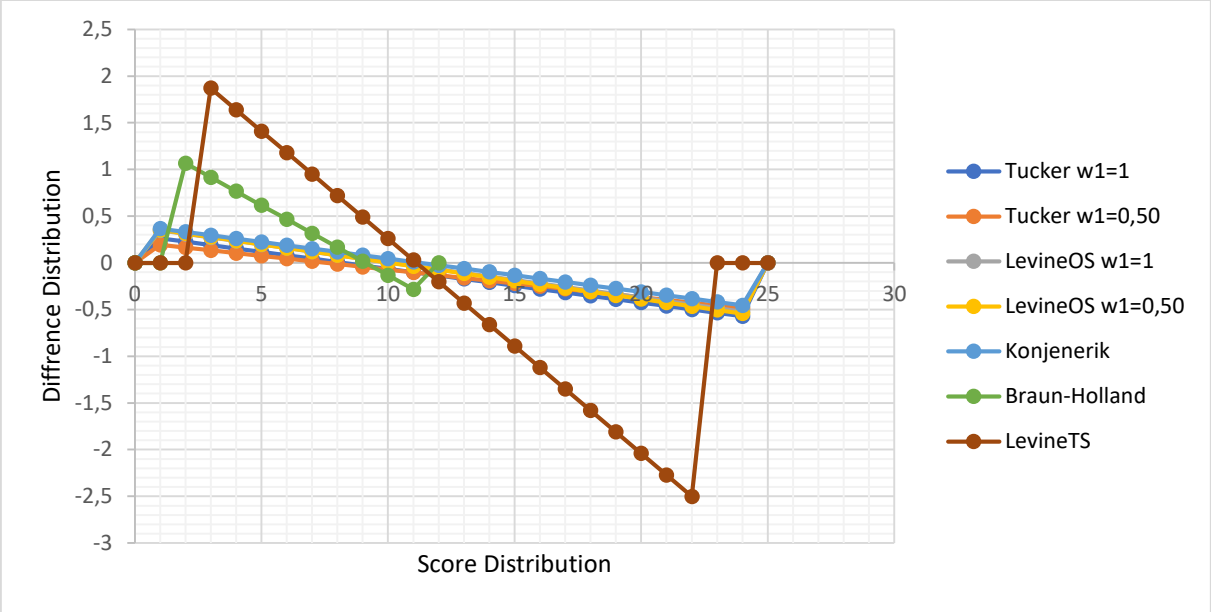
Equating functions for all LE methods in the study were obtained and the equating scores were calculated. Parameter values were found for the calculated equating scores. Calculated parameter values are given in Table 2.

***Table 2.*** *Parameter Values Considered When Using LE Method in CTT*

| w | Equating Methods | $\gamma_1$ | $\gamma_2$ | $m_S(X)$ | $m_S(Y)$ | $\sigma_S(X)$ | $\sigma_S(Y)$ |
|---|---|---|---|---|---|---|---|
| 0.5 | Tucker | 1.702 | 1.737 | 11.363 | 11.376 | 6.123 | 6.304 |
| 0.5 | LevineOS | 0.976 | 1.201 | 11.342 | 11.391 | 6.094 | 6.332 |
| 1 | Tucker | 1.702 | 1.737 | 11.314 | 11.525 | 6.0756 | 6.296 |
| 1 | LevineOS | 0.976 | 1.201 | 11.314 | 11.357 | 6.076 | 6.305 |
| - | LevineTS | 0.976 | 1.201 | 11.365 | 11.356 | 6.106 | 7.475 |
| - | Congeneric | 1.863 | 1.929 | 11.316 | 11.314 | 6.291 | 6.576 |
| - | Braun-Holland | 1.1499 | -1.3664 | 5.3706 | 4.8094 | 8.973 | 11.866 |

LE methods were applied to PISA 2012 data respectively using a common item non-equivalent groups design. The score distributions of LE methods using equating scores are given in Figure 1.

***Figure 1.*** *Score And Difference Distribution For CTT Equating Methods*




In Figure 1 above, the distribution of the difference between the equating scores and the raw scores is given graphically. In all linear methods equating scores tend to have lower values up to 12-13 raw scores, while equating scores for values greater than 13 raw scores tend to have higher values than raw scores. When the graph is examined, it is seen that the equating scores and the deviation from the raw scores in the LevineTS method are significantly different and higher than the other methods. Equating-Error_wg (v2.0) program was used to determine the error coefficients. Errors were calculated using the boostrap method in the program used. Error values for linear methods with 500 replications are given in Table 3.

***Table 3.*** *WMSE Values Obtained from LE Methods*

| LE methods | Weighs of Population | Error Values (Bootstrap method) | Error Values (Delta method) |
|---|---|---|---|
| Tucker | $w_1$=1 | 0.160 | 0.140 |
| | $w_1$=0.50 | 0.177 | 0.154 |
| LevineOS | $w_1$=1 | 0.171 | 0.152 |
| | $w_1$=0.50 | 0.168 | 0.150 |
| LevineTS | - | 0.345 | 0.340 |
| Braun-Holland | - | 0.229 | 0.169 |
| Classical Congeneric | - | 0.194 | 0.164 |

According to Kendall and Stuart (1977), the delta method is a widely used statistical method to derive standard error expressions. The delta method is used to derive the approximate standard error of a statistic which is a statistical function for which expressions for standard errors are already available. Equating errors were calculated using the Taylor expansion for the delta method in the research.

When Table 3 is examined. it is seen that the quantitative order of error values in bootstrap method and delta method did not change. When analyzed quantitatively, it is seen that the error values in the delta method are lower than the values obtained by the bootstrap method. Among the LE methods, the least error value was obtained from the Tucker internal ($w_1$=1) LE method. The maximum error was calculated from the LevineTS equating method. This result is consistent with the graphical representation in Figure 1. Least error $w_1$=1 in Tucker internal equating methods; The maximum error is seen when equal and $w_1$=0.5. In the Levine observed equating method, when equating errors are ranked according to the weights specified, the least error is $w_1$=0.5; the maximum error was found to be w1=1. When ordering from the method with the least errors to the method with the most errors, the order is as follows; Tucker internal($w_1$=1), Levine observed ($w_1$=0.50), Levine observed ($w_1$=1), Tucker internal($w_1$=0.50), classical congeneric, Braun-Holland and LevineTS equating.

### 3.2. The Results of Second Research Question

Smoothing methods should be tried in the equal percentage equating process. In the process, pre-smoothing and then post-smoothing methods were applied and equating scores were obtained. Equating scores obtained from C6 and beta4 methods in the pre-smoothing process are given in Appendix 9. When moments, fit indices and graphical distribution were examined, it was seen that C=6 polynomial degree was appropriate. In the pre-smoothing process, the equating scores obtained from the log-linear methods C6 and beta4 methods are presented in Appendix 9. When the In Appendix 9 was examined the raw scores of the X form are given between 0-25. Standard error values and equal percentile equating scores obtained without pre-smoothing are given. The scores obtained without pre-smoothing range from 0.1380 to 24.088. According to the log-LE calculated according to the C=6 polynomial degree, the equating scores ranged from -0.007 to 25.309, while in beta4 binominal equating scores were calculated between -0.164 and 25.044. The distribution of the log-linear method was within the standard error band with less deviation than the distribution of the beta4 method. The raw score moments for pre-smoothing results are given in Table 4.

***Table 4.*** *Raw Score Moments for Pre-Smoothing*

| Test Forms | μ | σ | Skewness | Kurtosis |
|---|---|---|---|---|
| Form X | 11.314 | 6.074 | 0.233 | 2.037 |
| Form Y | 11.426 | 6.350 | 0.186 | 1.935 |
| X form that equated to Y form | | | | |
| Unsmoothed | 11.423 | 6.346 | 0.187 | 1.934 |
| Beta4 | 11.426 | 6.345 | 0.1873 | 1.936 |
| Log-Linear C=6 | 11.424 | 6.344 | 0.185 | 1.932 |

Table 4 summarizes the unsmoothed, pre-smoothed and suitable polynomial functions. When the parameters Table 4 and the parameters obtained after customization are examined, it is seen that the values are as close as possible to each other.

After defining the appropriate polynomial function in the pre-smoothing method, which is one of the smoothing methods. S parameters are also tested for the final smoothing. Equating scores according to different S values for the final smoothing are given in Appendix 10, which shows that the

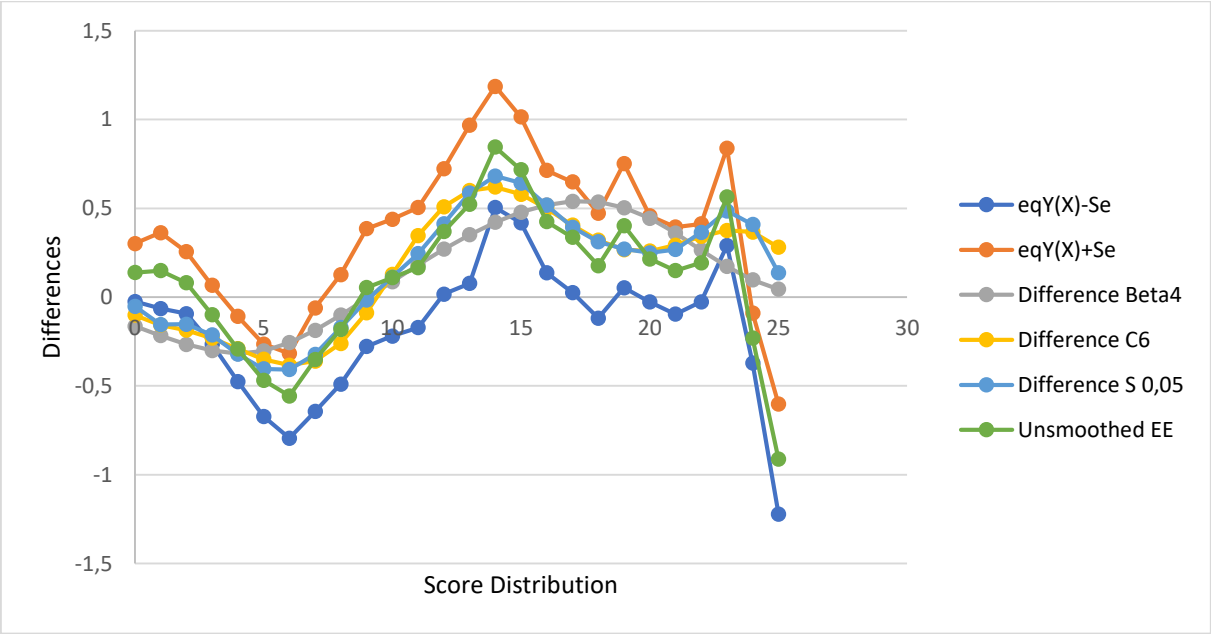
scores obtained according to different S smoothing degrees. Analytical processes for comparing smoothing methods are presented in Table 5 below.

