Equating TIMSS Mathematics Subtests with Nonlinear Equating Methods Using NEAT Design: Circle-Arc Equating Approaches

Burhanettin Ozdemir [i]
Siirt University

Abstract
The purpose of this study is to equate Trends in International Mathematics and Science Study (TIMSS) mathematics subtest scores obtained from TIMSS 2011 to scores obtained from TIMSS 2007 form with different nonlinear observed score equating methods under Non-Equivalent Anchor Test (NEAT) design where common items are used to link two or more test forms. The ultimate goal is to determine whether different forms of mathematics tests that administered in different years with anchor (common) items caused any inequalities with respect to students. In addition, results obtained from chained and frequency estimation based on equipercentile equating methods were compared to four different methods (Tucker, Levine, Braun-Holland and chained) based on a new nonlinear equating approach called circle-arc equating in order to see which method is the most appropriate for equating these forms. The results of different nonlinear equating methods were compared with respect to Root Mean Squared Error (RMSE) index, mean of bootstrap standard errors (MBSE) and mean of bootstrap bias. Results indicates that TIMSS 2007 mathematics tests were easier than TIMSS 2011 mathematics across the score scale which indicates that results were biased against to students participated to TIMSS 2007. Moreover, equating methods based on nonlinear circle-arc equating outperformed the equipercentile equating methods and presmoothing decreased both standard error and bias associated with each method.

Keywords: TIMSS mathematics subtest, Nonlinear Equating NEAT designs, Circle-arc equating approaches.

[i] **Burhanettin Ozdemir** is an Assistant Professor at Siirt University, Siirt, Turkey. He is also head of Educational Measurement and Evaluation Department. His specific research interests include adaptive testing (unidimensional and multidimensional computerized adaptive testing, multistage adaptive testing), item response theory models, test equating and R programing.

Correspondence: b.ozdemir025@gmail.com

Introduction

Large scale international assessments, such as the Trends in International Mathematics and Science Study (TIMSS) and Program for International Student Assessment (PISA), are used as an important assessment tool not only for making important policy decisions (Arim & Ercikan , 2014), but also to determine effectiveness of the present school curricula, students' achievement and effectiveness of the education system of participating countries (Keser, 2005; Uzun, Butuner & Yigit, 2010, Ozdemir, 2014). Another important aspect of these large scale assessments is that they are repeated in a certain period of time which enables policy makers and stakeholders in education to evaluate the educational developments and improvement within these period of time. In order to maintain comparability of these large scale assessments, equivalence of these forms administered in different years has to be satisfied. Otherwise, results obtained from different forms of these test administered in different years might cause biased inferences and decision made upon these assessment might not be valid and reliable (Eryaman & Schneider, 2017).

The Trends in International Mathematics and Science Study (TIMSS) is a well-known large scale assessment which aims to examine students' academic achievement based on some given variables in every 4-year-period. It not only aims to evaluate educational achievement of 4^{th} grade and 8^{th} grade students with respect to mathematics and science, but also gather comprehensive information from students, teachers and school principals about the teaching and learning of mathematics and science.

TIMSS is administered in approximately 50 countries and to thousands of students in each participating country. Limited version of TIMSS was first administered in 1995. In total, 59 countries participated in TIMSS 2007 and 57 countries with 4th grade students and 56 countries with 8th grade students participated in TIMSS 2011. TIMSS applied in different years basically consist of two sections that are mathematics and science, respectively. The content of mathematics consist of fraction, measurement, data representation, analysis and probability, proportionality, geometry and algebra. On the other hand, the content of science part consists of life science, physical science, earth science, biology, chemistry, and physics. Turkey only participated in TIMSS 1999 and 2007 at 8th grade level, while participated in TIMSS 2011 at both 4^{th} and 8^{th} grade levels (Erkan, 2013). Since these tests are administered in every 4-year circle, it is important to examine and check the statistical equivalence of test. Because violence of statistical equivalence of tests administered in different years may lead to biased inferences.

Ercıkan (2014) stated that the item level and test level comparability were related to item and construct bias where item bias was caused by translation/adaptation effects, differential familiarity with item context and content (Ercikan, 1998; Ercikan & McCreith, 2002; Hambleton et al., 2005), while construct bias was caused by factors such as conceptual inequivalence of the construct, inconsistency in theoretical definitions or the measurement of the construct across cultures (Ercikan & Lyons-Thomas, 2013; Geisinger, 1994; Hambleton, 1993, 1994; 2005; Hui & Triandis, 1985; Oliveri, Olson, Ercikan, & Zumbo, 2012; Reise, Widaman, & Pugh, 1993; Sireci, Bastari, & Allalouf, 1998; van de Vijver & Tanzer, 1997). Another source of bias was associated with methods in which differences in test administrations, differential familiarity of examinees with item and test formats and administering tests that measure same construct in different year might cause bias inferences with respect to students and stakeholders in education. At that point, test equating comes in handy and is used to determine and reduce methodological biases and inequalities.

Statistical procedures commonly used to determine the statistical equivalence of different test forms that aim to measure same traits or abilities are called test *equating* and *linking*. Especially, when the test forms are administered across more than one occasion or more than one examinee group, then security of test, statistical equivalence of test and overexposure of test items become major concern of test developers and policy makers. Although alternative test forms are used to

prevent item exposure, multiple test forms measuring the same construct but administered in different years may differ in difficulty levels. Therefore, test equating procedure is used to adjust for differences in difficulty levels of test form administered in the same year or different years in order to produce score scales that can be used interchangeably (Albano, 2014).

Equating methods aim to define statistical relationship between different test score distribution and score scales associated with different test forms that are constructed based on the same specification and similar statistical features. Equating functions related to each equating methods which defines these relationships convert scores from one scale directly to their equivalent values on another so that equated scores can be used interchangeably (Hambleton & Swaminathan, 1985; Holland & Dorans, 2006; Kolen & Brennan, 2004). On the other hand, linking is used to define statistical relationship between different test forms that are not constructed based on same specifications and equated scores can be considered as similar but cannot be used interchangeably. This study is restricted to nonlinear equating methods under non-equivalent anchor test (NEAT) design and test linking is the beyond the scope of this study(for more details about linking, see Holland & Dorans, 2006).

Equating Designs

Equating design has to be specified after determining the forms that will be equated. Equating design determines how test forms and individuals sampled; and how data was collected (Kolen & Brennan, 2004; Mao, von Davier & Rupp, 2006; Gok &Kelecioğlu, 2014; Albano, 2014). Determining proper equating design is also considered to be the most important step of equating (Holland & Dorans, 2006). There are three commonly used equating designs called "single group design", "random group design" and "non-equivalent anchor test (NEAT) design" (Crocker & Algina, 1986; Kolen & Brennan, 2004).

In the single-group design, there is only one sampled group that takes the both forms (X and Y) which will be equated. Since group ability remains unchanged, any observed differences are attributed to test forms. In the random group design, groups are drawn from the same population and each group takes different forms. Since samples are drawn randomly from the population, group ability is assumed to remain same and any observed differences in score distributions are attributed to test forms themselves.

