Comparing Performance of Different Equating Methods in Presence and Absence of DIF Items in Anchor Test

Neşe Gübeş [i]
Mehmet Akif Ersoy University

Şeyma Uyar [ii]
Mehmet Akif Ersoy University

Abstract

This study aims to compare the performance of different small sample equating methods in the presence and absence of differential item functioning (DIF) in common items. In this research, Tucker linear equating, Levine linear equating, unsmoothed and presmoothed (C=4) chained equipercentile equating, and simplified circle arc equating methods were considered. The data used in this study is 8th-grade mathematics test item responses which obtained from Trends in International Mathematics and Science Study (TIMSS) 2015 Turkey sample. Item responses from Booklet-1 (N=199) and Booklet-14 (N=224) are chosen for this study. Data analyses were completed in four steps. In the first step, assumptions for DIF detection and test equating methods were checked. In the second step, DIF analyses were conducted with Mantel Haenszel and logistic regression methods. In the third step, Booklet 1 was chosen as base form and Booklet 14 chosen as a new form, then test equating was conducted under common item nonequivalent groups design. Test equating was done in two phases: the presence and absence of DIF items in the common items. Equating results were evaluated based on standard error of equating (se), bias and RMSE indexes. DIF analyses showed that there were two sizeable DIF items in anchor test. Equating results showed that performances of equating methods are similar in presence and absence of DIF items from anchor test and there is no notable change in se, bias and RMSE values. While the circle arc equating method outperformed other equating methods based on se, 4-moment presmoothed chained equipercentile equating method outperformed other methods based on bias and RMSE evaluation criteria.

Keywords: Test Equating, Small Samples, Differential Item Functioning

DOI: 10.29329/ijpe.2020.248.8

---

[i] **Neşe Gübeş,** Assist. Prof., Education Faculty, Department of Educational Sciences, Mehmet Akif Ersoy University, ORCID: 0000-0003-0179-1986

Correspondence: nozturk@mehmetakif.edu.tr

[ii] **Şeyma Uyar,** Assist. Prof. Dr., Faculty of Education, Educational Sciences Department, Mehmet Akif Ersoy University

INTRODUCTION

In national and international testing programs, multiple forms of a single test are used to provide test security or to allow sampling a large of items without having each student answer all of the items. Alternative test forms which developed considering the same construct and blueprint almost differ somewhat in their difficulty. If one form is more difficult than other form, examinees would be expected to get higher scores from the easier form and get lower scores from the more difficult form. Test equating is required to remove effects on scores of these undesirable differences in test form difficulty (Dorans, Moses, & Eignor, 2010). As Kolen and Brennan (1995, p. 2) defined "equating is a statistical process that is used to adjust scores on test forms so that scores on the forms can be used interchangeably."

In testing programs like Trends in International Mathematics and Science Study (TIMSS), Programme for International Student Assessment (PISA) common-item nonequivalent groups (CINEG) design is used to equate test scores. In CINEG design, common items from different test forms are used to equate test forms. Like other statistical analysis methods, common item test equating is exposing to sampling error. One way to reduce sampling error is to conduct equating with large samples of examinees (Kurtz & Dwyer, 2013). Random equating error is directly related to sample size. The sample sizes required to conduct test equating accurately vary based on equating designs and equating methods. For example, "a random-groups design requires a much larger sample than a common-item design, which requires a larger sample than a single-group design" (Kim & Livingston, 2010, p. 286). Kolen and Brennan (2004) suggest that the minimum sample size for linear equating should be 400 and the minimum sample size for equipercentile equating should be 1500. However, large samples always may not be accessible in real test situations. Hence, a variety of methods has been recommended to cope with equating problem in small samples. These methods can be listed as identity, linear, chained equipercentile equating with log-linear presmoothing, circle arc, and synthetic equating (Babcock, Albano & Raymond, 2012). In this study, chained equipercentile equating, linear and circle arc methods are considered so information about these methods is given below.

Chained Equipercentile Equating

Chained equipercentile equating method is an alternative equipercentile equating method. Firstly, this method was described by Angoff (1971) and then Dorans (1990) named this method as chained equipercentile equating. In this method, Form X scores are equated to common items scores. Then scores of common items are equated to the Form Y scores. Assume that Form A is an anchor test for Form X and Form Y. Population P takes the Form X and Population Q takes the Form Y. The scores of Form X are equated to scores of anchor test A using examinees from Population P. Then anchor test A scores are equated to Form Y scores using Population Q. Because of including a chain of two equipercentile equating, it is called as chained equipercentile equating (Kolen & Brennan, 2004).