***Table 5**. Raw Score Moments Obtained from Smoothing Methods*

| Test Forms | $\hat{\mu}$ | $\hat{\sigma}$ | Skewness | Kurtosis |
|---|---|---|---|---|
| Form Y | 11.426 | 6.350 | 0.186 | 1.935 |
| Form X | 11.314 | 6.074 | 0.233 | 2.037 |
| X form that equated to Y form | | | | |
| Unsmoothed | 11.423 | 6.346 | 0.187 | 1.934 |
| Beta4 | 11.426 | 6.345 | 0.187 | 1.936 |
| Log-Linear C=6 | 11.424 | 6.344 | 0.185 | 1.932 |
| S=0.05 | 11.426 | 6.346 | 0.188 | 1.935 |

It is seen that beta4 pre-smoothing method is closer to the Y-form values when the moments are examined by looking at the table 5 values to decide which method is the most appropriate smoothing in the process of equating the old form to the new form. When the smoothing methods are analyzed analytically, it is seen that the cubic spline S 0.05 degree provides closer values for the four moments after the beta4 pre-smoothing method. The variation in moments was obtained at most in the Log-Linear C 6 pre-smoothing function. After the analytical process. smoothing methods in the error band gap were compared in the graphical analysis. The distribution of the scores in the error band regarding the score distribution between 0-25 is given in Figure 2 below.

***Figure 2.** Distribution of Difference Values of Smoothing Methods According to Standard Error Band*




In the figure 2, smoothing methods difference values between positive and negative equal percentile error values are given. The pre- and post-smoothing methods used for the raw score 0,1,2,24 and 25 were out of the error band value. In the raw scores of 3, 6, 14 and 18, the difference scores in the beta4 pre-smoothing method were out of the error band gap. The beta4 method, which has a score distribution that goes out of the error band range for the nine points specified, shows a more uniform distribution compared to other smoothing methods. When the difference score

distributions are examined, the second method that shows a uniform distribution in the difference score distributions within the error band values is the log-linear C6 method. When the difference distribution according to the cubic spline S 0.05 degree is examined, although the distribution is sharper than other smoothing methods, the sharpest distribution is obtained in the unsmoothed method.

After the smoothing method was decided, equal percentage equating methods were applied to the data set. It was used in the frequency estimation method as the first equal percentile equating method. In the frequency estimation method, like the linear methods observed by Tucker and Levine, equal percentage equating scores were calculated by using $w_1$=0.5 for the synthetic population and $w_1$=1 weights calculated by proportioning the number of persons between the two different forms, and the results are presented in Appendix 11.

Equating scores are presented using the frequency estimation method in the findings in Appendix 11. $w_1$=0.5 lowest score 0 highest score 11.6069; When the weight is $w_1$=1, the lowest score is 0 and the highest score is 11.6065. Synthetic population weight $w_1$=0.5; When $w_1$=0.1, negative values are generated against raw score 0,1, and 2, while values after raw score 3 (for 4,5,6,7,8,9,10 and 11) are higher than the specified values, obtained by the estimation method. When using different weights of the synthetic population for the raw score 12, values lower than 12 were obtained.

The mean, variance, slope and intercept values of these two calculated EE and the EE calculated without smoothing were calculated for the equating scores. These calculated values are presented in Table 6.

***Table 6.*** *Parameter Values Considered When Using EE in CTT*

| EE Method | Synthetic Population Weights | Slope | Intercept | $\mu_{(X)}$ | $\mu_{(Y)}$ | $\sigma_{(X)}$ | $\sigma_{(Y)}$ |
|---|---|---|---|---|---|---|---|
| Unsmoothed EE | | 0.993 | -0.307 | 4.156 | 3.820 | 10.890 | 10.738 |
| Frequency estimation | $w_1$=0.5 | 1.150 | -1.366 | 5.371 | 4.809 | 8.973 | 11.866 |
| | $w_1$=1 | 1.185 | -2.122 | 5.350 | 4.215 | 8.895 | 12.480 |

In the equal percentage equating method, firstly smoothing methods are tried. Equal percentile equating methods were tried by finding the beta 4 method, which is one of the smoothing methods, has less errors than the other methods. Standard error coefficients were calculated to estimate which of the EE methods used was more appropriate. Therefore, the Equating-Error_wg (v2.0) program was used with 500 replications for each method. Obtained error values are presented in Table 7.

***Table 7.*** *Error Coefficients of CTT EE Methods*

| EE Methods | Synthetic Population Weights | Error Values (Bootstrap method) | Error Values (Delta method) |
|---|---|---|---|
| Unsmoothed EE | - | 0.233 | 0.045 |
| Frequency estimation | $w_1$=0.5 | 0.159 | 0.038 |
| | $w_1$=1 | 0.120 | 0.040 |

When Table 7 is examined, it is seen that the error value calculated for the EE method without smoothing is 0.2330. When the program outputs are examined, the errors calculated with the bootstrap, the standard error of the equating for the frequency estimation method using the $w_1$=0.5 weight is 0.159; For the frequency estimation method using the $w_1$=1 weight, the standard error of the equating was found to be 0.120. For the new error values calculated according to the Taylor series function of the delta method, it was found to be 0.045 for the unsmoothed EE 0.040 for the frequency estimation $w_1$=1 weight, and 0.038 for the $w_1$=0.5. When the coefficients were examined, it was seen that the frequency estimation method using the weight of $w_1$=0.5, one of the equal percentage equating methods applied for the common item non-equivalent groups design, equated with less errors.

The Tucker internal, LevineOS, LevineTS, classical congeneric and Braun-Holland equating methods were applied to the scores of the participants who took booklet 1 and booklet 3, which contains PISA 2012 data. Equating was made by trying different synthetic population weights ($w_1$=0.5; $w_1$=1) in the linear methods observed by Tucker and Levine, one of the applied linear methods. In EE, frequency estimation methods ($w_1$=0.5; $w_1$=1) were tried. The frequency estimation method has been examined in detail in the context of different synthetic population weights ($w_1$=0.5; $w_1$=1) such as Tucker and LevineOS equating methods in LE. Equating-Error_wg (v2.0) program calculated the standard error of all CTT equating methods. All methods and their equating errors are presented in Table 8 below.

***Table 8.*** *Error Coefficients of Linear and EE Methods in CTT*

| | Equating Methods | Synthetic Population Weights | Error Values (Boostrap method) | Error Values (Delta method) |
|---|---|---|---|---|
| LE | Tucker | $w_1$=1 | 0.160 | 0.140 |
| | | $w_1$=0.50 | 0.177 | 0.154 |
| | LevineOS | $w_1$=1 | 0.171 | 0.152 |
| | | $w_1$=0.50 | 0.168 | 0.150 |
| | LevineTS | - | 0.345 | 0.340 |
| | Braun-Holland | - | 0.229 | 0.169 |
| | Classical Congeneric | - | 0.194 | 0.164 |
| EE | Unsmoothed EE | - | 0.233 | 0.045 |
| | Frequency Estimation | $w_1$=0.5 | 0.159 | 0.038 |
| | | $w_1$=1 | 0.120 | 0.040 |

In the findings in Table 8, the equating methods of the CTT and the equating errors of the equating methods are included. When the equating errors are examined, it is seen that the least equating error is obtained with the EE method, and the highest equating error is obtained with the linear matching method. When the equating error valuesare examined in detail, when a correct order is made from the method with the least errors to the method with the most errors, the order is as follows; frequency estimation method ($w_1$=0.5), Tucker internal($w_1$=1), LevineOS ($w_1$=0.5), Tucker internal ($w_1$=0.5), classical congeneric, frequency estimation ($w_1$=1), Braun -Holland, EE and LevineTS are equating without smoothing.

### 3.3 The Results of Third Research Question

Before equating in IRT, scale transformation was done by using common items. Moments for scale transformation (calibration) and transformation coefficients for characteristic curve transformation methods were calculated with ST 2.0. The findings of the transformation constants are given in Table 9.

***Table 9.*** *Conversion Coefficients and Conversion Constants Obtained from Calibration Methods*

| Calibration Methods | A | B |
|---|---|---|
| Mean-mean | 0.993 | -0.039 |
| Mean-standard deviation | 0.986 | -0.036 |
| Stocking-Lord | 0.957 | 0.010 |
| Haebara | 0.954 | 0.014 |

The calibration method with the least error scale values among the specified calibration methods was determined. Calculated error values are presented in Table 10.

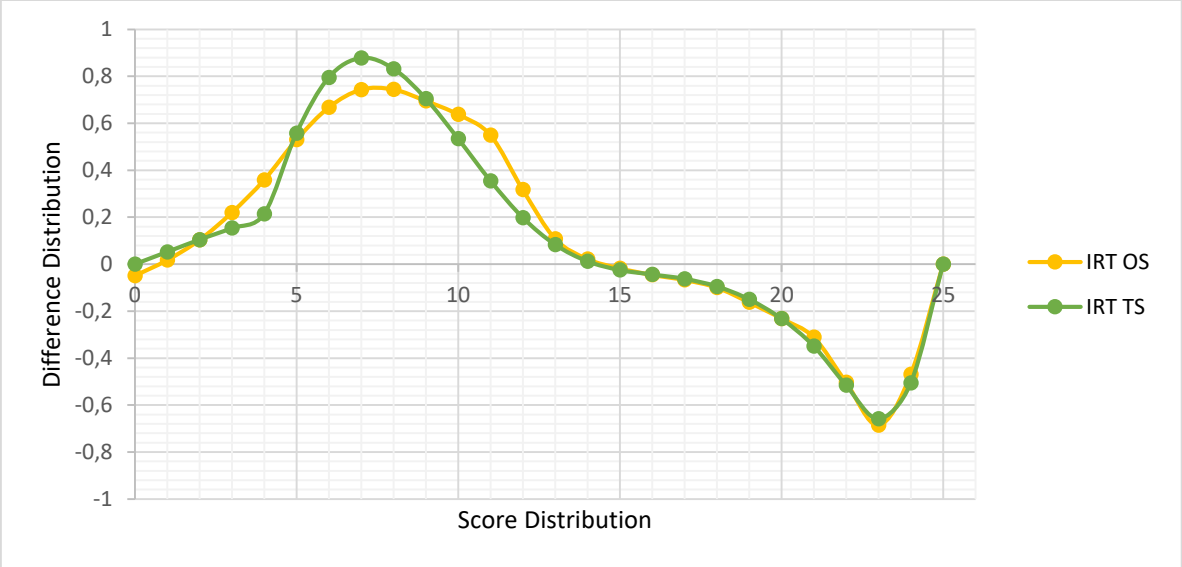
***Table 10.*** *Calculated Error Coefficients for Calibration Methods*

| Calibration Methods | Error Coefficients |
|---|---|
| Mean-mean | 0.039 |
| Mean-standard deviation | 0.036 |
| Stocking-Lord | 0.035 |
| Hebara | 0.038 |

When the error values in Table 10 are examined, the Stocking-Lord method, one of the calibration methods, allows the capabilities to be positioned on the same scale with the least error. The highest error was obtained with the mean-mean calibration method. After the calibration method was chosen, the abilities were brought to the same scale with the Stocking-Lord calibration method, and the scores obtained from the equating methods used in IRT are given in Appendix 12.