For both the single group design and the random group design, group abilities are assumed to be equal and constant. However, when the groups are not equivalent with respect to ability level, then the groups might be drawn from different populations (such as P and Q) and ability differences become a confounding factor. In order to solve these two problems arising from nonequivalent groups, anchor tests (V) that include common items are used for both groups. Ability differences appeared in groups are assumed to be removed or controlled by means of common items (Kolen & Brennan, 2004; Albano, 2014). In this study, NEAT design was used to equate TIMSS mathematics subtests administered in different years.

Equating types and methods

Generally, equating methods are classified as Classical Test Theory-based (CTT-based) equating methods and Item Response Theory-based (IRT-based) equating methods. These methods differ in required assumptions and mathematical functions being used to define relationship between score distributions. The most common equating methods based on CTT are called "linear equating", "mean equating" and "equipercentile equating" methods (Barnard, 1996; Kelecioğlu, 2014). Moreover, some researcher classifies the equating methods as *standard equating methods*, which can be found in Kolen and Brennan (2004), and von Davier et al. (2004), and *nonstandard equating methods* some of which are new methods while the others are extension of standard methods. For instance, von Davier (2011b) refers to hybrid methods such as local equating (van der Linden 2011) and the Levine nonlinear method (von Davier, Fournier-Zajac & Holland 2007) as nonstandard equating methods (Gonzales, 2014). Apart from that, some researcher defines the CTT-based

equating methods as equating types and classify them into two groups called *linear equating* that express scores on one scale with straight line and *nonlinear equating* that express scores on axis with curvilinear line (Albano, 2014).

Linear equating consist of *identity*, *mean* and *linear equating* and differ from one another in terms of intercept and slope. On the other hand, nonlinear equating consists of *equipercentile* and a new approach called *circle-arc equating* (Livingston & Kim, 2009). Nonlinear equating methods differs from one another in terms of the number of coordinates being estimated. In this study, mathematics subtest administered in different years were equated with nonlinear equating methods under NEAT design. A brief information about equating types (equipercentile and circle-arc equating) and equating methods based on these equating type is provided in the following section.

Equipercentile equating

Equipercentile equating constructs a nonlinear relationship between score scales of tests forms (X and Y) that are to be equated. It is more appropriate to use equipercentile equating methods when difficulty levels of forms differ and difference in difficulties fluctuates across the score scale. Assuming that scores on form X are equated to scale scores on form Y and let F(x) and G(y) be the associated cumulative distribution functions (CDF) of the scores. When CDF functions are set equal (F(x) = G(y)) and solved for y, then the formula for the equipercentile equating function is produced:

$$equipy(x) = G^{-1}[F(x)] \qquad (1)$$

The cumulative distribution function is approximated using percentile ranks, when the score scales are discrete in which this particular procedure is called continuizing the discrete score distributions (for details, see Kolen and Brennan 2004, ch. 2).

For nonlinear equating methods under NEAT design, a lot of techniques and functions have been developed for estimating the relationship between total scores on X and Y and the respective anchor scores on V. These techniques are all based on certain assumptions about the relationships between total scores and anchor scores for the populations that groups are sampled (P and Q). These techniques are referred to here as equating methods based on nonlinear equating under NEAT design. There are two common equating methods for equipercentile equating under NEAT design that are called *frequency estimation* and *chained equating* methods, respectively.

Frequency estimation involves a synthetic population taking forms X and Y are required. In frequency estimation method, conditional distribution of total scores on X for a given score point in V and the conditional distribution of total scores on Y for a given score point in V is the same across populations. On the other hand, chained equating method (Livingston, Dorans, and Wright 1990) can be applied to both linear and equipercentile equating under NEAT design. What differs chained equating from other methods is that it does not require synthetic populations but only requires an additional equating functions for estimating equivalent scores (for more details, see Livingston, Dorans, and Wright 1990; Holland and Dorans, 2006; Albano, 2014).

Circle-arc equating

As like equipercentile equating, circle-arc equating also defines a nonlinear relationship between score scales. Although the main idea behind circle-arc equating was first proposed by Divgi (1987), Livingston and Kim (2010) were those who proposed it as a nonlinear equating method and put into an equating framework. They suggest that when forms differ in difficulty, the relationship between test forms appear to be curvilinear. The main idea behind this method is first to define two end points and a middle point estimated from data, then constrain the estimated equating curve to pass through these points. The maximum and minimum possible scores on the test forms are set as the end points, while the middle point are determined by the mean scores of the test to be equated (Livingston & Kim, 2010). Therefore, the circle-arc equating function derives from the mathematical formulas restraining an arc of a circle to pass through these three pre-specified points.

Many different ways can be used to find the radius (r) of the circle and midpoint with the three known points. The formula for the radius of circle is as follows:

$$r = \sqrt{(x_c - x_1)^2 + \sqrt{(y_c - y_1)^2}} \quad (2)$$

Solving equation (2) for *y* produces the circle-arc equating function:

$$circy(x) = y_c \pm \sqrt{r^2 - (x - x_c)^2}, \quad (3)$$

where (x_c, y_c) is equal to estimated midpoints with different methods and r is the radius of the circle. Equating methods based on circle-arc equating under NEAT design apply only to estimation of this midpoint (x_c, y_c). Commonly used equating methods under NEAT design to estimate midpoints are *chained equating method* demonstrated by Livingston and Kim (2009), *Tucker, Levine* and *Braun-Holland* equating methods (Albano 2014). In this study, circle-arc equating was defined as equating type and *chained equating, Tucker, Levine* and *Braun-Holland* equating methods were used to estimate midpoints. Thus, four different equating methods based on circle-arc equating and two different equating methods based on equipercentile equating were used to equate TIMSS 2011 mathematics subtest to TIMSS 2007 mathematics subtest under NEAT design in order to determine best equating methods.

Generally, smoothing methods are used to reduce or remove irregularities caused by sampling error in the score distribution. Smoothing methods are also used to reduce irregularities caused by equipercentile equating function. There are two commonly used methods including polynomial loglinear presmoothing (Holland and Thayer, 2000) and cubic-spline postsmoothing (Kolen, 1984). In this study, loglinear presmoothing method, with polynomial degree equal to 3 (C=3), was used in order to see how presmoothing effect each nonlinear equating methods and distribution of equated scores.

Purpose of Study

The purpose of this study is to equate Trends in International Mathematics and Science Study (TIMSS) mathematics test scores obtained from TIMSS 2011 to scores obtained from TIMSS 2007 form with nonlinear observed score equating methods under Non-Equivalent Anchor Test (NEAT) design. The ultimate goal is to determine whether different forms of mathematics tests administered in different years with anchor (common) items caused any inequalities with respect to students. In addition, results obtained from different equating methods based on equipercentile equating methods were compared to a new nonlinear equating method called *circle-arc equating methods* so as to determine the most appropriate equating method.

Research questions

The main goal of this study is to equate mathematics tests administered in different years with nonlinear equating methods under NEAT design so as to determine whether administering different form cause any in equalities with respect to students. Therefore, research questions were as follows:

I. What is the relationship between observed scores and equivalent scores obtained from different equating methods based on equipercentile equating and circle-arc equating?
II. Do TIMSS mathematics test forms administered in different years cause any inequalities with respect to students?
III. How does presmoothing affect the distribution of observed scores and equated scores?
IV. How do nonlinear equating methods differ in terms of standard error, RMSE and bias values?
V. How does presmoothing affect the distribution of standard error, RMSE and bias values across the score scale?
VI. What is the best nonlinear equating methods to equate TIMSS mathematics subtests under NEAT design?