As Livingston (1993) reported smoothing in equipercentile equating is decreasing sample size requirements by about one half. Presmoothing and postsmoothing are two types of smoothing method which used in equipercentile equating. While the score distributions are smoothed in presmoothing, the equipercentile equivalents are smoothed in postsmoothing. Presmoothing can be done with a polynomial log-linear model or a strong true score model. In this study, we consider presmoothing with the polynomial log-linear model. For the polynomial log-linear presmoothing method, choosing the degree of the polynomial (C) is important because it limits how much smoothing is done. The C parameter is generally chosen from numbers from 1 to 10. After presmoothing, the fitted distribution has the moment preservation property. This means that first C moments of the fitted distribution are the same to sample distribution's first C moments. For instance, if C=2, the mean and standard deviation of the fitted distribution are the same to the mean and standard deviation of the observed distribution. Likelihood ratio chi-square goodness of fit statistic can be used for choosing the C parameter. For instance, the difference between chi-square statistics for C=3 and C=4 can be examined with one degree of freedom. A significant difference between chi-square values means that the more

complex model (C=4) fits the sample data more than the more simple model (C=3). If the two models fit the data adequately the simplest model should be chosen (Kolen & Brennan, 2004).

Linear Equating Methods

Linear equating assumes "apart from differences in means and standard deviations, the distributions of the scores on Form and Form Y are the same" (Crocker & Algina, 1986, p.458). Tucker and Levine are most prevalent (Kolen & Brennan, 2004) linear equating method in CINEG design and their use in small samples are supported by prior researches for small sample sizes (Parshall, Du Bose Houghtan, & Kromrey, 1995; Skaggs, 2005). In this study, we consider Tucker and Levine linear equating methods. Tucker equating was described by (Gulliksen, 1950) and he attributed it to Ledyard Tucker (as cited in Kolen & Brennan, 2004). Tucker equating method makes two important assumptions: regression slopes of the total test scores on the common item score for both populations are equal and variance on the common item score between both examinee populations are equal (Kurtz & Dwyer, 2013). Levine observed score equating is another equating method which used with CINEG design. There are three assumptions in Levine observed score equating which related to the observed scores in classical test theory: (1) there is a perfect correlation between the true scores of total test and true scores of anchor test in the old and new form populations, (2) the total test true scores' regression on to the anchor test true scores are assumed to be the same linear function for the old form and the new form populations, (3) the measurement error variance for X is the same for Populations 1 and 2 (Kolen & Brennan, 2004).

Circle-Arc Equating

Livingston and Kim (2008, 2009) suggested the circle arc method for small-sample data. This method has a curvilinear equating function. There are two kinds of circle arc equating method: the symmetric circle arc and simplified circle arc. Circle arc equating gets its equating function due to arc connecting three points in a Cartesian coordinate system (Babcock et al., 2012). "The upper end of the curve is determined by the maximum possible score on each form. The lower endpoint of the curve is determined by the lowest meaningful score on each form. The middle point on the curve is determined from the data, by equating at one point in the middle of the score distribution. If those three points happen to lie on a straight line, that line is the estimated equating curve. If three points do not lie on a straight line, they determine an arc of a circle." (Livingston & Kim, 2009, p.332). Livingston and Kim (2008) reported that the circle arc method typically yielded more precise and less biased results than other methods (mean, linear and smoothed equipercentile equating) in small samples.

Differential Item Functioning

The other issue that should be considered in national and international testing programs differential item functioning (DIF). The purpose of DIF analysis is to determine items that function differently for examinees which have the same underlying ability from different subgroups. DIF studies are usually carried out regarding to reference and focal groups that are established by considering manifest (observed) group characteristics such as gender and ethnicity. It is supposed that the observed groups are homogeneous subgroups. In line with this assumption, an item containing DIF is considered advantageous or disadvantageous for all individuals in any manifest groups. Therefore, with these studies, once a DIF item has been determined, there is little knowledge about the examinees for which the item functions differentially (Cohen & Bolt, 2005; De Ayala, Kim, Stapleton, & Dayton, 2002). However, there is a low relationship between the manifest characteristic associated with DIF and the actual advantaged or disadvantaged groups. Therefore, comparisons of item responses for manifest groups may lack sensitivity to determine the true source(s) of DIF (Cohen and Bolt, 2005; De Ayala et al., 2002; Oliveri, Ercikan, & Zumbo, 2013; Samuelsen, 2005).