When the findings presented in Appendix 12 are examined, the ability values could not be calculated for the 0, 1st, 2nd, 3rd and 25th scores for the TS equating based on IRT. The estimation totals of the calculated c parameter values were found to be 3.062 (Appendix 12). For this reason, ability values below 3 points were not estimated. When the chart above is examined, it is seen that the raw score for true score equating is lower than the scores equal to 15, and it is higher than the Y form up to 25 points including 15. The actual score equivalent of 25 raw scores was calculated as 25. The distributions according to the difference values between the equating scores and raw scores in IRT are presented in the Figure 3

***Figure 3.*** *Distribution of Difference Scores in IRT Observed and Actual Score Equating Methods*




When Figure 3 above is examined, the difference was calculated as 0, since the equating scores for raw scores 0 and 25 in IRT TS equating methods were equal to these extreme values. While there was a linear increase in difference scores up to raw score 4 for the IRT TS, the difference scores changed and increased rapidly after raw score 4. The difference value for the IRT TS reached the highest value at 7 for the raw score for the IRT TS. The raw score showed a significant decreasing trend to 14. Raw scores from 14 to 24 equating scores are lower than raw scores. In the IRT OS equating methods, the raw score was equating to the values of 0 and 25 by assigning the cutoff score. Just like the IRT TS, the raw score tends to increase up to 7, while the raw score tends to decrease up to 14, in the IRT OS equating. After the raw score of 14, the equating scores were calculated to be lower than the raw score. Equating scores were calculated with the IRT TS and IRT OS equating method equations and tried to be interpreted graphically. The standard error of equating of IRT equating methods is calculated. All methods and their equating errors are presented in Table 11.

***Table 11****. IRT Equating Methods and Error Coefficients of These Method*

| IRT Equating Methods | Error Coefficients |
|---|---|
| IRT TS | 0.0111 |
| IRT OS | 0.0118 |

In Table 11, the error coefficients included in the equating methods considered within the scope of the IRT are given. It was seen that the most reliable equating results with the least errors were obtained by the IRT TS equating method. The IRT TS equating method is a more appropriate equating method with less error than the IRT OS equating.

### 3.4. The Results of Fourth Research Question

Equating methods in CTT and IRT were compared according to the quantities of equating errors. All equating methods and their calculated equating errors are given in Table 12 by grouping them. The error values grouped according to the measurement theories and equating methods used are presented in Table 12.

***Table 12.*** *Error Coefficients of Equating Methods for CTT and IRT*

| Theory | Equating | Methods | Synthetic Population Weights | Error Values |
|---|---|---|---|---|
| CTT | LE | Tucker | $w_1$=1 | 0.1604 (Bootstrap) |
| | | | | 0.1404 (Delta) |
| | | | $w_1$=0.50 | 0.1765 (Bootstrap) |
| | | | | 0.1537 (Delta) |
| | | LevineOS | $w_1$=1 | 0.1709 (Bootstrap) |
| | | | | 0.1515 (Delta) |
| | | | $w_1$=0.50 | 0.1684 (Bootstrap) |
| | | | | 0.1502 (Delta) |
| | | LevineTS | - | 0.3448 (Bootstrap) |
| | | | | 0.3403 (Delta) |
| | | Braun-Holland | - | 0.2286 (Bootstrap) |
| | | | | 0.1689 (Delta) |
| | | Classical Congeneric | - | 0.1943 (Bootstrap) |
| | | | | 0.1636 (Delta) |
| | EE | Unsmoothed EE | - | 0.2330 (Bootstrap) |
| | | | | 0.0445 (Delta) |
| | | Frequency Estimation | $w_1$=0.5 | 0.1589 (Bootstrap) |
| | | | | 0.03814 (Delta) |
| | | | $w_1$=1 | 0.1995 (Bootstrap) |
| | | | | 0.04012 (Delta) |
| IRT | | IRT TS | - | 0.0111 |
| | | IRT OS | - | 0.0118 |

When Table 12 is examined, it is seen that all the equating methods used for this study of the two measurement theories used have equating error values. While determining the most appropriate method for equating in other sub-problem statements before this sub-problem statement, Table 12 was examined in line with the comments made. First of all, when the methods in CTT were examined, it was found that the Tucker equating method, which used $w_1$=1 synthetic weight from LE methods, obtained scores equating with the least error, and the least incorrectly equating scores were obtained with the frequency estimation method, which was one of the EE methods, where $w_1$=0.5 synthetic weight was used. When CTT equating methods are compared by looking at their error values, it is seen that the frequency estimation method, which is one of the EE methods, equates with less errors. For LevineTS equating, which is one of the LE methods, the scores equaled with the most errors were obtained. Examination of the TS and OS equating methods calculated from IRT showed that the IRT TS equating achieved equating scores with fewer errors. As for the error coefficients in Table 12, the equating methods belonging to IRT are obtained with less errors than all the equating methods in CTT. When these methods, which are equated with the least error in both theories, are examined in terms of error quantities, it is seen that the most appropriate equating scores are obtained with the least error in the IRT TS method.

## 4. Discussion and Conclusion

### 4.1. Conclusions and Discussions on The First Research Question

Equating error values were examined for LE methods based on CTT. When the error values obtained are compared quantitatively, the order of the methods with the least errors to the methods with the most errors is as follows; Tucker internal ($w_1$=1). LevineOS ($w_1$=0.5), LevineOS ($w_1$=1), Tucker internal ($w_1$=0.5), classical congeneric, Braun-Holland, and LevineTS equating. When the equating error values are examined, it is seen that the most appropriate equating method with the least error for LE methods is Tucker internal ($w_1$=1), and the Levine total score equating method has the highest error.

One of the LE methods, Levine's true score equating method was found to be the worst equating method. For the Levine true score, it was observed that the equating scores at the extreme values were more differentiated than the raw score, and the calculated difference values were different compared to other linear methods. Although Levine's actual mean score and the observed mean score are derived from the assumption that the observed mean score is similar, it is seen that the estimated error value is too high due to the difference values obtained in this study. Theoretically, the true score is obtained by adding the plus and minus error value to the observed score (Spearman, 1907). It is striking that the difference values between the equating score and the raw score for the LevineTS in the research are large. When the distribution of the difference scores is examined, the fact that the change is high is explained in the findings section of the research. When the variability of the difference values is interpreted for the error distribution, it can be concluded that the Levine true score is the method with the most errors. Similar results in the literature are in line with the results of the study conducted by Chen et al., (2011). It was concluded that the difference scores for the LevineTS did not produce a linear function but increased the error value. Contrary to this result, Hanson et al., (1993) found in their study that the Levine true score had less error than the Levine observed, İnal & Akın Arıkan (2017) found the similar result that Tucker has less equating error than Levine methods.

### 4.2. Conclusions and Discussions on The Second Research Question

For the second research question of the research, EE methods related to CTT were applied. Before the EE method, it was decided which of the smoothing methods was appropriate. In the smoothing methods. C 6 polynomial function and beta4 binomial method were found suitable for pre-smoothing, while S 0.05 degree was chosen for final smoothing. It was investigated which smoothing method had less errors and it was seen that the best smoothing method was beta4 binominal pre-smoothing, while the C 6 degree pre-smoothing method contained the most errors. The information that the beta4 binomial function used in the pre-smoothing for EE has less errors is in line with the results of the study by Livingstone (1993), Kahraman (2012) and Tan (2015).

Equating equations were found by using the frequency estimation method ($w_1$=1; $w_1$=0.5). Equating scores were calculated with the obtained equations. When the equating scores calculated in the frequency estimation method for different weights were examined, the equating scores calculated against the raw scores of 0,1 and 2 were equating to 0 using the cut-off score. For all EE methods used, a strong positive correlation was found between equating scores and raw scores. Equating error values were examined for EE equating methods based on CTT. When the error values found are compared quantitatively, the order of the methods with the least errors to the methods with the most errors is as follows; frequency estimation ($w_1$=0.5), frequency estimation ($w_1$=1) and unsmoothed EE. When the error values of the EE methods in the CTT were examined, it was seen that the most appropriate

equating method with the least error was frequency estimation ($w_1$=0.5) and the most error was the non-smoothed EE method.

The results obtained from the second research question of the research Hagge et al., (2011), Livingstone et al., (1990), Livingstone (1993), Livingston and Feryok (1987), Skaggs (2005). This is in line with the studies of Kolen (1988). In these studies, it was stated that the frequency estimation method produced more accurate results than other EE methods. It has been determined that equating scores calculated by frequency estimation method tend to give more accurate results when a large sample is used within the scope of the research (Livingstone & Feryok, 1987).

As a LE method, equating scores were obtained by using Tucker internal, LevineTS, LevineOS, classical congeneric model and Braun-Holland equating methods. In Classical Test Theory, frequency estimation and unsmoothed EE methods were used for EE, and equating scores were calculated. Equating score distributions are explained in the results of the first and second research questions above. When the presented error values are compared quantitatively, the order of the methods with the least errors to the methods with the most errors is as follows; frequency estimation ($w_1$=0.5), Tucker internal($w_1$=1), LevineOS ($w_1$=0.50), LevineOS ($w_1$=1), Tucker internal ($w_1$=0.5), classical congeneric, frequency estimation ($w_1$=1), Braun-Holland and LevineTS equating. When the equating error values in the CTT were examined, it was seen that the most appropriate equating method with the least error was frequency estimation ($w_1$=0.5), and the LevineTS equating method with the most errors. When the CTT equating methods used are compared, it is concluded that EE is suitable with less errors.

Kolen and Brennan (1995), Mutluer and Nartgün (2017), Pektaş and Kılınç (2016) and von Davier (2008), found in their research that EE produces more accurate results than LE method and the result of this research shows similarity with the result that it has fewer errors. The results of this research do not overlap with the results of Wang. et al. (2008), or Kelecioğlu and Gübeş (2013). In these studies, it was understood that the LE method produced more accurate results. In the literature, Kolen and Brennan (2014) found that EE produces more accurate results in large samples; It has been clearly stated that the difficulty differences between the forms make more harmonious equating since they involve the conversion process with percentiles in the drawn curves. Çörtük (2022) found the EE method is more accurate for equating process.

### 4.3. Conclusions and Discussions on The Third Research Question

The error coefficients of the calibration methods used in the same scale conversion process were calculated. When the calculated error coefficients were compared quantitatively, the highest error was obtained from the mean-mean method, and the least error was obtained from the Stocking-Lord method. Aksekioğlu (2017), Demirus (2015), Karkee and Wright (2004), Kilmen (2010), Spearman (1907), Stocking and Lord (1982) and Yurtçu and Güzeller (2018). It was stated that scale conversion processes based on item characteristic curves are more appropriate. It has been stated that the characteristic curve methods have a structure that eliminates the mismatch (Stocking & Lord, 1982). Stocking-Lord is more durable in the differences of the ability parameter (Keller, 2007). On the contrary, it has been observed in the studies conducted by Gök (2012), Gündüz (2015) and Tanberkan Suna (2018) that the mean-mean calibration method is also suitable. With Salmaner Doğan (2022) research found that Stocking-Lord calibration method is suitable when the difficulty among the forms is less.

After determining the appropriate calibration method, true and observed score equating based on IRT was made. When the equating error values were compared quantitatively, it was

concluded that the IRT true score equating method produced a more robust solution and equated with less errors.

When the available literature is scanned, IRT is observed in the studies of Aksekioğlu (2017), Hagge et al. (2011), Han et al. (1997), Lord and Wingersky (1984), Tanberkan Suna (2018). They found that the score had fewer equating errors. In the true score equating process, it is accepted that the true score is a combination of the observed and true score. It assumes that individuals at the same ability level have the same true score in the equating process. In the observed score equating, a particular group is focused. The score distribution of this group is placed on a common scale by ensuring that its characteristics are equal (von Davier, 2008). Based on this explanation, Gündüz (2015), and Kumlu (2019) IRT continued to work with the TS equating method, and IRT reported that the true score had fewer errors as a result of the study. Keller (2007) stated in their studies that IRT parameters calculated in calculations related to the actual score give more consistent results. In addition, Kolen & Brennan (1995) explained that the superiority of the IRT TS equating method over the IRT OS equating method is that it is easy to calculate, and the transformation obtained can be obtained independently of the group's ability distribution, and its limitation is that it equates the true scores that do not exist in practice.