Methodology

Research Model

The model of this study was casual comparative research; since it aimed to investigate the statistical relationship and differences between two mathematics tests which assumed to measure same construct and administered in different years by different nonlinear equating methods under NEAT design.

Study groups

In this study, data from TIMSS 2011 and TIMSS 2007 mathematics subtests administered to eighth grade Turkish students were used. The ultimate goal of TIMSS survey is to determine student achievement in mathematics and science in the participating countries. TIMSS 2007 mathematics test was comprised of 14 booklets with 115 items and 4498 eight-grade Turkish student participated in total. On the other hand TIMSS 2011 mathematics test was comprised of 14 booklets with 119 items and 6928 Turkish students participated.

Raw scores obtained from TIMSS 2011 mathematics subtest booklet 12 were equated to TIMSS 2007 mathematics subtest booklet 14 with two different nonlinear equating methods named as equipercentile equating and circle-arc methods, respectively, under NEAT design. TIMSS 2007 mathematics test booklet 14 was administered to 323 students and TIMSS 2011 mathematics tests booklet 12 was administered to 495 students. Both booklets contain 23 dichotomous items and 7 out of 23 items are anchor (common) items. It is assumed that each booklet used for equating represent the other booklets administered in the same year, since booklet administered in the same year contain anchor items and the total scores are equated before revealing the results.

Data Analysis

For equipercentile equating, two different equating methods named as "frequency estimation" and "chained equating" were used, while for circle-arc method, four different equating methods named as "Tucker", "Levine", "Braun-Holland" and "chained equating" were used, respectively. In addition, observed scores obtained from both forms were presmoothed before equating in order to examine effect of presmoothing on different equating methods. The results of different nonlinear equating methods were compared with respect to Root Mean Squared Error (RMSE) index, mean of bootstrap standard errors (MBSE) and bias. The R package called "equate" (Albano 2014) was used to conduct equating analysis.

Assumptions of Observed Score Equating

Test forms that are being equated based on Classical Test Theory (CTT) have to satisfy three main assumptions of equating. These assumptions are unidimensionality, equal reliability, and equivalent difficulty levels (Angoff, 1984; Dorans & Holland, 2000; Kolen and Whitney, 1982). Therefore before conducting the analysis, assumptions of unidimensionality, equal reliability and equivalent difficulty levels were checked.

A test is assumed to be unidimensional only when there is just one dominant factor or ability being measured by items (Hambleton et al., 1991). In order to check unidimensionality of each form, factor analysis was conducted. Table 1 presents the results of factor analyses related to each TIMSS mathematics subtest.

Table 1. *Factor analysis results associated with TIMSS mathematics subtests*

	TIMSS 2007 Booklet 14			**TIMSS 2011 Booklet 12**		
Factors	Eigen Values	Explained Variance	Cumulative Variance (%)	Eigen Values	Explained Variance	Cumulative Variance (%)
1	7,301	**31,744**	31,8	7,113	**30,926**	30,926
2	1,392	6,053	37,797	1,226	5,329	36,255
3	1,222	5,313	43,111	1,096	4,767	41,022

Table 1 presents Eigen values, explained variance and cumulative variance related to first three factors of each form based on factor analysis results. Results indicate that the first factor (dominant factor) of TIMSS 2007 mathematics subtest explained 31.74 percent of total variance, while the first factor of TIMSS 2011 mathematics subtest explained 30,93 percentage of the total variance. Büyüköztürk (2007) suggests that when the first factor of a test explains 30% (or more) of total variance, then this test is considered to be unidimensional. According to results in Table 1, both forms measured same construct and were unidimensional and therefore, the unidimensionality assumption of equating was satisfied.

Table 2 presents the results associated with the other two assumption of equating named as equal reliability and equivalent difficulties, respectively.

Table 2. *Descriptive statistics associated with each form and assumptions*

	Mean difficulties	Reliability (α)	Variance	Standard deviation	Mean	Skewness	Kurtosis
TIMSS 2011 Booklet 12	0.391	0.892	33,693	5,804	8,96	0,745	-0,425
TIMSS 2007 Booklet 14	0.389	0.891	33,441	5,782	9,00	0,721	-0,514

Table 2 shows that Cronbach alpha reliability (α) of TIMSS 2011 booklet 14 and TIMSS 2007 booklet 12 were 0.891 and 0.892, respectively. Z statistics based on difference between two reliability coefficients was not statistically significant ($z=-0.072$; $p>0.05$). On the other hand, mean difficulty of TIMSS 2011 booklet 14 was 0.389 and mean difficulty of TIMSS 2007 booklet 12 was 0.391. Z statistics based on difference between two ratio was not statistically significant ($z=-0.092$; $p>0.05$). Therefore, reliability coefficients and mean difficulties of two test forms were assumed to be almost identical. As a result, the three main assumptions of equating based on CTT were satisfied.

Descriptive statistics shown in Table 2 also indicates that both test forms were positively skewed with negative kurtosis statistics. In addition, mean, variance, reliability and mean difficulties of each form were compared by using Z test procedure. As a result, difference between two forms related to mean, variance, reliability and mean difficulties were not statistically significant ($p=0,05$) which indicates that each forms were almost identical.

Findings

In this study, raw scores obtained from TIMSS 2011 mathematics subtest booklet 12 were equated to TIMSS 2007 mathematics subtest booklet 14 under NEAT design with two different nonlinear equating methods named as equipercentile equating and circle-arc methods, respectively. In addition, raw scores were presmoothed with log-linear presmoothing method in order to examine how smoothing effects distribution of raw scores and the standard error, RMSE and bias related to each non-linear equating method.

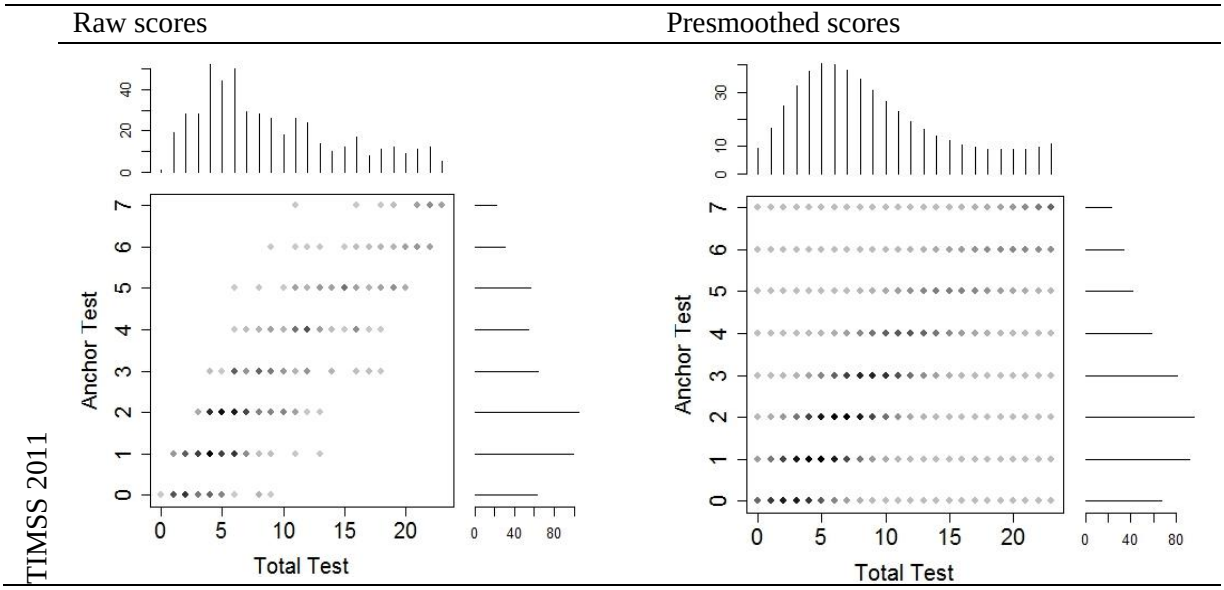
Raw scores of TIMSS 2011 mathematics subtest and equivalent scores of TIMSS 2011 equated to TIMSS 2007 mathematics subtest scores with different equipercentile and circle-arc equating methods *without presmoothing* under NEAT design were presented in Appendix A. On the other hand, equivalent scores obtained from different equipercentile and circle-arc equating methods with *presmoothing* under NEAT design were presented in Appendix B.