Latent group means that to set the group membership to an unknown homogeneous subgroup which can be determined by mixture modeling (McLachlan & Peel, 2000). In mixture modeling, while

the item functions the same in a latent group, it functions differently among latent groups (Fieuws, Spiessens, & Draney, 2004). So the use of mixture IRT models can overcome this problem that rises with the use of manifest groups. They can make it possible to detect latent groups for which the items function differently (Cohen & Bolt, 2005). In this study, DIF analyses were conducted based on latent classes.

Latent DIF

The mixture model is defined by Rost (1990) as a "Mixture Rasch" model. It is a combination of latent class and the Rasch model. In this model, it is assumed that a population of examinees can be grouped into several discrete latent classes based on examinees response patterns. With this model, item parameters can be simultaneously estimated with individual's ability and the class he/she belongs (Alexeev, Templin, & Cohen, 2011; Cohen and Bolt, 2005; Mislevy and Verhelst, 1990; Rost, 1990). In these models, each latent class fits Rasch model but classes have different item difficulty parameters. Therefore, MixIRT models can simultenously determine subpopulations that display qualitative differences and quantify the differences in the ability within the groups (Mislevy & Verhelst, 1990; Rost, 1990). According to the model, the possibility of giving a correct answer is as follows.

$$P\left(y_{ijg}=1 \mid g,\theta_{jg}\right)=\frac{1}{1+\exp[-(\theta_{jg}-\beta_{jg})]} \qquad (1)$$

In equation 1, g=1,…,G refers to index with specified latent class; j=1,…,J refers to index with specified responders; θ_{jg}: j. refers to examinees latent ability in g latent class; β refers to difficulty parameter of i. item in class g.

If the DIF detection is carried out during the test construction process, the test developers usually delete flagged items from the test. However, in many situations DIF can be detected after data have been collected. In these situations, deleting DIF detected items may not be a good idea because item deletion can affect test reliability and validity negatively (Elosua & Hambleton, 2018). Hu and Dorans (1989) reported that deleting both minimal and sizable DIF items resulted in different scaled scores after IRT true score re-equating and Tucker re-equating. They also noted that the deleting item itself had a noticeable effect on scale score and the effect size of the DIF item had a less prominent effect on the scale scores (cited in Kolen & Brennan, 2005). Therefore, it is important to determine DIF items during test equating process and apply methods which reduce the effect of these items on test equating constants (Hidalgo-Montesinos & Lopez-Pina, 2002).

In literature, there are some researches (Atalay-Kabasakal & Kelecioğlu, 2015; Chu, 2002; Demirus & Gelbal, 2016; Turhan, 2006; Yurtçu & Güzeller, 2018) which have been compared equating methods in presence of DIF items. In these studies, IRT equating methods were compared in the presence or absence of DIF items. There is not any study which compares small samples equating methods in the presence and absence of DIF items in tests. Therefore this study aims to compare the performance of different small sample equating methods in the presence and absence of DIF in common items.

METHOD

In this study, performance of existing small equating methods was compared in the presence and absence of DIF in common items with using real data. Therefore, this study was designed as descriptive survey. In descriptive survey design, there is not any attempt to change or influence the study situation, existing situation is described (Karasar, 2009).

Data

The data used in this study is 8th-grade mathematics test item responses which obtained from Trends in International Mathematics and Science Study (TIMSS) 2015 Turkey (N=6079) sample. Item responses from Booklet-1 and Booklet-14 are chosen for this study. There are 14 dichotomous scored items common for both booklets. Booklet 1 consists of totally 32 dichotomously scored items and the maximum score which can be obtained are 32. Booklet 14 consists of totally 26 dichotomous scored and 1 polytomous scored (0-1-2) items and the maximum score which can be taken from Booklet 14 is 28. There are 199 students who took Booklet 1 and 224 students who took Booklet 14.

Table 1. Summary of Data

	Common Items	Total of Items	Maximum Score
Booklet-1 (N=199)	14 MC	32 MC	32
Booklet-14 (N=224)	14 MC	26 MC + 1 PS	28

Note. MC: Multiple Choice Item; PS: Polytomous Scored Items

Data Analyses

In this study, data analyses were completed in four steps. In the first step, the confirmatory factor analyses are carried out for Booklet 1 and Booklet 14 to assure the unidimensionality requirement for DIF detection and test equating methods. In the second step, DIF analyses are conducted with Mantel Haenszel (MH) and logistic regression (LR) methods based on latent class. In the third step, Booklet 1 is chosen as a base form and Booklet 14 chosen as a new form, then test equating is conducted under common item nonequivalent groups design. Test equating is done in two phases: the presence of DIF items in anchor test and removing sizable (C level) DIF items from anchor test. In this study, Tucker linear equating, Levine observed score equating, unsmoothed chained equipercentile equating, chained equipercentile equating with presmoothing (C=4), and simplified circle arc equating methods are considered. These equating methods are chosen because the researches showed that they gave accurate equating results in small samples (Babcock et. al, 2011, Kim & Livingston, 2010). Equating results are evaluated based on the standard error of equating, bias and root mean squared error (RMSE) index which provided from 1000 bootstrapped samples.