### 4.4. Conclusions and Discussions on The Fourth Research Question

In this research, equating methods within the scope of CTT and IRT are included. Among the CTT LE methods, LevineTS equating method was the worst equating method, while the Tucker equating method with a synthetic population weight of 1 was determined as the best equating method with the least error. Among the CTT and EE methods, it was determined that the frequency estimation method, which was processed with the synthetic population weight of 0.5, was the worst equating method with the highest equating error value without smoothing, and the best equating method with the least error. When the TS and OS methods are taken into consideration and IRT equating methods are compared, it has been determined that the TS equating method with the least error in IRT is a more appropriate and powerful equating method. When the error values of the equating methods in CTT and IRT are examined quantitatively, the order from the one with the least error to the equating method with the most error is as follows; IRT TS, IRT OS, frequency estimation ($w_1$=0.5), frequency estimation ($w_1$=1), Tucker ($w_1$=1), LevineOS ($w_1$=1), LevineOS ($w_1$=0. 50), Tucker ($w_1$=1), classical congeneric, Braun-Holland, unsmoothed EYE, LevineTS.

Comparison of theories has been considered as the aim of many studies, Petersen et al. (1983), Lord and Wingersky (1984), Han et al. (1997), Hagge et al. (2011) Liu and Kolen (2011) and also Tanberkan Suna (2018) compared IRT and CTT equating methods in their study. It has been seen that the results obtained and the results of this research are in parallel with the IRT equating methods, giving more accurate results with less errors. On the other hand, the results don’t support Olaginan et al (2022) research. The difference in this research is mainly about the sample size. If the sample size is not big enough, IRT equating process includes large error values.

In this part of the research, suggestions are explained for those who will work on test equating. As a new study subject. scores from different booklets can be recalculated with the IRT TS equating method as a result of this research. Although PISA focuses on a different area every three years, the scores obtained from the learning area outside the target learning area of PISA on the date specified during the research process can be equated. In this research, a study was carried out on 4 countries in line with the purpose of the research. It is recommended to perform a new equating study with a new sample representing the 65 countries participating in the PISA 2012 application or with all the scores

related to the universe. In PISA, the scores of individuals are calculated over the Rasch model (OECD, 2014). In the synchronization process used, 3PLMs were used for IRT. According to the results of the research, it is recommended to recalculate student scores using 3PLM and recalculate the test equating process and success order. Within the scope of the study, it was observed that the conversion process to scale scores with the Stocking-Lord method in the IRT-based equating process led to equating with the least error. For this reason, it is recommended that practitioners conduct equating by performing scale conversion with Stocking-Lord. Based on the results of this research, which was carried out to compare the theoretical equating methods, it is recommended to use this method in the case of being free from errors in the IRT TS equating method and in the equating of scores for different situations.

**References**

Aksekioğlu, B. (2017). *Madde tepki kuramına dayalı test eşitleme yöntemlerinin karşılaştırılması: PISA 2012 fen testi örneği* (Yayın No. 454879) [Yüksek lisans tez, Akdeniz üniversitesi]. YÖK. https://tez.yok.gov.tr/UlusalTezMerkezi/

Angoff, W.H. (1987). Technical and practical issues in equating: A discussion of four papers. *Applied Psychological Measurement, 11,* 291-300.

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.

Brossman, B. G., & Lee, W.C. (2013). Observed score and true score equating procedures for multidimensional item response theory. *Applied Psychological Measurement, 37*(6), 460-481. https://doi.org/10.1177/0146621613484083

Büyüköztürk, Ş., Çakmak, E. K., Akgün, Ö. E., Karadeniz, Ş., & Demirel, F. (2008). *Bilimsel araştırma yöntemleri.* Pegem Akademi.

Chen, H. H., Livingston, S. A., & Holland, P. W. (2011). Generalized equating functions for NEAT designs. *Statistical models for test equating, scaling and linking*, 185-200.

Cook, L. L.,& Eignor, D. R, (1991). IRT equating methods. *Educational Measurement: Issues And Practice, 10*(3), 37-45.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Harcourt Brace Javonich College.

Çörtük, M. (2022). *Çok kategorili puanlanan maddelerden oluşan testlerde Klasik Test Kuramı ve Madde Tepki Kuramı'na dayalı test eşitleme yöntemlerinin karşılaştırılması* (Yayın No. 743619) [Yüksek lisans tez, Akdeniz üniversitesi]. YÖK. https://tez.yok.gov.tr/UlusalTezMerkezi/

Demirus. K. B. (2015). *Ortak maddelerin değişen madde fonksiyonu gösterip göstermemesi durumunda test eşitlemeye etkisinin farklı yöntemlerle incelenmesi* (Yayın No. 399468) [Doktora tezi, Ankara üniversitesi]. YÖK. https://tez.yok.gov.tr/UlusalTezMerkezi/

Dorans, N. J., & Holland, P. W. (2000). Population invariance and the equatability 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

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists.* Lawrence Erlbaum.

Felan, G. D. (2002). Test Equating: Mean, Linear, Equipercentile, and Item Response Theory. *Annual Meeting of the Southwest Educational Research Association*, 1-24.

Gök, B. (2012). *Denk olmayan gruplarda ortak madde deseni kullanılarak madde tepki kuramına dayalı eşitleme yöntemlerinin karşılaştırılması* (Yayın No. 321947) [Doktora tezi, Ankara üniversitesi]. YÖK. https://tez.yok.gov.tr/UlusalTezMerkezi/

Gulliksen, H. (1950). The reliability of speeded tests. *Psychometrika, 15*(3), 259-269.

Gündüz, T. (2015). *Test eşitlemede Madde Tepki Kuramına dayalı yetenek parametresine yönelik ölçek dönüştürme yöntemlerinin karşılaştırılması* (Yayın No. 429524) [Yüksek Lisans Tezi, Gazi Üniversitesi]. YÖK. https://tez.yok.gov.tr/UlusalTezMerkezi/

Hagge, S. L., Liu, C., He, Y., Powers, S. J., Wang, W., & Kolen, M. J. (2011). A comparison of IRT and traditional equipercentile methods in mixed-format equating. *Mixed-Format Tests: Psychometric Properties With A Primary Focus On Equating, 1*, 19-50.

Hambleton, R. K.,& Swaminathan, H. (2013). *Item response theory: Principles and applications.* Springer Science & Business Media.

Han, T., Kolen, M., & Pohlmann, J. (1997). A comparison among IRT true-and observed-score equatings and traditional equipercentile equating. *Applied Measurement in Education, 10*(2), 105-121.

Hanson, B. A., Zeng, L., & Colton, D. A. (1994). *A comparison of presmoothing and postsmoothing methods in equipercentile equating* (Vol. 94. No. 4). American College Testing Program.

Hanson, B. A., Zeng, L., & Kolen, M. J. (1993). Standard errors of Levine linear equating. *Applied Psychological Measurement, 17*(3), 225-237.

İnal, H. & Akın Arıkan, Ç. (2017). An investigation of group invariance in test equating according to gender. *Journal of Measurement and Evaluation in Education and Psychology, 8*(1), 128-145.

Kahraman, H. (2012). *Düzgünleştirilmiş puanların eşitleme hatasına etkisi* (Yayın No. 314954) [Yüksek lisans tezi, Hacettepe Üniversitesi]. YÖK. https://tez.yok.gov.tr/UlusalTezMerkezi/

Kane, M. T., Mroch, A. A., Suh, Y., & Ripkey, D. R. (2009). Linear equating for the NEAT design: Parameter substitution models and chained linear relationship models. *Measurement, 7*(4), 125-146. https://doi.org/10.1080/15366360903418022

Karasar, N. (2005). *Bilimsel araştırma yöntemi.* Nobel Yayın Dağıtım.

Karkee, T. B., & Wright, K. R. (2004). *Evaluation of linking methods for placing three-parameter logistic item parameter estimates onto a one-parameter scale.* Online Submission.

Keller, R. R. (2007). *A comparison of item response theory true score equating and item response theory-based local equating.* University of MassachusettsAmherst.

Kendall, M., & Stuart, A. (1977). *The advanced theory of statistics.* Macmillan.

Kelecioğlu, H.. & Öztürk Gübeş, N. (2013). Comparing linear equating and equipercentile equating methods using random groups design. *International Online Journal of Educational Sciences, 5*(1), 227-241.

Kilmen, S. (2010). *Madde Tepki Kuramına dayalı test eşitleme yöntemlerinden kestirilen eşitleme hatalarının örneklem büyüklüğü ve yetenek dağılımına göre karşılaştırılması* (Yayın No. 279926) [Doktora tezi, Ankara Üniversitesi]. YÖK. https://tez.yok.gov.tr/UlusalTezMerkezi/

Kolen, M. J. (1988). Traditional equating methodology. *Educational measurement: Issues and practice, 7*(4), 29-37.

Kolen, M., & Brennan, R. (1995). *Test equating: methods and practices*. Springer-Verlag.

Kolen, M. J., & Brennan. R. L. (2014). *Test equating. scaling. and linking: Methods and practice.* Springer Science and Business Media.

Kumlu, G. (2019). *Test ve alt testlerde eşitlemenin farklı koşullar açısından incelenmesi* (Yayın No. 584462) [Doktora tezi, Hacettepe Üniversitesi]. YÖK. https://tez.yok.gov.tr/UlusalTezMerkezi/

Levine, R. (1955). *Equating the score scales of alternate forms administered to samples of different ability.* ETS Research Bulletin Series, 1955(2), i-118.

Liu, C., & Kolen, M. J. (2011). A comparison among IRT equating methods and traditional equating methods for mixed-format tests. *Psychometric properties with a primary focus on equating, 1*, 75-94.

Livingston, S. A. (1993). Small-sample equating with log-linear smoothing. *Journal of Educational Measurement, 30*(1), 23-39.

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.

Livingston, S. A., & Feryok, N. J. (1987). Univariate vs. bivariate smoothing in frequency estimation equating. *ETS Research Report Series, 1987*(2), 1-21.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Erlbaum.

Lord, F. M., & Wingersky, M. S. (1984). Comparison of IRT true-score and equipercentile observed-score" equatings". *Applied Psychological Measurement, 8*(4), 453-461.

Lucke, J. F. (2005). The α and the ω of congeneric test theory: An extension of reliability and internal consistency to heterogeneous tests. *Applied Psychological Measurement, 29*(1), 65-81. https://doi.org/10.1177/0146621604270882

MEB (2013). *PISA 2012 ulusal ön raporu*. Ankara: MEB Yenilik ve Eğitim Teknolojileri Genel Müdürlüğü.http://odsgm.meb.gov.tr/test/analizler/docs/pisa/pisa2012-ulusal-on-raporu.pdf sayfasından erişilmiştir.

MEB (2015). *PISA 2012 araştırması ulusal nihai raporu.* https://drive.google.com/file/d/0B2wxMX5xMcnhaGtnV2x6YWsyY2c/view?pref=2 &pli=1 Erişim tarihi: 5 Ocak 2017

Mutluer, C., & Nartgün, Z. (2017). Test equating study concerning to ALES (Academic Personnel and Postgraduate Education Entrance Exam) scores obtained at different times in a year. *European Journal of Education Studies, 12*, 96–120.