The results of equipercentile equating showed that equivalent scores of TIMSS 2011 mathematics subtest ranged between -0.38 and 23.08. However, when raw scores were presmoothed, equivalent scores ranged between 0 and 23.08 which implies that equipercentile equating with presmoothing yielded more accurate result. On the other hand, the results of different circle-arc

equating methods indicate that equivalent scores of TIMSS 2011 mathematics subtest ranged between 0 and 23 for both raw and presmoothed scores. This finding indicates that equivalent score interval remained same with circle-arc equating methods. Therefore, one can conclude that circle-arc equating method yielded more accurate results compare to equipercentile equating method.

According to results in Appendix A and B, all raw scores of TIMSS 2011 mathematics subtest were smaller than TIMSS 2007 mathematics subtest equivalent scores based on circle-arc equating results which indicates that there was a linear relationship between mathematics subtest raw scores and TIMSS 2011 mathematics subtest equivalent scores. Therefore, it can be concluded that TIMSS 2007 mathematics subtest was easier than TIMSS 2011 mathematics subtest along the score scale. This equating result is an indicator of inequality with respect to student attended in 2011 TIMSS mathematics subtest caused by administration of different forms in different years.

On the other hand, there was nonlinear relationship between raw scores and equivalent scores obtained from equipercentile equating methods without presmoothing. However, when the raw scores were presmoothed, all raw scores of TIMSS 2011 mathematics subtest appeared to be smaller than equivalent scores based on equipercentile equating results indicating that there was a linear relationship between raw scores and TIMSS 2011 mathematics subtest equivalent scores. When compared to results of equating without presmoothing, equipercentile equating methods with presmoothing yielded more accurate and consistent results. As like equating results obtained from circle-arc equating, equipercentile equating results also indicate that different mathematics subtest applied in different years caused inequalities with respect to students.



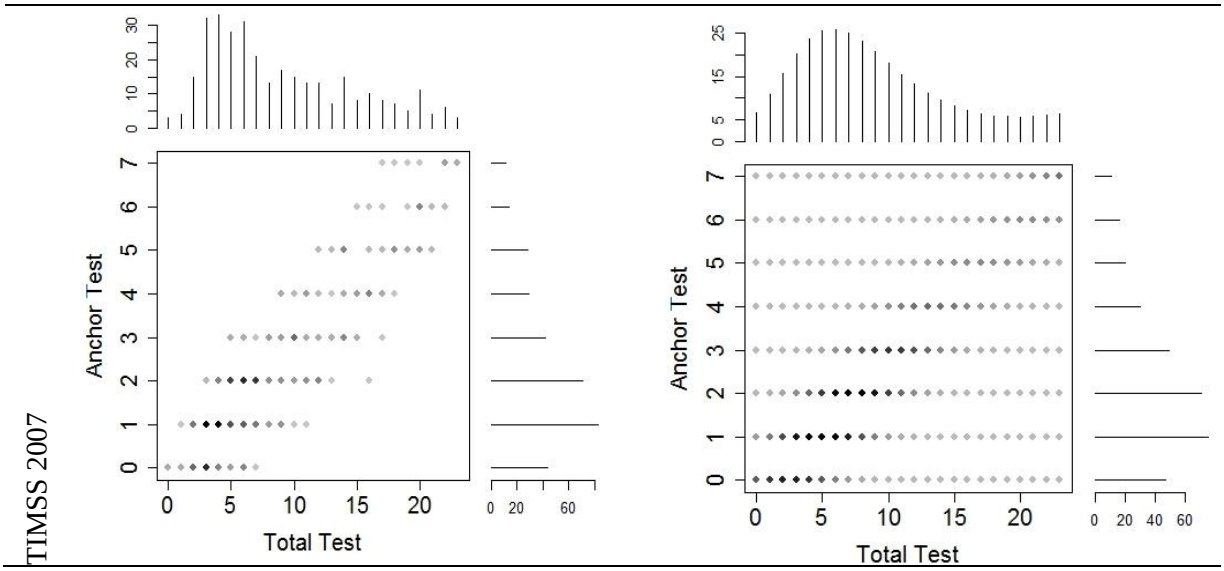


Figure 1. *Distribution of raw and presmoothed scores related to TIMSS mathematics subtests*

Figure 1 depicts frequency distribution of raw scores without presmoothing on the left hand side and distribution of presmoothed scores related to each test forms on the right hand side. In addition, Figure 1 shows the frequency distribution of total test scores on the horizontal line and frequency distribution of anchor test scores on the vertical line. With the help of graphs in the Figure 1, one can easily observe the effect of presmoothing on the raw scores for each test forms with anchor items.

Figure 2 depicts distribution of mean equated scores versus total observed scores. Red lines in the Figure 2 represent the equated scores related to equipercentile equating methods, while green and yellow lines represent the equated scores related to different circle-arc equating methods. In addition, one can observe the effect of presmoothing on the relation between mean equated scores and total scores for each non-linear equating methods by examining the graph on the right hand side.