The Mplus (Muthen & Muthen, 1998-2012) computer program is to assess unidimensionality assumption; the WINMIRA (reference) computer program is used to find how many latent groups exist in the data and to estimate item parameters MixRasch analysis; "difR" R package (Magis, Beland and Raiche, 2015) is used to find DIF items across latent class; "equate" R package (Albano, 2017) is used for test equating and calculating bootstrapped standard error, bias, and RMSE indexes. In "equate" package of R, as reported Albano (2017) "standard errors are calculated as standard deviations over replications for each score point; bias is the mean equated score over replications, minus the criterion; RMSE is the square root of the squared standard error and squared bias combined." (p. 5).

RESULTS

Descriptive Statistics and Testing Assumptions

The descriptive statistics for Booklet 1 and Booklet 14 are reported in Table 2. As seen in Table 2, Booklet 1 has 32 multiple choices (MC) items and Booklet 14 has 26 MC and 1 polytomous scored (0-1-2) items. In both forms, the 14 MC items are common. The mean for Booklet 1 is 14.51 and for Booklet 14 is 12.61 and mean test difficulty is equal for both forms ($z=0.00$, $p>0.05$). Cronbach alpha reliability is .93 for Booklet 1 and .91 for Booklet 14 and there is no statistical difference between forms' reliability level ($z=.75$, $p>.05$). Item discrimination of items in each form was calculated by using point-biserial correlation coefficient. The mean of point-biserial correlations was the same and .50 for Booklet 1 and Booklet 14 (see Table 2).

Table 2. Descriptive Statistics for Booklet 1 and Booklet 14

	Booklet 1	Booklet 14
N	199	214
Number of items	32 MC	26 MC + 1 PC
Common items	14 MC	14 MC
Minimum score	3	1
Maximum score	32	27
Mean	14.51	12.61
Mode	7	7
Median	11.00	11.50
SD	8.24	7.14
Mean difficulty	.45	.45
Mean r_{pb}	.50	.50
Cronbach's Alpha	.93	.91

Note. N: Total number of students; SD: Standard deviation; rpb: Point biserial correlation

Prior to DIF analyses and equating test forms, confirmatory factor analyses (CFA) is carried out for Booklet 1 and Booklet 14 by Mplus (Muthén & Muthén, 1998-2012). Comparative fit index (CFI), Tucker Lewis index (TLI) and Root mean square error of approximation (RMSEA) indexes for Booklet 1 (CFI= .99, TLI= .99, RMSEA= .027) and Booklet 14 (CFI= .99, TLI= .99, RMSEA= .019) support that each form measures a unidimensional trait (Byrne, 2010; Hu & Bentler, 1999; Kline, 2005).

Estimation of Model Parameters

To determine fitted latent classes to the model, results of model comparison criteria for mixture Rasch solutions given in Table 3 is examined.

Table 3. Results of Model Comparison Information Criteria for Mixture Rasch Solutions

Number Of Classes	BIC
1	6505.23
2	6452.29
3	6512.80

In the MixIRT model applications, information criteria Akaike information criterion (AIC) and Bayesian information criterion (BIC) have been widely used to select the model. Li, Cohen, Kim, and Cho (2009) suggested that the smallest BIC result should be used to determine the number of classes. Based on BIC values in Table 3, we can say that a model with two latent classes' best fit the data.

Results of DIF Analyses

DIF analyses were conducted by using Mantel Haenszel (MH) and logistic regression (LR) methods based on two latent classes. The DIF results are reported in Table 4.

Table 4. The Results of DIF Analyses

	MH			**Logistic Regression**		
	ΔMH	p	DIF Level	R^2	p	DIF Level
Item 7	-3.63	.02	C	.070	.00	C
Item 13	12.68	.00	C	.047	.00	B

As seen in Table 4, the DIF analyses results showed that two items (7 and 13) sizable (C level) DIF based on Mantel Haenszel and logistic regression methods. In Mantel Haenszel method, Dorans and Holland's (1993) effect size and in Logistic regression method Gierl, Khaliq and Boughton's (1990) DIF cut points are used.