OECD (2014). *PISA 2012 technical report.* OECD Publishing https://www.oecd.org/pisa/pisaproducts/PISA-2012-technical-report-final.pdf Erişim tarihi: 10 Mart 2018

Olanigan, N. A., Adediwura, A. A., & Ogunsanmi, O. A. (2022). Linear and separate calibration methods of equating continuous assessment scores of public and private elementary schools. *Journal of Integrated Elementary Education, 2(*2), 117-129.

Özdemir, B. (2017). Equating TIMSS mathematics subtests with nonlinear equating methods using neat design: circle-arc equating approaches. *International Journal of Progressive Education, 13*(2), 116-132.

Pektaş, S., & Kılınç, M. (2016). PISA 2012 matematik testlerinden iki kitapçığın gözlenen puan eşitleme yöntemleri ile eşitlenmesi. *Mehmet Akif Ersoy Üniversitesi Eğitim Fakültesi Dergisi,* (40), 432-444.

Petersen, N. S., Cook, L., & Stocking, M. L. (1983). IRT versus conventional equating methods: A comparative study of scale stability. *Journal of Educational Statistics, 8*(2), 137-156.

Petersen, N. S., Kolen, M. J., & Hoover, H. D. (1989). Scaling. norming. and equating. *Educational Measurement, 3*, 221-262.

Salmaner Doğan, R. (2022). *Meta analitik test eşitleme yönteminin çeşitli değişkenler açısından incelenmesi: TIMMS 2015 örneği* (Yayın No. 765807) [Doktora tezi, Gazi Üniversitesi]. YÖK. https://tez.yok.gov.tr/UlusalTezMerkezi/

Sezer Başaran, E. (2023) *Farklı ortak değişkenlerle test eşitlemenin ortak maddeli test eşitlemeyle karşılaştırılması* (Yayın No. 788504) [Doktora tezi, Gazi Üniversitesi]. YÖK. https://tez.yok.gov.tr/UlusalTezMerkezi/

Skaggs, G. (2005). Accuracy of random groups equating with very small samples. *Journal of Educational Measurement, 42*(4), 309-330. https://doi.org/10.1111/j.1745-3984.2005.00018.x

Spearman, C. (1907). Demonstration of formular for true measurement of correlation. *American Journal of Psychology, 18*, 161-169.

Stocking, M. L., & Lord, F. M. (1982). *Developing a common metric in item response theory*. ETS Research Report Series, 1982(1), i-29.

Tan, Ş. (2015). Küçük örneklemlerde beta4 ve polynomial loglineer öndüzgünleştirme ve kübik eğri sondüzgünleştirme metotlarının uygunluğu. *Gazi Üniversitesi Gazi Eğitim Fakültesi Dergisi, 35*(1), 123-151.

Tanberkan-Suna, H. (2018). *Grup değişmezliği özelliğinin farklı eşitleme yöntemlerinde eşitleme fonksiyonları üzerindeki etkisi* (Yayın No. 527064) [Doktora tezi, Gazi Üniversitesi]. YÖK. https://tez.yok.gov.tr/UlusalTezMerkezi/

Thorndike,R. L. (1982). *Applied psychometrics.* Houghton Mifflin.

Yurtçu, M., & Güzeller, C. O. (2018). Investigation of equating error in tests with differential item functioning. *International Journal of Assessment Tools in Education, 5*(1), 50-57. https://doi.org/10.21449/ijate.316420

von Davier, A. A. (2008). New results on the linear equating methods for the nonequivalent-groups design. *Journal of Educational and Behavioral Statistics, 33*(2), 186-203.

von Davier, A. A., & Kong, N. (2005). A unified approach to linear equating for the nonequivalent groups design. *Journal of Educational and Behavioral Statistics, 30*(3), 313-342.

Wang, T., Lee, W.C., Brennan, R. L., & Kolen, M. J. (2008). A comparison of the frequency estimation and ed equipercentile methods under the common-item nonequivalent groups design. *Applied Psychological Measurement, 32*(8), 632-651.

Zhu, W. (1998). Test equating: What, why, how? *Research Quarterly for Exercise and Sport, 69*(1), 11-23.

## Geniş Özet

### Giriş

Farklı zaman ve koşullarda yapılan sınavlardan elde edilen sonuçlar kişilerin kuruma yerleşmesi, kurumda yükselmesi, eğitim düzeyi hakkında bilgi vermek amaçlı kullanılmaktadır. Bu nedenle yapılan sınav sonuçları birden fazla yıllarda da geçerliği korunduğu için test sürecinin standart uygulama koşullarına sahip olması istenir. Bu standart koşullarda test sonuçlarının karşılaştırılabilmesi için eşitlenmesi ve birbiri yerine dönüşümü sağlanmalıdır. Standart ve eşitlenebilir özellikler için farklı sınav sonuçlarının birbiri yerine kullanılması için istatistiksel süreç test eşitleme ile mümkündür. Bu araştırmada ortak maddeli eşdeğer olmayan grup deseni kullanılarak puanlar eşitlenmeye çalışmıştır. Bu araştırmada KTK ve MTK bünyesindeki eşitleme yöntemleri karşılaştırılarak en az hata değerinin hangi eşitleme yönteminden elde edildiği belirlenmeye çalışılmıştır. Klasik test kuramında lineer eşitleme Tucker ($w_1 = 1$, $w_1 = 0.5$), Levine gözlenen puan ($w_1 = 1$, $w_1 = 0.5$), Levine gerçek puan, klasik konjenerik ve Braun-Holland yöntemleri kullanılmıştır. Klasik Test Kuramına bağlı eşityüzdelikli eşitleme (EYE) yöntemleri için düzgünleştirilmeden EYE, ön düzgünleştirme (C 6 polinom derecesi, beta4), son düzgünleştirme (S = 0.05). frekans kestirim yöntemleri seçilmiştir. Madde Tepki Kuramına dayalı eşitleme yapabilmek için öncelikle kalibrasyon yapılmıştır daha sonra gerçek ve gözlenen puan eşitleme yöntemleri uygulanmıştır.

### Yöntem

Çalışmada KTK ve MTK'daki eşitleme yöntemlerinden en az hata değerine sahip olan eşitleme yönteminin belirlenmesine odaklanıldığından betimsel araştırma niteliği taşımaktadır.

Bu araştırmada PISA 2012 testine katılan en iyi performans gösteren ülke olarak Şangay / Çin, en başarısız ülke (kitapçık eşleşmesi koşulu için) Peru, ortalama altında yer alan Türkiye, genel başarı düzeyindeki Finlandiya ülkelerinde bulunan ve kitapçık1 (N = 1921) - kitapçık3 (N = 1900) için toplam 3821 kişi bulunmaktadır. Kitapçık 1 ve kitapçık 3 için 13 madde ankor madde olarak. 12 madde ise ankor olamayan madde olarak ele alınmıştır.

KTK'daki eşitleme yöntemlerinden en az hata değerine sahip olanını belirlemek için WMSE (Weigted Mean Square Error - Ağırlıklandırılmış Hata Kareleri Ortalaması - AHKO), MTK'da ise kalibrasyon sürecindeki ölçek dönüştürme yöntemleri ve eşitleme yöntemlerinin en az hata değerini belirlemek için RMSE (Root Mean Square Error- Hata Kareleri Ortalamasının Karekökü) katsayıları hesaplanılmıştır. Belirtilen hata katsayıları boostrap kökenli bir sonuç verdiği için Newton-Raphson yöntemine dayalı Delta hata katsayıları da ayrıca raporlaştırılmıştır.

### Bulgular

Araştırma sürecine öncelikle eşitleme varsayımları kontrol edilerek sürece başlanmıştır. Tüm eşitlenmiş puanlarda ham puan uç değerlerine sabitlenerek fark değerleri hesaplanmıştır. Fark değerleri, ham puanlardan bu değere karşılık gelen eşitlenmiş puan çıkartılarak hesaplanmıştır.

KTK'daki Tucker yöntemi öncelikle $w_1 = 1$ sentetik evren ağırlığına göre ele alınmıştır. 0 - 8 puan aralığında eşitlenmiş puanlar, ham puanlardan; 9 - 25 puan aralığında ise ham puanlar eşitlenmiş puanlardan daha düşük değer almıştır. Sentetik evren ağırlığı olarak seçilen 1 değeri için hesaplanan fark puanları -0.571 ile 0.193 arasında değişmektedir. Tucker lineer yönteminde ikinci ağırlık olarak $w_1 = 0.50$'e göre eşitlenmiş puanlar hesaplanmıştır. 0-7 puan aralığında eşitlenmiş puanlar, ham puanlardan; 8 - 25 puan aralığında ise ham puanlar eşitlenmiş puanlardan daha düşük değer almıştır. Tucker lineer eşitlemede $w_1 = 0.5$ sentetik evren ağırlığı için fark puanları -0.4856 ile 0.1929 arasında değer almıştır.

Levine gözlenen eşitleme $w_1 = 1$ sentetik evren ağırlığında 0-10 puan aralığında eşitlenmiş puanlar, ham puanlardan; 11 - 25 puan aralığında ise ham puanlar eşitlenmiş puanlardan daha düşük değer almıştır. Levine $w_1 = 1$ için fark puanları ise en düşük -0.5219 ve en yüksek 0.3465 arasında değerlerini almıştır. Levine gözlenen eşitleme $w_1 = 0.5$ sentetik evren ağırlığında 0 - 10 puan aralığında eşitlenmiş puanlar, ham puanlardan; 11 - 25 puan aralığında ise ham puanlar eşitlenmiş puanlardan daha düşük değer almıştır. Levine $w_1 = 0.5$ için fark puanları ise en düşük -0.543 ile en yüksek 0.354 arasında değerlerini almıştır.

Levine gözlenen puan eşitlemeden sonra gerçek puan eşitleme yöntemi kullanılmış ve bu yöntemde sentetik evren ağırlıkları kullanılmamıştır. Levine gerçek eşitlemede 0 - 11 puan aralığında eşitlenmiş puanlar, ham puanlardan; 12 - 25 puan aralığında ise ham puanlar eşitlenmiş puanlardan daha düşük değer almıştır. Levine gerçek puan eşitlemede fark puanları ise en düşük 2.5013 ve en yüksek 1.8666 değerleri hesaplanmıştır

Klasik konjenerik eşitleme yönteminde 0 - 11 puan aralığında eşitlenmiş puanlar, ham puanlardan; 12-25 puan aralığında ise ham puanlar eşitlenmiş puanlardan daha düşük değer almıştır. Ham puanlar ve eşitlenmiş puanlar arasındaki fark puanları ise -0.4542 ile 0.3667 arasındadır.

Braun-Holland yöntemi kullanıldığında 0-9 puan aralığında eşitlenmiş puanlar ham puanlardan; 10-25 puan aralığında ise ham puanlar eşitlenmiş puanlardan daha düşük değer almıştır. Braun-Holland yöntemindeki fark puanları 1.067 ile -0.283 değerleri arasında değişkenlik göstermiştir. Lineer eşitleme yöntemlerinde hata değeri en az Tucker içsel ($w_1$=1) için en fazla ise Levine gerçek puan eşitleme tarafından üretilmiştir.