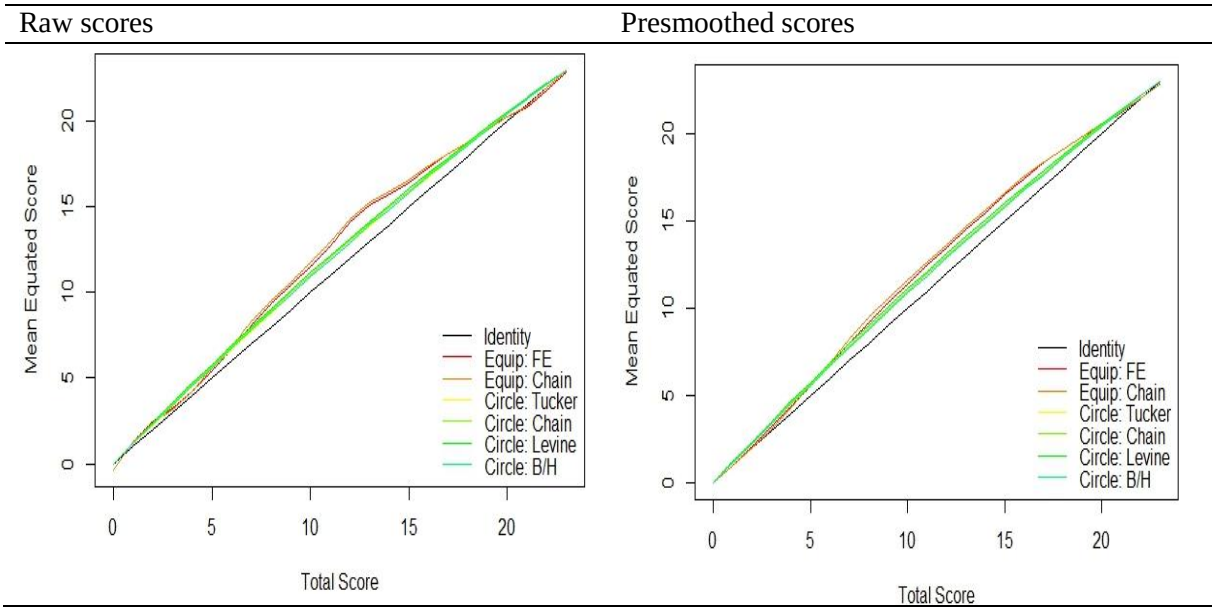


Figure 2 *Distribution of Mean Equated scores versus Total scores*

Figure 2 shows that relationship between mean equated scores and observed scores was almost identical for equipercentile equating methods. Likewise, circle-arc equating methods yielded

similar results. However, compared to the relationship between mean equated scores and observed scores associated with equipercentile equating methods, circle-arc equating methods yielded more consistent results. Without presmoothing the relationship between mean equated scores and observed scores associated with equipercentile equating became more divergent around the mean. On the other hand, both circle-arc and equipercentile equating methods yielded similar results when observed scores were presmoothed. Thus, it can be assumed that presmoothing increased accuracy of observed score equating for each nonlinear equating method.

Table 3. *Equating results associated with nonlinear equating methods without presmoothing*

Equating type	Equating method	MBSE	bias	w.bias	RMSE
Equipercentile Equating	Frequency E.	0.790	0.906	0.038	1.309
	Chained	0.853	0.984	0.041	1.414
Circle-Arc Equating	Tucker	0.237	0.598	0.027	0.643
	Chained	0.240	0.669	0.030	0.711
	Levine	0.263	0.744	0.034	0.789
	Braun-Holland	0.241	0.600	0.027	0.646

Table 3 presents RMSE, MBSE and bias statistics associated with different equipercentile equating and circle-arc methods without presmoothing.

The results given in Table 3 indicate that when the results of different equipercentile equating methods were compared, frequency estimation equipercentile equating method without presmoothing yielded smaller MBSE (0.790), RMSE (1,309) and bias (0.906) statistics and outperformed the chained equipercentile method. On the other hand, when the results of different circle-arc equating methods were compared, Tucker circle-arc equating method yielded smallest BMSE (.237), RMSE (0.643) and bias (0.598) statistics. Although, Tucker circle-arc equating method outperformed other equating methods, Braun-Holland and chained circle-arc methods yielded almost similar results. Table 3 also shows that, regardless of methods being used, the circle-arc equating methods yielded relatively small equating errors and bias statistics compared to equipercentile equating methods. Therefore, it can be concluded that circle-arc equating methods without presmoothing outperformed the equipercentile equating methods.

Table 4 presents RMSE, MBSE and bias statistics associated with different equipercentile equating and circle-arc methods with presmoothing based on bootsrap method.

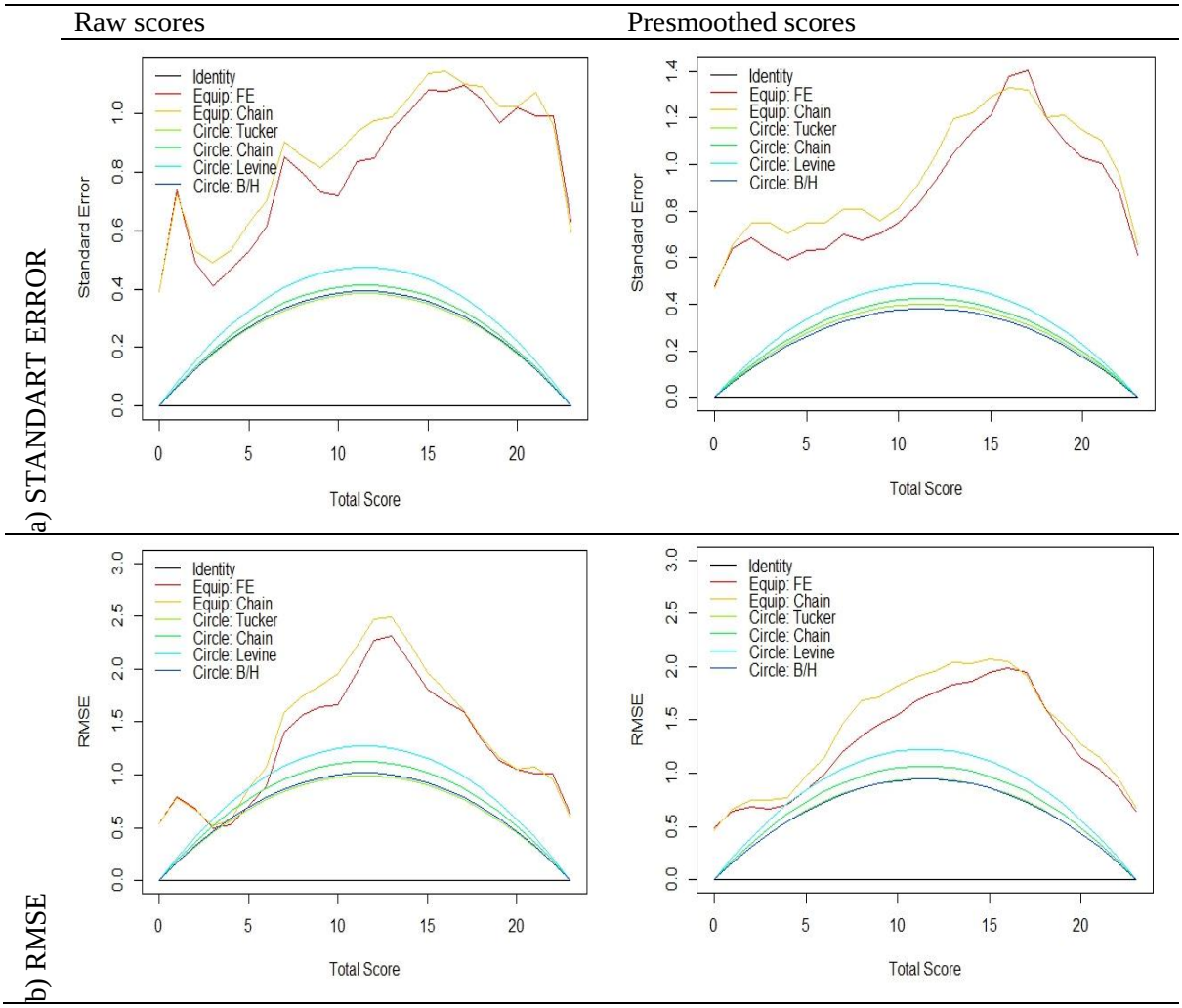
Table 4. *Equating results associated with nonlinear equating methods with presmoothing*