Results of Equating

In this study, five equating methods are considered: Tucker linear equating, Levine observed score equating, chained equipercentile equating no smooth, chained equipercentile equating with presmoothing (C=4), and simplified circle arc equating methods with nominal weights. As mentioned before for the polynomial log-linear presmoothing method, choosing the degree of the polynomial (C) is important. For this study, we compare chi-square values under different smoothing parameter. The moments and fit statistics for presmoothing are given in Table 5.

Table 5. The Moments and Fit Statistics

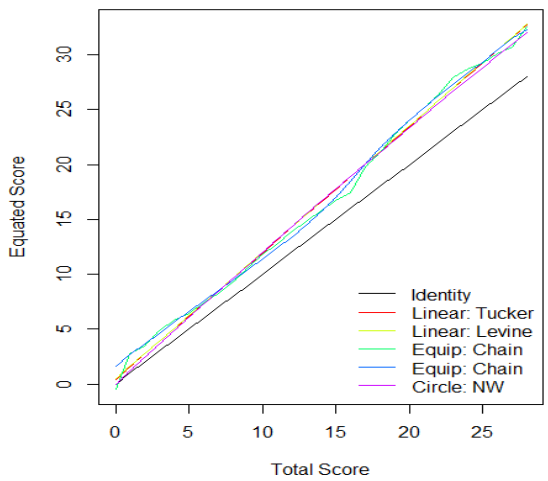
Form	Smoothing Parameter	$\bar{\mu}$	$\bar{\sigma}$	$\overline{sk}$	$\overline{ku}$	$X^2(df)$	$X_C^2 - X_{C+1}^2$
Booklet 1	C=5	14.51	8.22	.57	1.91	16.12 (27)	2.38
	C=4	14.51	8.22	.57	1.91	16.93 (28)	0.81
	C=3	14.51	8.22	.57	2.42	76.15 (29)	59.22
	C=2	14.51	8.22	.16	2.13	105 (30)	28.97
	C=1	14.51	9.45	.19	1.85	120 (31)	15.63
Booklet 14	C=5	12.60	7.13	.28	1.74	15.43 (27)	0.68
	C=4	12.60	7.13	.28	1.74	15.73 (28)	0.29
	C=3	12.60	7.13	.28	2.43	75.63 (29)	59.90
	C=2	12.60	7.13	.31	2.47	75.78 (30)	0.15
	C=1	12.60	9.15	.44	2.06	121.88 (31)	46.10

Note. The chosen C parameter for presmoothing is shown in boldface.

As seen in Table 5, for Booklet 1, C=4 the overall X^2 statistic is not significant ($X^2_{(28)}$ =16.93, p>.05) and the difference statistics for chi-square $X^2_{C=}4$-$X^2_{C=5}$ equals .81 and it is not significant at .05 level for one degree of freedom (X^2<3.84). Based on results, for C≥4 model fit the data and C=5 do not improve the fit of data. For Booklet 14, as seen in Table 5, C=4 the overall X^2 statistic is not significant ($X^2_{(28)}$ =15.73, p>.05) and the difference statistic for chi-square $X^2_{C=4}$-$X^2_{C=5}$ equals .29 and it is not significant at .05 level for one degree of freedom (X^2<3.84). Based on results, again for C≥4 model fit the data and C=5 do not improve the data fit. For Booklet 1 and Booklet 14, C=4 is chosen for presmoothing.

In this study, Booklet 1 is a base form and Booklet 14 is a new form and test equating is conducted under CINEG design. Test equating is done in two phases: the presence of DIF items in the anchor test and removing sizeable DIF items from anchor test. The Figure 1 shows that equated scores versus total scores in the presence and absence of sizeable DIF items in the anchor test.