Lineer eşitleme yönteminden sonra KTK'da eşityüzdelikli eşitleme yöntemlerine göre eşitlenmiş puanlar elde edilmiştir. Ön düzgünleştirme yapılmadan elde edilen puanlar 0.138 ile 24.088 arasında değişmektedir. Öncelikle düzgünleştirme süreci için ön düzgünleştirme yöntemlerinden log-lineer polinom derecesine karar verilmiştir. Momentlerin uyumu hem analitik hem de grafiksel olarak incelendiğinde C 6 polinominal dereceye göre log-lineer ön düzgünleştirme yöntemi kullanılmıştır. C 6 polinominal derecede log-lineer ve beta4 yöntemlerine göre ön düzgünleştirmede 0-9 puan aralığında eşitlenmiş puanlar ham puanlardan; 10-25 puan aralığında ise ham puanlar eşitlenmiş puanlardan daha düşük değerler alınmıştır. C 6 polinominal derecesine eşitlenmiş puanlar -0.007 ile 25.309 arasında değişirken. beta4 binominal eşitlemede ise -0.1637 ile 25.044 arasında değerlere sahiptir. Son düzgünleştirme için analitik ve grafiksel çözüm incelendiğinde kübik spline S 0.05 derecesi en uygun eşitleme derecesi belirtmektedir. S 0.05 son düzgünleştirme sürecinde 0 - 8 puan aralığında eşitlenmiş puanlar ham puanlardan; 9-25 puan aralığında ise ham puanlar eşitlenmiş puanlardan daha düşük değerler alınmıştır.

Frekans kestirim yöntemi için sentetik evren ağırlığı $w_1 = 0.5$; $w_1 = 0.1$ için 0 - 2 puan aralığında eşitlenmiş puanlar ham puanlardan; 3 - 12 puan aralığında ise ham puanlar eşitlenmiş puanlardan daha düşük değerler alınmıştır. Ham puan 12 için sentetik evrenin farklı ağırlıkları kullanıldığında ise 12'den daha düşük değerler elde edilmiştir.

KTK'na bağlı eşitleme yöntemlerinin eşitleme hataları incelendiğinde en düşük hata değerinden en yüksek hata değerine doğru sıralama şu şekildedir; frekans kestirim yöntemi ($w_1 = 0.5$), Tucker içsel ($w_1 = 1$), Levine gözlenen ($w_1 = 0.5$), Tucker içsel ($w_1 = 0.5$), klasik konjenerik, frekans kestirim ($w_1 = 1$), Braun-Holland, düzgünleştirme yapılmadan EYE ve Levine gerçek puan eşitlemedir.

MTK'ya dayalı eşitleme yapmak için öncelikle aynı ölçek üzerinde puanların yerleştirilmesi gerekmektedir. Bu sebeple kalibrasyon yöntemleri incelenmiş ve en az hata ile Stocking-Lord kalibrasyon yöntemine karar verilmiştir. MTK gerçek ve gözlenen puan eşitleme yöntemlerinde 0 - 15 puan aralığında puan aralığında eşitlenmiş puanlar ham puanlardan; 16 - 25 puan aralığında ise ham puanlar eşitlenmiş puanlardan daha düşük değerler alınmıştır. Hata katsayıları incelendiğinde en düşük eşitleme hatası MTK gerçek puan eşitleme yöntemindedir. Kuramsal olarak eşitleme yöntemleri incelendiğinde MTK gerçek ve gözlenen eşitleme yöntemleri, KTK'daki tüm eşitleme yöntemlerinden daha az hata ile eşitlenmiş puanların elde edildiği görülmektedir

**Sonuç ve Öneriler**

KTK bünyesindeki lineer ve eşityüzdelikli ve MTK gerçek ve gözlenen puan eşitleme yöntemleri kullanılarak PISA 2012 matematik testi kitapçık1 ve kitapçık3 test puanları eşitlenmiştir. Lineer eşitlemede Tucker $w_1$=1, $w_1$=0.5 sentetik evren ağırlıkları, Levine gözlenen $w_1 = 1$, $w_1 = 0.5$ sentetik evren ağırlıkları, Levine gerçek, klasik konjenerik, Braun-Holland yöntemleri kullanılmıştır. Lineer eşitleme yöntemlerinde en az eşitleme hatası sentetik evren ağırlığı $w_1$= 1 Tucker içsel yöntemi ile en fazla hata ise Levine gerçek puan eşitleme yöntemlerinden elde edilmiştir.

Eşityüzdelikli eşitlemede ise düzgünleştirme yapılmadan, ön düzgünleştirme için C 6 polinominal derecesine dayalı log-lineer, beta4, son düzgünleştirme S 0.05 kübik spline derecesi ve $w_1$= 1, $w_1$= 0.5 sentetik evren ağırlıkları kullanılarak frekans kestirim yöntemleri ile eşitlenmiş puanlar hesaplanmıştır. Bu eşitleme yöntemleri arasında frekans kestirim ($w_1$= 0.5) olduğu, en fazla hatanın düzgünleştirme yapılmamış EYE yöntemi olduğu görülmüştür. KTK'ya dayalı lineer ve eşityüzdelikli eşitleme yöntemleri hata değerlerine göre kıyaslandığında en az hata eşityüzdelikli eşitleme yönteminde olduğu görülmüştür.

MTK'da eşitleme yönteminden önce kalibrasyon yöntemine karar verilmiştir. En uygun kalibrayon yöntemi en az hata ile Stocking-Lord ile sağlanmıştır. MTK gerçek ve gözlenen puan eşitleme yöntemleri karşılaştırıldığında en az hata MTK gerçek puan eşitleme yöntemidir. MTK gerçek puan eşitleme yönteminin daha robust bir çözüm ürettiği, daha az hata ile eşitleme yaptığı sonucuna varılmıştır. İncelenen tüm kuramlara dayalı eşitleme yöntemleri en az hataya göre sıralandığında MTK gerçek puan eşitleme, MTK gözlenen puan eşitleme, frekans kestirim ($w_1$= 0.5), frekans kestirim ($w_1$= 1), Tucker ($w_1$= 1), Levine gözlenen ($w_1$= 1), Levine gözlenen ($w_1$=0.50), Tucker ($w_1$= 1), klasik konjenerik, frekans kestirim ($w_1$= 1), Braun-Holland, düzgünleştirme yapılmamış EYE, Levine gerçek puan eşitleme yöntemi şeklindedir. Bu çalışma bulguları doğrultusunda ortak maddeli eşdeğer olmayan gruplar deseninde yapılan eşitleme sürecinde en az hata ile MTK gerçek puan eşitleme yönteminden elde edilmiştir.

Bu araştırmadan elde edilen sonuçlarla birlikte PISA 2012 için farklı kitapçıklar farklı ortak madde oranları gözetilerek incelenebilir. PISA 2012'de yer alan diğer okuryazarlık türleri (fen ve teknoloji ve okuma becerisi) için eşitleme süreçleri raporlaştırılabilir. Çalışma verilerinde çoklu puanlama 0-1 matrisine dönüştürülerek incelenmiştir. Başka bir çalışmada çoklu puanlamaya dayalı olarak eşitleme yöntemleri kıyaslanabilir. Araştırmada ele alınan ortak maddeli eşitleme yöntemleri yerine okuryazarlık puanları ile düşük, orta ve yüksek korelasyon veren kodeğişkenlerine bağlı eşitleme süreci ele alınabilir. PISA 2012 testinde yer alan tek boyutluluk varsayımının ihlalini vurgulayarak gerçek

verilere dayalı simülasyon yapılarak tek boyutluluk ihlalinde eşitleme yöntemleri kıyaslanabilir. Araştırmada çoktan seçmeli test maddeleri veri olarak değerlendirilmiştir. Karma test formatında eşitleme yöntemleri denenebilir. Ortak maddenin testin toplam madde sayısına göre farklı oranlarına göre değiştirilip en uygun ortak madde oranı ve bu süreçte kullanılacak eşitleme yöntemine karar verilebilir.

## Appendixes

**Appendix 1-Distribution of Students Receiving Booklets by Countries**

| | Booklet 1 | Booklet 3 |
|---|---|---|
| China (QCN) | 442 | 439 |
| Indonesia (IDN) | 424 | 417 |
| Finland (FIN) | 683 | 669 |
| Turkey (TUR) | 372 | 375 |
| Total | 1921 | 1900 |
| General Total | 3821 | |

**Appendix 2**

***Comparison of parameters according to t test***

| Parameter | Booklets | n | x | Sx | df | t | p |
|---|---|---|---|---|---|---|---|
| a | Booklet 1 | 13 | 2.198 | 0.583 | 24 | 0.243 | 0.769 |
| | Booklet 3 | 13 | 2.138 | 0.661 | | | |
| b | Booklet 1 | 13 | 0.458 | 0.796 | 24 | 0.312 | 0.468 |
| | Booklet 3 | 13 | 0.369 | 0.644 | | | |
| c | Booklet 1 | 13 | 0.129 | 0.093 | 24 | 0.513 | 0.991 |
| | Booklet 3 | 13 | 0.109 | 0.1 | | | |

***Comparison of mean score according to t test***

| Booklets | n | X | Sx | df | t | p |
|---|---|---|---|---|---|---|
| Booklet 1 | 1921 | 11.31 | 6.075913 | 3819 | 0.554 | 0.58 |
| Booklet 3 | 1900 | 11.43 | 6.351331 | | | |

***Reliability Coefficients of Booklets***

| | Booklet 1 | Booklet 3 |
|---|---|---|
| KR-20 | 0.902 | 0.910 |

Fisher Z Coefficients of Booklets

| | Booklet 1 | Booklet 3 |
|---|---|---|
| Fisher Zr | 1.472 | 1.528 |

| Booklets | Dimension | $\chi^2$ (sd) | RMSEA | GFI | CFI | NNFI | SRMR | λ | ε |
|---|---|---|---|---|---|---|---|---|---|
| Booklet 1 | Unidimension | 11890.546 (275) | 0.05 | 0.991 | 0.969 | 0.952 | 0.0531 | 0.42-0.90 | 0.10-0.60 |
| Booklet 3 | Unidimension | 11821.59 (275) | 0.05 | 0.99 | 0.965 | 0.95 | 0.0517 | 0.43-0.90 | 0.10-0.60 |

**Appendix 3-Items and Codes in Booklets**

| Item type | Booklet | Items | Codes of Items |
|---|---|---|---|
| Non-anchor/common items | Booklet 1 | MATH-P2012- An advertising Column Q1 | PM00GQ01 |
| | | MATH-P2012-Speeding Fines Q1 | PM909Q01 |
| | | MATH-P2012-Speeding Fines Q2 | PM909Q02 |
| | | MATH-P2012-Speeding Fines Q3 | PM909Q03 |
| | | MATH-P2012-Roof Truss Design Q1 | PM949Q01T |
| | | MATH-P2012-Roof Truss Design Q2 | PM949Q02T |
| | | MATH-P2012-Roof Truss Design Q3 | PM949Q03T |
| | | MATH-P2012-Migration Q1 | PM955Q01 |
| | | MATH-P2012-Migration Q2 | PM955Q02 |
| | | MATH-P2012-Migration Q3 | PM955Q03 |
| | | MATH-P2012-Bike Rental Q2 | PM998Q02T |
| | | MATH-P2012-Bike Rental Q4 | PM998Q04T |
| | Booklet 3 | MATH-P2000-Pipelines Q1 | PM273Q01T |
| | | MATH-P2003-Lotteries Q1 | PM408Q01T |
| | | MATH-P2003-Transport Q1 | PM420Q01T |
| | | MATHP2003-TheThermometer Criket Q1 | PM446Q01 |
| | | MATHP2003-TheThermometer Criket Q2 | PM446Q02 |
| | | MATH-P2003-Tile Arrangement Q1 | PM447Q01 |
| | | MATH-P2003-The Fence Q1 | PM464Q01T |
| | | MATH-P2003-Telephone Rates Q1 | PM559Q01 |
| | | MATH-P2003-Computer Game Q1 | PM800Q01 |
| | | MATH-P2003-Carbon Dioxide Q1 | PM828Q01 |
| | | MATH-P2003-Carbon Dioxide Q2 | PM828Q02 |
| | | MATH-P2003-Carbon Dioxide Q3 | PM828Q03 |
| Anchor items | Booklet 1 ve Booklet 3 | MATH-P2012-Apartment Purchase Q1 | PM00FQ01 |
| | | MATH-P2012-Drip Rate Q1 | PM903Q01 |
| | | MATH-P2012-Drip Rate Q3 | PM903Q03 |
| | | MATH-P2012-Charts Q1 | PM918Q01 |
| | | MATH-P2012-Charts Q2 | PM918Q02 |
| | | MATH-P2012-Charts Q5 | PM918Q05 |
| | | MATH-P2012-Sailing Ships Q1 | PM923Q01 |
| | | MATH-P2012-Sailing Ships Q3 | PM923Q03 |
| | | MATH-P2012-Sailing Ships Q4 | PM923Q04 |
| | | MATH-P2012-Sauce Q2 | PM924Q02 |
| | | MATH-P2012-Revolving Door Q1 | PM995Q01 |
| | | MATH-P2012-Revolving Door Q2 | PM995Q02 |
| | | MATH-P2012-Revolving Door Q3 | PM995Q03 |