Equating type	Equating method	MBSE	bias	w.bias	RMSE
Equipercentile Equating	Frequency estimation	0.905	0.897	0.036	1.324
	Chained	0.986	0.944	0.039	1.437
Circle-Arc Equating	Tucker	0.233	**0.561**	**0.026**	**0.607**
	Chained	0.251	0.631	0.029	0.679
	Levine	0.286	0.705	0.032	0.761
	Braun-Holland	**0.231**	0.575	**0.026**	0.619

When different methods based on equipercentile equating were compared, chained equipercentile equating method with presmoothing yielded smaller MBSE (0.905), RMSE (1.324) and bias (0.897) statistics and outperformed the frequency estimation equipercentile method. On the other hand, Braun-Holland circle-arc equating method yielded smallest BMSE (.231), while Tucker circle-arc method yielded smallest RMSE (0.607) and bias (.0561) values respectively. In addition, presmoothing decreased MBSE, RMSE and bias values substantially for circle-arc equating methods which indicates that presmoothing decreased both random and systematic errors associated with each equating method. On the other hand, presmoothing increased MBSE and RMSE while decreased bias values for equipercentile equating methods which indicates that presmoothing increased random error and decreased systematic errors associated with each equipercentile equating method.

In general, circle-arc equating methods outperformed equipercentile equating methods and presmoothing decreased MBSE, RMSE and bias values substantially for equating methods based on circle-arc equating. Therefore, Tucker and Braun-Holland circle-arc equating methods with presmoothing can be considered as the most appropriate method for equating TIMSS mathematics subtest administered in different years.

Figure 3 presents the distribution of standard error (SE), RMSE and bias statistics along with the score scale for each equipercentile and circle-arc equating methods. In addition, distribution of SE, RMSE and bias statistics along with the presmoothed score scale were given on the right hand side so as to see effect of presmoothing. Yellow and red lines in each graph represent findings related to equipercentile equating methods, while green and blue lines represent findings related to circle-arc equating methods.



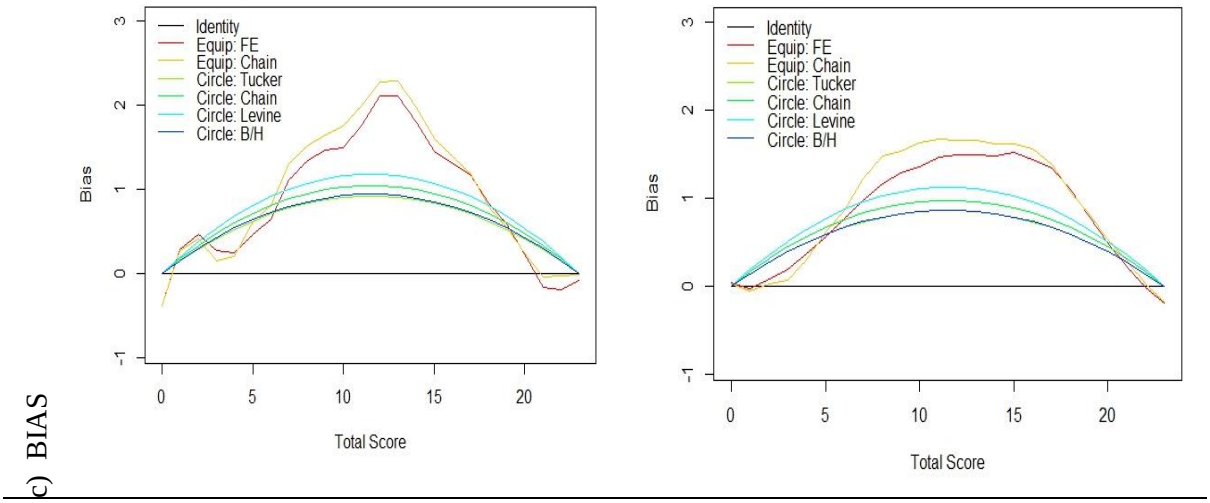


Figure 3. *Distribution of SE, RMSE and Bias values associated with each equating method*

Graphs related to distribution of standard errors in Figure 3 indicates that standard errors of circle-arc equating methods were relatively smaller than standard errors of equipercentile equating methods along with the score scale. In addition, standard errors of circle-arc equating methods had curvilinear distribution, while standard errors of equipercentile equating methods did not have a certain distribution pattern. Moreover, highest and lowest observed scores had the smallest standard errors, whereas observed scores around the mean had the highest standard errors for circle-arc equating methods. For equipercentile equating methods, as the observed score increased so did the standard errors. On the other hand, when the observed scores were presmoothed, standard errors associated with both circle-arc and equipercentile equating methods tended to decrease somewhat.

Graphs related to distribution of RMSE in Figure 3 indicates that RMSE of circle-arc equating methods were relatively smaller than RMSE of equipercentile equating methods along with the score scale in general. However, both circle-arc and equipercentile equating methods yielded similar RMSE values when observed scores were equal (or close) to 5. As like standard errors, RMSE values of circle-arc equating methods had curvilinear distribution, while RMSE of equipercentile equating methods did not have a certain distribution pattern. Moreover, highest and lowest observed scores had the smallest RMSE, whereas observed scores around the mean had the highest RMSE values for both equipercentile and circle-arc equating methods. On the other hand, when the observed scores were presmoothed, RMSE associated with both circle-arc and equipercentile equating methods tended to decrease somewhat.

Graphs related to distribution of equating bias associated with each equating methods in Figure 3 indicates that equating bias of circle-arc equating methods were relatively smaller than equating bias of equipercentile equating methods along with the score scale, in general. However, when observed scores were equal or smaller than 5 and greater than 20, equipercentile equating methods yielded smaller bias values compared to circle-arc equating methods. As like standard errors and RMSE distribution, equating bias of circle-arc equating methods had curvilinear distribution, while bias of equipercentile equating methods did not have a certain distribution. Moreover, highest and lowest observed scores had the smallest bias, whereas observed scores around the mean had the largest bias values for both equating methods based on equipercentile and circle-arc equating. On the other hand, bias values associated with both circle-arc and equipercentile equating methods tended to decrease somewhat when the observed scores were presmoothed.

Conclusion and discussion

Large scale tests, such as TIMSS, require administering different forms each period of time since traits being measured remains same. However, educational testing services must take into

account the psychometric and practical issues that may cause inequalities and biased measurement. This study aimed to check statistical equivalence of TIMSS mathematics subtest test administered different years and to determine the most appropriate nonlinear equating method.

Equating results indicated that TIMSS 2007 mathematics subtest was easier than TIMSS 2011 mathematics subtest across the score scale and there was a nonlinear relationship between raw scores and equivalent scores of TIMSS 2011 mathematics subtest which indicates that results were biased against to students participated in TIMSS 2007. This is an indicator of methodological bias (Sireci Patsula & Hambleton, 2005) caused by using different test forms which aims to measure same construct and affects the comparability of test scores (Arim & Erçıkan, 2014). These biased score might also affect the educational inferences and decisions made upon these scores. Therefore, the scores obtained from TIMSS mathematics tests administered in different years cannot be used interchangeable.