a) DIF items in anchor test



b) After removed DIF items from anchor test

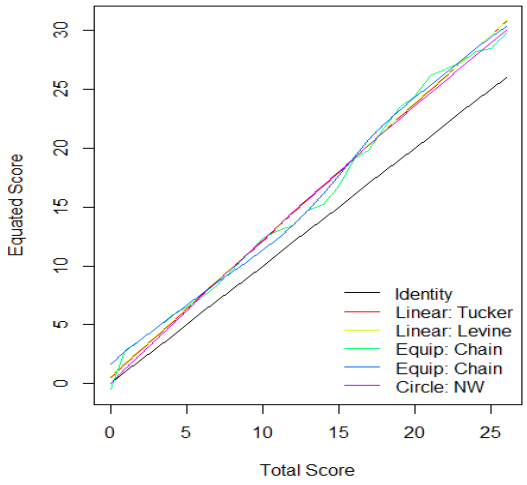


Figure 1. Equated Scores versus Total Scores

In Figure 1, the red line represents Tucker linear equating, the yellow line represents Levine observed score equating, the green line represents chained unsmoothed equipercentile equating, the blue line represents chained presmoothed (C=4) equipercentile equating and the purple line represents simplified circle arc equating method. Also in Figure 1, the black line belongs to identity equating and it means that there is no equating between old form and new form scores. As seen for both conditions Tucker, Levine and circle arc methods yield similar equated scores; their lines in the graphics are almost top of each other. In unsmoothed chained equipercentile equating method, there are some irregularities between equated scores and total scores (see the green line in Figure 1). The random error in estimating equivalent scores causes these irregularities. As seen in Figure 1, these irregularities got lost in presmoothed equipercentile equating (see the blue line).

The performance of different equating methods in presence and absence of DIF items in anchor test was evaluated based on standard errors of equating, bias and RMSE values which provided from 1000 bootstrapped samples and reported in Table 6.

Table 6. Equating Results

Equating Methods	Presence of DIF Items in Anchor Test			Absence of DIF Items in Anchor Test		
	se	Bias	RMSE	se	Bias	RMSE
Tucker Linear Equating	0.36	0.52	0.63	0.35	0.52	0.63
Levine Observed Score Equating	0.37	0.52	0.63	0.39	0.51	0.64
Unsmoothed Chained Equipercentile Equating	0.77	0.54	0.94	0.72	0.61	0.94
Presmoothed (C=4) Chained Equipercentile Equating	0.49	0.19	0.52	0.51	0.17	0.53
Simplified Circle - Arc Equating	0.19	0.69	0.72	0.19	0.72	0.74

As seen in Table 6, when the common items include DIF items, simplified circle arc equating method has the least (.19) standard error of equating (se) and unsmoothed chained equipercentile equating method has the largest (.77) standard error of equating. On the other hand, when we consider bias as a criterion simplified circ-arc method has the largest (.69) amount of bias, 4-moments presmoothing chained equipercentile equating has the smallest amount of bias value. Levine and Tucker equating methods have the same (.52) and smaller bias values than unsmoothed equipercentile equating method. We can say that Tucker linear equating and Levine observed score equating methods show similar and better performance than the unsmoothed chained equipercentile equating method. According to last criteria of RMSE values, again smoothed chained equipercentile equating method has the smallest (.52) RMSE value and the unsmoothed equipercentile equating method has the largest (.94) RMSE value. Tucker linear equating and Levine observed score linear equating methods had the same (.63) and smaller RMSE values than simplified circle-arc equating method (.72). Again we can say that Tucker linear equating and Levine observed score linear equating methods show similar and better performance than the simplified circle-arc equating method.

After removing two sizeable DIF items from anchor test, the similar results have been found (See Table 6). Again based on se criteria, the simplified circle arc method was the best and the unsmoothed chain equipercentile equating method was the worst. On the other hand, based on bias criteria the best equating method is presmoothed equipercentile equating method and the worst one is simplified circle arc equating method. Concerning RMSE values, the best one is again presmoothed chained equipercentile equating method and the worst one is unsmoothed chained equipercentile equating method. Also, according to results, we can say that performances of equating methods are similar with the presence and not presence of DIF items in anchor test and we can say that there is no notable change in se, bias and RMSE values.

Another result of this study is that whether or not common items include DIF items, unsmoothed chained equipercentile equating method has larger se, bias and RMSE values than presmoothed (C=4) chained equipercentile equating method.

CONCLUSION AND DISCUSSION

In this study, five equating methods were considered: Tucker linear equating, Levine observed score linear equating, unsmoothed chained equipercentile equating, chained equipercentile equating with presmoothing (C=4), and simplified circle arc equating methods with nominal weights. Equating methods were compared in two phases: the presence of DIF items in anchor test and removing sizeable DIF items from anchor test. The results show that performances of equating methods are similar to presence and absence of DIF items in anchor test and there is no notable change in se, bias and RMSE values. Also, results show that according to the standard error of equating criteria, the circle arc equating method outperformed other equating methods but based on bias evaluation criteria its performance was the worst one in both situations.