**Appendix 4- Equating Scores and Difference Values Obtained Using the Tucker Internal Partner Equating Method**

| Tucker-Internal | | | | |
|---|---|---|---|---|
| Raw Scores | Equating Scores for $w_1$=1 | Difference | Equating Scores for $w_1$=0.5 | Difference |
| 0 | 0 | 0 | 0 | 0 |
| 1 | 0.7374 | 0.2626 | 0.8071 | 0.1929 |
| 2 | 1.7737 | 0.2263 | 1.8366 | 0.1634 |
| 3 | 2.8099 | 0.1901 | 2.8661 | 0.1339 |
| 4 | 3.8462 | 0.1538 | 3.8956 | 0.1044 |
| 5 | 4.8824 | 0.1176 | 4.9251 | 0.0749 |
| 6 | 5.9187 | 0.0813 | 5.9546 | 0.0454 |
| 7 | 6.9549 | 0.0451 | 6.9841 | 0.0159 |
| 8 | 7.9912 | 0.0088 | 8.0136 | -0.0136 |
| 9 | 9.0275 | -0.0275 | 9.0431 | -0.0431 |
| 10 | 10.0637 | -0.0637 | 10.0726 | -0.0726 |
| 11 | 11.1 | -0.1000 | 11.1021 | -0.1021 |
| 12 | 12.1362 | -0.1362 | 12.1316 | -0.1316 |
| 13 | 13.1725 | -0.1725 | 13.1611 | -0.1611 |
| 14 | 14.2087 | -0.2087 | 14.1906 | -0.1906 |
| 15 | 15.245 | -0.245 | 15.2201 | -0.2201 |
| 16 | 16.2813 | -0.2813 | 16.2496 | -0.2496 |
| 17 | 17.3175 | -0.3175 | 17.2791 | -0.2791 |
| 18 | 18.3538 | -0.3538 | 18.3086 | -0.3086 |
| 19 | 19.39 | -0.39 | 19.3381 | -0.3381 |
| 20 | 20.4263 | -0.4263 | 20.3676 | -0.3676 |
| 21 | 21.4625 | -0.4625 | 21.3971 | -0.3971 |
| 22 | 22.4988 | -0.4988 | 22.4266 | -0.4266 |
| 23 | 23.5351 | -0.5351 | 23.4561 | -0.4561 |
| 24 | 24.5713 | -0.5713 | 24.4856 | -0.4856 |
| 25 | 25 | 0 | 25 | 0 |

**Appendix 5- Equating Scores and Difference Values Obtained Using the LevineOS Equating Method**

| Levine OS | | | | |
|---|---|---|---|---|
| Raw Sores | Equating Scores for $w_1$=1 | Difference | Equating Scores for $w_1$=0.5 | Difference |
| 0 | 0 | 0 | 0 | 0 |
| 1 | 0.6535 | 0.3465 | 0.6459 | 0.3541 |
| 2 | 1.6913 | 0.3087 | 1.6849 | 0.3151 |
| 3 | 2.729 | 0.271 | 2.7239 | 0.2761 |
| 4 | 3.7668 | 0.2332 | 3.7629 | 0.2371 |
| 5 | 4.8045 | 0.1955 | 4.8019 | 0.1981 |
| 6 | 5.8423 | 0.1577 | 5.841 | 0.159 |
| 7 | 6.8801 | 0.1199 | 6.88 | 0.12 |
| 8 | 7.9178 | 0.0822 | 7.919 | 0.081 |
| 9 | 8.9556 | 0.0444 | 8.958 | 0.042 |
| 10 | 9.9933 | 0.0067 | 9.997 | 0.003 |
| 11 | 11.0311 | -0.0311 | 11.036 | -0.036 |
| 12 | 12.0688 | -0.0688 | 12.075 | -0.075 |
| 13 | 13.1066 | -0.1066 | 13.114 | -0.114 |
| 14 | 14.1444 | -0.1444 | 14.153 | -0.153 |
| 15 | 15.1821 | -0.1821 | 15.192 | -0.192 |
| 16 | 16.2199 | -0.2199 | 16.2311 | -0.2311 |
| 17 | 17.2576 | -0.2576 | 17.2701 | -0.2701 |
| 18 | 18.2954 | -0.2954 | 18.3091 | -0.3091 |
| 19 | 19.3331 | -0.3331 | 19.3481 | -0.3481 |
| 20 | 20.3709 | -0.3709 | 20.3871 | -0.3871 |
| 21 | 21.4087 | -0.4087 | 21.4261 | -0.4261 |
| 22 | 22.4464 | -0.4464 | 22.4651 | -0.4651 |
| 23 | 23.4842 | -0.4842 | 23.5041 | -0.5041 |
| 24 | 24.5219 | -0.5219 | 24.5431 | -0.5431 |
| 25 | 25 | 0 | 25 | 0 |

**Appendix 6- Equating Scores and Difference Values Obtained Using the Levine True Score Equating Method**

| | LevineTS | |
|---|---|---|
| Raw Scores | Equating Scores | Difference |
| 0 | 0 | 0 |
| 1 | 0 | 0 |
| 2 | 0 | 0 |
| 3 | 1.1301 | 1.8699 |
| 4 | 2.3602 | 1.6398 |
| 5 | 3.5902 | 1.4098 |
| 6 | 4.8203 | 1.1797 |
| 7 | 6.0504 | 0.9496 |
| 8 | 7.2804 | 0.7196 |
| 9 | 8.5105 | 0.4895 |
| 10 | 9.7406 | 0.2594 |
| 11 | 10.9706 | 0.0294 |
| 12 | 12.2007 | -0.2007 |
| 13 | 13.4307 | -0.4307 |
| 14 | 14.6608 | -0.6608 |
| 15 | 15.8909 | -0.8909 |
| 16 | 17.1209 | -1.1209 |
| 17 | 18.3510 | -1.3510 |
| 18 | 19.5811 | -1.5811 |
| 19 | 20.8111 | -1.8111 |
| 20 | 22.0412 | -2.0412 |
| 21 | 23.2713 | -2.2713 |
| 22 | 24.5013 | -2.5013 |
| 23 | 25 | 0 |
| 24 | 25 | 0 |
| 25 | 25 | 0 |

**Appendix 7- Equating Scores and Difference Values Obtained Using the Classical Congeneric Equating Method**

| | Classical Congeneric | |
|---|---|---|
| Raw Scores | Equating Scores | Difference |
| 0 | 0 | 0 |
| 1 | 0.6333 | 0.3667 |
| 2 | 1.6690 | 0.3310 |
| 3 | 2.7047 | 0.2953 |
| 4 | 3.7404 | 0.2596 |
| 5 | 4.7761 | 0.2239 |
| 6 | 5.8118 | 0.1882 |
| 7 | 6.8475 | 0.1525 |
| 8 | 7.8832 | 0.1168 |
| 9 | 8.9188 | 0.0812 |
| 10 | 9.9545 | 0.0455 |
| 11 | 10.9902 | 0.0098 |
| 12 | 12.0259 | -0.0259 |
| 13 | 13.0616 | -0.0616 |
| 14 | 14.0973 | -0.0973 |
| 15 | 15.1330 | -0.1330 |
| 16 | 16.1687 | -0.1687 |
| 17 | 17.2044 | -0.2044 |
| 18 | 18.2401 | -0.2401 |
| 19 | 19.2758 | -0.2758 |
| 20 | 20.3115 | -0.3115 |
| 21 | 21.3472 | -0.3472 |
| 22 | 22.3829 | -0.3829 |

**Appendix 8- Equating Scores Obtained from the Braun-Holland Method**

| | Braun-Holland | |
|---|---|---|
| Raw Scores | Equating Scores | Difference |
| 0 | 0 | 0 |
| 1 | 0 | 0 |
| 2 | 0.9335 | 1.0665 |
| 3 | 2.0835 | 0.9165 |
| 4 | 3.2334 | 0.7666 |
| 5 | 4.3833 | 0.6167 |
| 6 | 5.5333 | 0.4667 |
| 7 | 6.6832 | 0.3168 |
| 8 | 7.8331 | 0.1669 |
| 9 | 8.9831 | 0.0169 |
| 10 | 10.133 | -0.133 |
| 11 | 11.2829 | -0.2829 |
| 12 | 12 | 0 |

**Appendix 9- Values Obtained from EE Method Based on Pre-Smoothing**

| | | EE | | |
|---|---|---|---|---|
| X form score | Standard Error | Unsmoothed | Log-Lineer (C=6) | Beta4 |
| 0 | 0.1630 | 0.1384 | -0.0681 | -0.1637 |
| 1 | 0.2137 | 1.1492 | 0.8742 | 0.7833 |
| 2 | 0.1742 | 2.0807 | 1.8140 | 1.7339 |
| 3 | 0.1655 | 2.9008 | 2.7460 | 2.6994 |
| 4 | 0.1831 | 3.7078 | 3.6794 | 3.6815 |
| 5 | 0.2035 | 4.5315 | 4.6317 | 4.6969 |
| 6 | 0.2384 | 5.4438 | 5.6234 | 5.7431 |
| 7 | 0.2916 | 6.6477 | 6.6708 | 6.8132 |
| 8 | 0.3081 | 7.8182 | 7.7791 | 7.8985 |
| 9 | 0.3316 | 9.0535 | 8.9379 | 8.9916 |
| 10 | 0.3292 | 10.1093 | 10.1243 | 10.0871 |
| 11 | 0.3379 | 11.1660 | 11.3084 | 11.1810 |
| 12 | 0.3525 | 12.3693 | 12.4626 | 12.2700 |
| 13 | 0.4455 | 13.5228 | 13.5677 | 13.3511 |
| 14 | 0.3409 | 14.8450 | 14.6131 | 14.4215 |
| 15 | 0.2978 | 15.7167 | 15.6000 | 15.4783 |
| 16 | 0.2881 | 16.4249 | 16.5391 | 16.5187 |
| 17 | 0.3112 | 17.3366 | 17.4492 | 17.5392 |
| 18 | 0.2956 | 18.1770 | 18.3534 | 18.5353 |
| 19 | 0.3496 | 19.4024 | 19.2760 | 19.5034 |
| 20 | 0.2425 | 20.2149 | 20.2384 | 20.4440 |
| 21 | 0.2452 | 21.1492 | 21.2519 | 21.3607 |
| 22 | 0.2197 | 22.1928 | 22.3106 | 22.2648 |
| 23 | 0.2737 | 23.5643 | 23.3847 | 23.1723 |
| 24 | 0.1414 | 23.7691 | 24.4196 | 24.0965 |
| 25 | 0.3099 | 24.0878 | 25.3089 | 25.0442 |