Kan (2011) examined OKS tests administered in different years so as to determine whether scores obtained from these two forms could be treated as interchangeable and whether these two form caused any advantage or disadvantage on examinees performance. He equated the raw scores on 2005 OKS form to 2003 OKS form with linear equating under single group design. Similar to results obtained from this study, Kan (2011) also found that scores obtained from two different OKS tests administered in different years could not be used interchangeably.

When it comes to comparison of nonlinear equating methods, results indicates that the new circle-arc methods outperformed the equipercentile methods and yielded more consistent results with smaller MBSE, RMSE and bias values. In addition, when observed scores were presmoothed, MBSE, RMSE and bias values associated with circle-arc methods were decreased substantially. Unlike circle-arc equating, standard error and RMSE values associated with equipercentile equating tended to increase, while bias values associated with equipercentile equating tended to decrease in the case of presmoothing. This result suggests that presmoothing tended to increase random errors and decrease systematic errors with respect to equipercentile equating methods. Hanson, Zeng and Colton (1994) stated that both presmoothing and postsmoothing methods could improve estimation of the equipercentile equating function. However, Parshall and his colleague (1995) found that decreases in sample size caused substantial increases in standard errors indicating that equipercentile equating methods were negatively affected from decrease in sample size. Livingston and Kim (2009) suggested that one could prefer circle-arc equating to equipercentile equating when the samples were too small for equipercentile equating.

Equivalent observed scores obtained from circle-arc equating methods had least standard error, RMSE and bias statistics regardless of methods being used. Some other studies (Butler and Hanson, 1997; Zhu, 1998) yielded parallel result. Hanson, Zeng and Colton (1994); and Kelecioglu and Ozturk (2013) found that presmoothing and postsmoothing improved estimated equipercentile function and reduced equating error in random group design.

As a result, nonlinear Tucker and Braun-Holland circle-arc equating methods with presmoothing were considered to be the most appropriate equating methods for TIMSS dataset administered in different years, since they yielded the least standard random error, bias and RMSE coefficients. Demir and Güler (2014) examined the statistical equivalence of different PISA 2009 science tests administered at the same time with different equating methods under NEAT design. In this study, PISA 2009 science tests were equated with Tucker linear equating, Levine linear equating, and frequency prediction and Braun-Holland linear equating methods. They found that among these linear equating methods, Braun-Holland linear equating method was the most appropriate for PISA 2009 science tests which supports the results of present study.

Results of this study indicate that different methods based on circle-arc equating outperformed the other nonlinear observed score equating methods based on equipercentile equating.

. In this study real data set obtained from TIMSS mathematics subtests were used. However more research should be conducted in order to see how different conditions, such as different sample sizes, equating designs and postsmoothing affect the circle-arc equating methods. Moreover, it is suggested to compare the performance of circle-arc equating methods with other linear observed and true score equating methods.

References

Albano, A. D., (2014). equate: Observed-Score Linking and Equating. R package version 2.0-3, URL http://CRAN.R-project.org/package=equate.

Angolf, W. H. (1984). *Scales, norms, and equivalent scores*. Princeton, NJ: ETS.

Büyüköztürk, Ş. (2007), *Sosyal Bilimler için Veri Analizi El Kitabı*, Ankara: Pegem A Yayıncılık.

Crocker, L., & Algina, J. (1986) *Introduction to Classical and Modern Test Theory*, Harcourt Brace Jovanovich College Publishers: Philadelphia

Divgi, D. R. (1987). *A stable curvilinear alternative to linear equating* (Report CRC 571). Alexandria, VA: Center for Naval Analyses.

Demir, S., & Güler, N. (2014). Ortak maddeli denk olmayan gruplar desenine ilişkin test eşitleme çalışması. *International Journal of Human Sciences, 11*(2), 190-208. doi: 10.14687/ijhs.v11i2.2870

Dorans, J. N. & Holland, P. W. (2000). Population in variance and the equitability of tests: Basic theory and the linear case. *Journal of Educational Measurement,* 37, 281-306.

Erkan, SS. S. (2013). A comparison of the educational systems in Turkey and Singapore and 1999-2011 TIMSS test results. *Procedia - Social and Behavioral Sciences* 106 (2013) 55 – 64

Ercikan, K. (1998). Translation Effects in International Assessments. *International Journal of Educational Research, 29,* 543-553.

Ercikan, K. (2002). Disentangling Sources of Differential Item Functioning in Multilanguage Assessments. *International Journal of Testing,* 4, 199-215.

Ercikan, K., & Lyons-Thomas, J. (2013). Adapting tests for use in other languages and cultures. In K. Geisinger (Ed.), *APA handbook of testing and assessment in psychology, Volume 3,* (pp. 545-569). American Psychological Association: Washington, DC.

Ercikan, K., & McCreith, T. (2002). Effects of Adaptations on Comparability of Test Items and Test Scores. In D. Robitaille & A. Beaton (Eds.) *Secondary analysis of the TIMSS results: A synthesis of current research* (pp. 391-407). Dordrecht, the Netherlands, Kluwer Academic Publishers.

Eryaman, M. Y. & Schneider, B. (2017). *Evidence and Public Good in Educational Policy, Research and Practice.* New York: Springer.

Geisinger, K. F. (1994). Cross-cultural normative assessment: Translation and adaptation issues influencing the normative interpretation of assessment instruments. *Psychological Assessment, 6,* 304-312.

Gonzalez J (2014). SNSequate: Standard and Nonstandard Statistical Models and Methods for Test Equating. R package version 1.1-1, URL http://CRAN.R-project.org/package=SNSequate.

Gök, B. & Kelecioğlu, H., (2014). Comparison of IRT Equating Methods Using the Common-Item Nonequivalent Groups Design. *Mersin University Journal of the Faculty of Education* Vol. 10, Issue 1, April 2014, pp. 120-136

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage.

Hambleton, R. K. (1993). Translating achievement tests for use in cross-cultural studies. *European Journal of Psychological Assessment, 9,* 57-68.

Hambleton, R. K. (1994). Guidelines for adapting educational and psychological tests: A progress report. *European Journal of Psychological Assessment*, 10, 229-244.

Hambleton, R. K. (2003). Advances in translating and adapting educational and psychological tests. *Language Testing, 20,* 127-134.

Hambleton, R. K. (2005). *Issues, Designs, and Technical Gmdelines for Adapting Tests into Multiple Languages and Cultures*. In R.K. Hambleton, P. Merenda, & C. Spielberger (Eds.). *Adapting educational and psychological tests for cross-cultural assessment* (pp. 93-115). Hillsdale, NJ: Lawrence Erlbaum.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston: Kluwer.

Holland PW, Dorans NJ (2006). "Linking and Equating." In RL Brennan (ed.), *Educational Measurement*, 4 edition, pp. 187{220. Westport, CT: Greenwood.

Hui, C. H., & Triandis, H. C. (1985). Individualism-collectivism: A study of cross-cultural researchers. *Journal of Cross-Cultural Psychology,* 17, 225-248.

Kan, A. (2011). Test Equating: Checking Statistical Equivalance of OKS Test Edition. *Education and* Science. 36(160), 38-51.