As Kolen and Brennan (2004) reported standard error of equating is the standard deviation of equivalent scores over replications of the equating process and random error indexed by the standard error of equating. Standard error equating is closely related with sample size and as the sample size becomes larger it becomes smaller. The result of this study showed that the circle arc method has the minimum se among other equating methods. The circle-arc method especially suggested for small samples (Livingston, & Kim, 2009) and in their study, Kim and Livingston (2010) showed that in small samples the circle arc method clearly outperformed other equating methods (chained equipercentile, Levine, chained linear, Tucker) based on bias, RMSD and, se evaluation indexes. The results of our study supported Kim and Livingston (2010) only in terms of random equating error. Also, among other equating methods the unsmoothed chained equipercentile equating has the largest se value and we can say that based on random equating error its performance was the worst. As seen in Figure 1 there are some irregularities between equated scores and total scores and Kolen and Brennan (2004) noted that the reason for these irregularities is random equating error. Also in our study, the results of unsmoothed chained equipercentile equating method based on se support this view.

Based on bias and RMSE evaluation criteria, smoothed chained equipercentile equating method is the best equating method and unsmoothed chained equipercentile equating method is the worst method. In one of a simulation study, Aşiret and Sünbül (2016) shows that for sample size 200 presmoothed equipercentile equating method produced more accurate results than other methods (linear, circle-arc, mean). Our study results supports this finding with real data which has sample size roughly 200. Tucker linear equating and Levine observed score equating methods show similar and better performance than the unsmoothed chained equipercentile equating method. To all evaluation indexes, smoothed chained equipercentile equating has lower values than the unsmoothed equipercentile equating method. We can say that presmoothing tended to decrease random and systematic equating error as in shown other studies (Aşiret & Sünbül, 2016; Özdemir, 2017; Kelecioğlu & Öztürk Gübeş, 2013; Livingston, 1993; Skaggs, 2005). As a result, we can also say that presmoothed chained equipercentile equating yields more precise and accurate equating results than unsmoothed chained equipercentile equating as is assumed (Kolen and Brennan, 2004).

RECOMMENDATIONS

This study is limited with 8th-grade mathematics data from Booklet 1 and Booklet 14 in TIMSS 2015 Turkey sample. The results showed that performances of equating methods are similar to the presence and not the presence of DIF items in anchor test and there is no notable change in the error of equating. This result should be interpreted carefully and in further researches effects of DIF on small sample equating methods should be examined with real and simulated data sets more detailed.

Acknowledgements or Notes

This study was presented at the 6th International Congress on Measurement and Evaluation in Education and Psychology, KOSOVA.

REFERENCES

Albano, A. (2017). *equate: Observed –score linking and equating.* [Computer software].

Alexeev, N., Templin, J., & Cohen, A. (2011). Spurious latent classes in mixture rasch model. *Journal of Educational Measurement*, *48*(3), 313-332.

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 508-600). Washington, DC: American Council on Education.

Aşiret, S., & Sünbül, S. Ö. (2016). Investigating test equating methods in small samples through various factors. *Kuram ve Uygulamada Eğitim Bilimleri*, 16(2), 647-668.

Atalay-Kabasakal, K. & Kelecioğlu, H. (2015). Effect of differential item functioning on testequating.*Educational Sciences: Theory & Practice, 15*(5), 2015, 1229-1246.

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational Measurement* (2nd ed.). Washington, DC: American Council on Education.

Babcock, B., Albano, A., & Raymond, M. (2012). Nominal weights mean equating: A method for very small samples. *Educational and Psychological Measurement, 72*(4), 608-628.

Byrne, B. M. (2010). *Structural equation modeling with AMOS*, (2nd ed.). New York: Routledge.

Chu, K. L. (2002). *Equivalent group test equating with the presence of differential item functioning* (Doctoral dissertation). Available from ProOuest Dissertations and Theses database.

Cohen, A.S., & Bolt, D. M. (2005). A mixture model analysis of differential item functioning *Journal of Educational Measurement, 42*(2), 133-148.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Mason, OH: Cengage Learning.

De Ayala, R.J., Kim, S.H., Stapleton, L.M., & Dayton, C.M. (2002). Differential item functioning: A mixture distribution conceptualization. *International Journal of Testing,3*(4), 243-276.

Demirus, K. B., & Gelbal, S. (2016). The study of the effect of anchor items showing or not showing differential item functioning to test equating using various methods. *Journal of Measurement and Evaluation in Education and Psychology 7*(1), 182-201.

Dorans, N. J. (1990). Equating methods and sampling designs. *Applied Measurement in Education*, *3*(1), 3-17.

Dorans, N. J., & Holland, P. W. (1993). DIF detection and description: Mantel-Haenszel and standardization. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 35-66). Hillsdale, NH: Erlbaum.