**Appendix 10-Raw Score Conversions for Post-Smoothing**

| Raw Scores | Standard Error | Unsmooted | S=0.01 | S=0.05 | S=0.10 | S=0.20 | S=0.30 | S=0.40 | S=0.50 | S=0.75 | S=1.00 | LE |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 0.163 | -0.143 | -0.065 | -0.052 | -0.045 | -0.054 | -0.068 | -0.081 | -0.091 | -0.092 | -0.091 | -0.401 |
| 1 | 0.2137 | 0.787 | 0.806 | 0.845 | 0.866 | 0.837 | 0.795 | 0.757 | 0.726 | 0.724 | 0.727 | 0.644 |
| 2 | 0.1742 | 1.887 | 1.874 | 1.847 | 1.823 | 1.787 | 1.759 | 1.735 | 1.715 | 1.726 | 1.728 | 1.689 |
| 3 | 0.1655 | 2.825 | 2.813 | 2.787 | 2.760 | 2.741 | 2.738 | 2.736 | 2.736 | 2.763 | 2.765 | 2.735 |
| 4 | 0.1831 | 3.663 | 3.667 | 3.679 | 3.684 | 3.702 | 3.723 | 3.742 | 3.76 | 3.801 | 3.802 | 3.780 |
| 5 | 0.2035 | 4.574 | 4.569 | 4.596 | 4.630 | 4.684 | 4.725 | 4.76 | 4.791 | 4.838 | 4.839 | 4.825 |
| 6 | 0.2384 | 5.571 | 5.571 | 5.592 | 5.631 | 5.701 | 5.752 | 5.795 | 5.831 | 5.876 | 5.875 | 5.870 |
| 7 | 0.2916 | 6.657 | 6.670 | 6.678 | 6.698 | 6.760 | 6.807 | 6.848 | 6.883 | 6.914 | 6.912 | 6.916 |
| 8 | 0.3081 | 7.847 | 7.848 | 7.828 | 7.817 | 7.855 | 7.888 | 7.917 | 7.943 | 7.951 | 7.949 | 7.961 |
| 9 | 0.3316 | 9.044 | 9.023 | 8.983 | 8.964 | 8.974 | 8.986 | 8.998 | 9.01 | 8.989 | 8.986 | 9.006 |
| 10 | 0.3292 | 10.118 | 10.116 | 10.112 | 10.117 | 10.104 | 10.091 | 10.083 | 10.079 | 10.026 | 10.023 | 10.052 |
| 11 | 0.3379 | 11.181 | 11.198 | 11.245 | 11.270 | 11.231 | 11.194 | 11.168 | 11.147 | 11.064 | 11.059 | 11.097 |
| 12 | 0.3525 | 12.368 | 12.364 | 12.413 | 12.417 | 12.344 | 12.287 | 12.245 | 12.21 | 12.101 | 12.096 | 12.142 |
| 13 | 0.4455 | 13.523 | 13.601 | 13.586 | 13.536 | 13.430 | 13.362 | 13.309 | 13.265 | 13.138 | 13.133 | 13.188 |
| 14 | 0.3409 | 14.845 | 14.778 | 14.682 | 14.595 | 14.480 | 14.411 | 14.358 | 14.309 | 14.175 | 14.17 | 14.233 |
| 15 | 0.2978 | 15.717 | 15.692 | 15.642 | 15.578 | 15.490 | 15.434 | 15.388 | 15.343 | 15.212 | 15.206 | 15.278 |
| 16 | 0.2881 | 16.435 | 16.485 | 16.517 | 16.507 | 16.468 | 16.435 | 16.402 | 16.366 | 16.248 | 16.243 | 16.324 |
| 17 | 0.3112 | 17.341 | 17.340 | 17.393 | 17.420 | 17.430 | 17.422 | 17.406 | 17.381 | 17.284 | 17.28 | 17.369 |
| 18 | 0.2956 | 18.251 | 18.289 | 18.311 | 18.345 | 18.391 | 18.405 | 18.404 | 18.391 | 18.32 | 18.317 | 18.414 |
| 19 | 0.3496 | 19.377 | 19.308 | 19.271 | 19.297 | 19.360 | 19.39 | 19.401 | 19.399 | 19.356 | 19.354 | 19.460 |
| 20 | 0.2425 | 20.263 | 20.262 | 20.249 | 20.278 | 20.346 | 20.383 | 20.401 | 20.407 | 20.392 | 20.39 | 20.505 |
| 21 | 0.2452 | 21.204 | 21.224 | 21.268 | 21.299 | 21.354 | 21.387 | 21.406 | 21.416 | 21.427 | 21.427 | 21.550 |
| 22 | 0.2197 | 22.312 | 22.360 | 22.364 | 22.365 | 22.386 | 22.404 | 22.417 | 22.427 | 22.463 | 22.464 | 22.596 |
| 23 | 0.2737 | 23.564 | 23.539 | 23.485 | 23.455 | 23.433 | 23.429 | 23.432 | 23.44 | 23.498 | 23.501 | 23.641 |
| 24 | 0.1414 | 24.349 | 24.403 | 24.409 | 24.404 | 24.383 | 24.372 | 24.371 | 24.379 | 24.44 | 24.444 | 24.686 |
| 25 | 0.3099 | 25.088 | 25.134 | 25.136 | 25.135 | 25.128 | 25.124 | 25.124 | 25.126 | 25.147 | 25.148 | 25.732 |

**Appendix 11-Frequency Estimation Method Results According to Different Weight Values of the Synthetic Population**

| Synthetic Weights | Raw Scores | Equating Scores | Difference | Synthetic Weights | Equating Scores | Difference |
|---|---|---|---|---|---|---|
| $w_1$= 0.5; $w_2$=0.5 | 0 | 0 | 0 | $w_1$=1; $w_2$=0 | 0 | 0 |
| | 1 | 0 | 0 | | 0 | 0 |
| | 2 | 0 | 0 | | 0 | 0 |
| | 3 | 3.9136 | -0.9136 | | 3.9172 | -0.9172 |
| | 4 | 4.9173 | -0.9173 | | 4.9215 | -0.9215 |
| | 5 | 5.9513 | -0.9513 | | 5.9593 | -0.9593 |
| | 6 | 6.0923 | -0.0923 | | 6.0947 | -0.0947 |
| | 7 | 7.2091 | -0.2091 | | 7.2133 | -0.2133 |
| | 8 | 8.1629 | -0.1629 | | 8.1648 | -0.1648 |
| | 9 | 9.0895 | -0.0895 | | 9.0912 | -0.0912 |
| | 10 | 10.1874 | -0.1874 | | 10.1892 | -0.1892 |
| | 11 | 11.4419 | -0.4419 | | 11.4434 | -0.4434 |
| | 12 | 11.6069 | 0.3931 | | 11.6065 | 0.3935 |

**Appendix 12- Scores Obtained from IRT TS and IRT OS Equating Methods**

| X Scores | IRT θ | | | IRT OS | |
|---|---|---|---|---|---|
| | | IRT TS | Difference | | Difference |
| 0 | ….. | 0 | 0 | 0.0489 | -0.0489 |
| 1 | ….. | 0.9477 | 0.0523 | 0.9824 | 0.0176 |
| 2 | ….. | 1.8955 | 0.1045 | 1.8965 | 0.1035 |
| 3 | ….. | 2.8462 | 0.1538 | 2.78 | 0.22 |
| 4 | -1.3124 | 3.7859 | 0.2141 | 3.6408 | 0.3592 |
| 5 | -0.9112 | 4.4421 | 0.5579 | 4.4699 | 0.5301 |
| 6 | -0.647 | 5.205 | 0.795 | 5.3319 | 0.6681 |
| 7 | -0.439 | 6.1223 | 0.8777 | 6.2561 | 0.7439 |
| 8 | -0.2673 | 7.1676 | 0.8324 | 7.2557 | 0.7443 |
| 9 | -0.1206 | 8.2952 | 0.7048 | 8.3046 | 0.6954 |
| 10 | 0.0103 | 9.4658 | 0.5342 | 9.362 | 0.638 |
| 11 | 0.131 | 10.645 | 0.355 | 10.4504 | 0.5496 |
| 12 | 0.2444 | 11.8017 | 0.1983 | 11.6826 | 0.3174 |
| 13 | 0.3516 | 12.9166 | 0.0834 | 12.8915 | 0.1085 |
| 14 | 0.4535 | 13.9877 | 0.0123 | 13.9777 | 0.0223 |
| 15 | 0.5513 | 15.0248 | -0.0248 | 15.0174 | -0.0174 |
| 16 | 0.6467 | 16.0439 | -0.0439 | 16.0452 | -0.0452 |
| 17 | 0.7418 | 17.0623 | -0.0623 | 17.067 | -0.067 |
| 18 | 0.8393 | 18.0946 | -0.0946 | 18.0992 | -0.0992 |
| 19 | 0.9431 | 19.1495 | -0.1495 | 19.1623 | -0.1623 |
| 20 | 1.0595 | 20.2319 | -0.2319 | 20.231 | -0.231 |
| 21 | 1.1992 | 21.3487 | -0.3487 | 21.311 | -0.311 |
| 22 | 1.3847 | 22.5154 | -0.5154 | 22.5026 | -0.5026 |
| 23 | 1.6807 | 23.6584 | -0.6584 | 23.6854 | -0.6854 |
| 24 | 2.1527 | 24.5061 | -0.5061 | 24.4691 | -0.4691 |
| 25 | ….. | 25 | 0 | 25 | 0 |

## Yayın Etiği Beyanı

Bu araştırmanın planlanmasından, uygulanmasına, verilerin toplanmasından verilerin analizine kadar olan tüm süreçte "Yükseköğretim Kurumları Bilimsel Araştırma ve Yayın Etiği Yönergesi" kapsamında uyulması belirtilen tüm kurallara uyulmuştur. Yönergenin ikinci bölümü olan "Bilimsel Araştırma ve Yayın Etiğine Aykırı Eylemler" başlığı altında belirtilen eylemlerden hiçbiri gerçekleştirilmemiştir. Bu araştırmanın yazım sürecinde bilimsel, etik ve alıntı kurallarına uyulmuş; toplanan veriler üzerinde herhangi bir tahrifat yapılmamıştır. Bu çalışma herhangi başka bir akademik yayın ortamına değerlendirme için gönderilmemiştir.

## Araştırmacıların Katkı Oranı Beyanı

Birinci yazar Ceren Mutluer %70, ikinci yazar Prof. Dr. Mehtap Çakan %30 oranında katkı sağlamıştır.

## Çatışma Beyanı

Araştırmanın yazarları arasında herhangi bir çıkar çatışması bulunmamaktadır. Ayrıca yazarlar, diğer kişi, kurum ya da kuruluşlarla herhangi bir çıkar çatışması içinde olmadıklarını beyan ederler.