Kelecioğlu, H.,& Öztürk Gübeş, N. (2013). Random Grup Deseni ile Yapılan Doğrusal Eşitleme ve Eşit Yüzdelikli Eşitleme Yöntemlerinin Karşılaştırılması. *International Online Journal of Educational Sciences*, 5(1), 227-241.

Keser, Ö. F. (2005). Recommendations towards developing educational standards to improve science education in Turkey. *The Turkish Online Journal of Educational Technology* – TOJET, 4(1), Article 6

Kolen MJ, Brennan RL (2004). *Test Equating, Scaling, and Linking*. New York, NY: Springer-Verlag.

Kolen, M. J., Whitney, D. R. (1982). Comprison of four procedures for equating the tests general educational development. *Journal of Educational Measurement*, 19(4), 279–293.

Livingston SA, Dorans NJ,Wright NK (1990).What Combination of Sampling and Equating Methods Works Best? *Applied Measurement in Education*, 3, 73-95.

Livingston SA, Kim S (2009). "The Circle-Arc Method for Equating in Small Samples." *Journal of Educational Measurement,* 46, 330{343.

Livingston SA, Kim S (2010). Random-groups Equating With Samples of 50 to 400 Test Takers. *Journal of Educational Measuremen*t, 47, 175{185.

Oliveri, M., Olson, B., Ercikan, K., & Zumbo, B. (2012). Methodologies for investigating item- and test- level construct comparability in international large-scale assessments. *International Journal of Testing, 12,* 203-223.

Ozdemir, B. (2014). A comparison of IRT-based methods for examining differential item functioning in TIMSS 2011 mathematics subtest. *Procedia - Social and Behavioral Sciences* 174 (2015) 2075 – 2083

Parshall, C.G., Houghton, P. D. B., & Kromrey, J. D. (1995). Equating error and statistical bias in small sample linear equating. *Journal of Educational Measurement*, *32*, 37–54.

Reise, S. P. Widaman, K. F., & Pugh, R. H. (1993). Confirmatory factor analysis and item response theory: Two approaches for exploring measurement invariance. *Psychological Bulletin, 114,* 552566.

Sireci, S. G., Bastari, B., & Alallouf, A. (1998). *Evaluating construct equivalence across adapted tests.* Invited paper presented at the meeting of the American Psychological Association, San Francisco.

Sireci, S.G., Patsula, L., & Hambleton, R. K. (2005). Statistical methods for identifying flawed items in the test adaptations process. In R. K. Hambleton, P. Merenda, & C. Spielberger (Eds.). *Adapting educational and psychological tests for cross-cultural assessment* (pp. 93-115). Hillsdale, NJ: Lawrence Erlbaum.

Uzun, S. Bütüner, Ö. S ve Yiğit, N. (2010). A Comparison of the Results of TIMSS 1999-2007: The Most Successful Five Countries-Turkey Sample. *Elementary Education Online*, 9 (3), 1174–1188.

van der Linden W (2011). "Local Observed-Score Equating." In A von Davier (ed.), statistical Models for Test Equating, Scaling, and Linking, pp. 201{223. Springer-Verlag.

van de Vijver, F., & Tanzer, N. K. (1997). Bias and equivalence in cross-cultural assessment. *European Review of Applied Psychology*, *47*, 263-279

von Davier AA, Holland PW, Thayer DT (2004). *The Kernel Method of Test Equating*. Springer-Verlag.

von Davier AA (2011b). *Statistical Models for Test Equating, Scaling, and Linking*. Springer-Verlag.

von Davier AA, Fournier-Zajac S, Holland PW (2007). \An Equipercentile Version of the Levine Linear observed Score Equating Function Using the Methods of Kernel Equating." *Research Report RR*-87-31, Educational Testing Service, Princeton NJ.

Appendix A. Equated scores associated with different nonlinear equating methods under NEAT design without presmoothing

	Equipercentile Equating		Circle-Arc Equating			
Raw Scores	Frequency	Chained	Tucker	Chain	levine	Braun Holland
0	**-0.38**	**-0.38**	**0.00**	**0.00**	**0.00**	**0.00**
1	1.54	1.52	1.15	1.06	1.19	1.15
2	2.61	2.55	2.29	2.10	2.37	2.29
3	3.27	3.17	3.42	3.15	3.52	3.42
4	4.22	4.19	4.53	4.19	4.66	4.53
5	5.44	5.56	5.63	5.22	5.79	5.63
6	6.55	6.72	6.71	6.25	6.89	6.71
7	8.15	8.30	7.78	7.28	7.98	7.78
8	9.38	9.48	8.84	8.30	9.04	8.84
9	10.55	10.69	9.88	9.31	10.10	9.88
10	11.62	11.80	10.91	10.32	11.13	10.91
11	12.78	13.17	11.92	11.33	12.15	11.92
12	14.11	14.24	12.92	12.33	13.15	12.92
13	15.11	15.29	13.91	13.32	14.13	13.91
14	15.84	15.95	14.88	14.31	15.10	14.88
15	16.42	16.50	15.84	15.30	16.04	15.84
16	17.35	17.41	16.78	16.28	16.98	16.78
17	18.24	18.29	17.71	17.25	17.89	17.71
18	19.07	19.15	18.63	18.22	18.79	18.63
19	19.81	19.82	19.53	19.19	19.66	19.53
20	20.26	20.25	20.42	20.15	20.52	20.42
21	21.02	20.94	21.29	21.10	21.37	21.29
22	22.10	22.08	22.15	22.06	22.19	22.15
23	**23.08**	**23.08**	**23.00**	**23.00**	**23.00**	**23.00**

Appendix B. Equated scores of different nonlinear equating methods under NEAT design with presmoothing

	Equipercentile Equating		Circle-Arc Equating			
Raw Scores	Frequency	Chained	Tucker	Chained	Levine	Braun-Holland
0	**0.01**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**
1	1.07	1.03	1.15	1.06	1.19	1.15
2	2.18	2.09	2.29	2.10	2.37	2.29
3	3.31	3.17	3.42	3.15	3.52	3.42
4	4.46	4.43	4.53	4.19	4.66	4.53
5	5.63	5.69	5.63	5.22	5.79	5.63
6	6.80	6.91	6.71	6.25	6.89	6.71
7	7.98	8.12	7.78	7.28	7.98	7.78
8	9.14	9.37	8.84	8.30	9.04	8.84
9	10.29	10.46	9.88	9.31	10.10	9.88
10	11.42	11.60	10.91	10.32	11.13	10.91
11	12.53	12.73	11.92	11.33	12.15	11.92
12	13.62	13.75	12.92	12.33	13.15	12.92
13	14.67	14.78	13.91	13.32	14.13	13.91
14	15.68	15.77	14.88	14.31	15.10	14.88
15	16.63	16.69	15.84	15.30	16.04	15.84
16	17.54	17.58	16.78	16.28	16.98	16.78
17	18.39	18.42	17.71	17.25	17.89	17.71
18	19.20	19.22	18.63	18.22	18.79	18.63
19	19.98	19.99	19.53	19.19	19.66	19.53
20	20.73	20.73	20.42	20.15	20.52	20.42
21	21.49	21.48	21.29	21.10	21.37	21.29
22	22.26	22.26	22.15	22.06	22.19	22.15
23	**23.08**	**23.08**	**23.00**	**23.00**	**23.00**	**23.00**