Dorans, N. J., Moses, T. P., & Eignor, D. R. (2010). *Principles and practices of test score equating* (ETS Research Report No. RR-10-29). Princeton, NJ: ETS.

Elosua, P., & Hambleton, R. K. (2018). Psychological and educational test scorecomparability across groups in the presence of item bias. *Journal of Psychology and Education, 13*(1), 23-32.

Fieuws, S., Spiessens, B., & Draney, K. (2004). Mixture models. In P. de Boeck & M. Wilson (Eds.), *Explanatory Item Response Models: A Generalized Linear and Nonlinear Approach*. (pp.317-340). New York: Springer.

Gierl, M., Khaliq, S. N., & Bougthon, K. (1999). Gender differential item functioning in mathematics and science: Prevalence and policy implications. Paper presented at the *Improving large-scale assessment in education*. Symposium conducted at the Annual Meeting of Canadian Society for the Study of Education, Canada.

Gulliksen, H. (1950). *Theory of mental tests.* New York: Wiley Hidalgo-Montesinos, M. D., & Lopez-Pina, J. A. (2002). Two-stage equating in differential item functioning detection under the graded response model with the Raju area measures and the lord statistic. *Educational and Psychological Measurement, 62*(1), 32–44.

Hu, L. & Bentler, P. (1999). Cutoff criteria for fit indexes in covariance structure analysis Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*, 1-55.

Karasar, N. (2009). *Bilimsel araştırma yöntemi: Kavramlar, ilkeler, teknikler.* Ankara: Nobel Yayınları.

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.

Kim, S. & Livingston, S. A. (2010). Comparisons among small sample equating methods in a common item design. *Journal of Educational Measurement, 47*(3), 286-298.

Kline, R. (2005). *Principles and practices of structural equation modeling* (2n ed.). New York: Guilford Press.

Kolen, M. J., & Brennan, R. L. (1995). *Test equating: Methods and Practices*. New York: Springer Verlag.

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

Kurtz, A. M., & Dwyer, A. C. (2013). *Small sample equating: Best practices using a SAS Macro*. Retrieved from http://analytics.ncsu.edu/sesug/2013/BtB-11.pdf

Li, F., Cohen, A. S., Kim, S.-H., & Cho, S.-J. (2009). Model selection methods for dichotomous mixture IRT models. *Applied Psychological Measurement, 33*(5), 353-373. doi: 10.1177/0146621608326422

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

Livingston, S. A., & Kim, S. (2008). *Small sample equating by the circle-arc method* (ETS Research Report No. RR-08-39). Princeton, NJ: ETS

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

Magis, D., Beland, S., & Raiche, G. (2015). *difR: Collection of methods to detect dichotomous differential item functioning (DIF)*. [Computer software].

Mclachlan, G. & Peel, D., (2000). *Finite Mixture Models.* John Wiley & Sons, Inc. New York.

Mislevy, R. J. & Norman, V. (1990). Modeling item responses when different subjects employ different solution strategies. *Psychometrika, 55*(2), 195-215.

Muthén, L. K., & Muthén, B. O. (1998-2012). *Mplus user's guide* (7th ed.). Los Angeles, CA: Muthén & Muthén.

Oliveri, M. E., Ercikan, K. Zumbo, B. (2013). Analysis of Sources of Latent Class Differential Item Functioning in International Assessments. *International Journal of Testing*, *13*(3), 272–293. doi: 10.1080/15305058.2012.738266

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.

Parshall, C. G., Du Bose, P., Houghton, P., & Kromrey, J. D. (1995). Equating error and statistical bias in small sample linear equating. *Journal of Educational Measurement, 32*(1), 37–54.

Rost, J. (1990). Rasch models in latent classes: An integration of two approaches to item analysis. *Applied Psychological Measurement*, *14*(3), 271-282. doi: 10.1177/014662169001400305

Samuelsen, K. M. (2005). *Examining differential item functioning from a latent class perspective*. (Doctoral dissertation, Faculty of Graduate School of the University of Maryland, College Park). Retrieved from https://drum.lib.umd.edu/bitstream/handle/1903/2682/umi-umd-2604.pdf?sequence=1&isAllowed=y

Skaggs, G. (2005). Accuracy of random groups equating with very small samples. *Journal of Educational Measurement, 42*(4), 309–330.

Turhan, A. (2006). *Multilevel 2PL item response model vertical equating with the presence of differential item functioning* (Doctoral Dissertation). Available from ProOuest Dissertations and These database.

Von Davier, M. (2001). *WINMIRA* [Computer Software]. Groningen, the Netherlands: ASCAssessment Systems Corporation. USA and Science Plus Group.

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.