International Journal of Progressive Education, Volume 14 Number 2, 2018

The Effect of Mini and Midi Anchor Tests on Test Equating

Çiğdem Akın Arıkan [i]
Hacettepe University

Selahattin Gelbal [ii]
Hacettepe University

Abstract

The main purpose of this study is to compare the test forms to the midi anchor test and the mini anchor test performance based on item response theory. The research was conducted with using simulated data which were generated based on Rasch model. In order to equate two test forms the anchor item nonequivalent groups (internal anchor test) was used in this research. The data were generated and analyzed using the R software. 100 replications were done for each different condition. The results obtained from this simulation study were evaluated according to the equating error (RMSE) and bias (BIAS) criterions. In the study, midi anchor generally produced better equating results than the mini anchor test with a few exceptions for most conditions.

Keywrods: test equating, anchor test design, equating error, bias

DOI: 10.29329/ijpe.2018.139.11

---

[i] **Çiğdem Akın Arıkan**, Res. Assist Dr., Hacettepe University, Department of Education Faculty, Turkey.

Correspondence: cakinarikan@hacettepe.edu.tr

[ii] **Selahattin Gelbal**, Prof. Dr., Hacettepe University, Department of Assessment and Evaluation in Education, Turkey. Email: sgelbal@gmail.com

Introduction

In most testing applications, it is needed to develop different forms of the same test for test safety or testing of individuals on distinct days. Different forms are used to prevent examinees from obtaining preliminary information about questions. Test forms should be equated to compare scores from different test forms and to eliminate possible advantages or disadvantages of individuals who take different forms. Angoff (1971) defines equating as conversion of a test form's unit system into another test form's unit system; while Kolen and Brennan (2004) refer to it as a statistical process which helps using scores obtained from the tests interchangeably by arranging differences between the forms having identical test characteristics.

Test equating consists of three steps. The first stage is about selection of equating design. Equating designs are classified as single-group design, balanced groups design, and anchor-item equivalent and nonequivalent groups designs (internal and external anchor) (Livingston, 2004). In single-group design, the simplest equating design, different forms of the test are applied to the same group. In balanced groups design, the group is divided into two groups to take both forms to be equated. The first group is given form I and II, respectively; while the other group is given form II and I, respectively. In equivalent groups design, examinees with equivalent distribution of capability receive different test form (Cook and Eignor, 1991). In this design, the first student receives Form X, the second student takes Form Y, and the third student takes Form X again and so on. The individuals who receive form X and Y create an equivalent group. Lastly, in nonequivalent groups design, individuals are given different test which comprises of anchor items representing the test forms in terms of content and statistical properties(Livingston, 2004; Kolen and Brennan, 2004).

Once the equating design is determined, decision is made regarding which equating method(s) to use. Equating methods are divided into two main groups: methods based on traditional equating methods and item response theory (IRT). Traditional type includes average, linear, and equipercentile equating; while the other type includes IRT true-score and IRT observed-score equatings. In the latter type of equating based on IRT, once observed score distribution of each form is estimated with IRT models, scores are equated by using equipercentile equating methods. In the former type of IRT-based methods, it is assumed that the true scores associated with a specific θ ability of a form are equal to the true scores associated with the same θ value in the other form (Tong and Kolen, 2005). Since the tests are applied to two different groups, the parameters obtained should be placed on the same scale. For this purpose, the anchor items in the test forms are used for converting scales. In IRT-based equating; scale conversion contains the concurrent calibration (Lord, 1980) and separate calibration methods (Stocking and Lord, 1983; Haebara, mean-mean and mean-sigma). In the concurrent calibration, the item parameters obtained from the two forms are estimated together and it is assumed that the anchor items in both forms have the same parameters (Nozawa, 2008). In the case of separate calibrations, different item parameters are estimated for both test forms. Due to the uncertainty of scale in IRT, the estimated parameters may not be equal and thus they cannot be compared. For this; the equating coefficients A (slope) and B (intercept) are calculated based on a(item discrimination) and b(item difficulty) parameters of the anchor items. With these coefficients, θ value in one form is converted to θ value in the other form. Conversion of the ability parameter from test X into test Y is performed with the equation below with constants A and B.

$$\theta_{Xi} = A\theta_{Yi} + B \qquad (1)$$

Conversion of item parameters from test X to Y is carried out with the following equation.

$$a_{Yj} = a_{Xj}/A \qquad (2)$$

$$b_{Xj} = Ab_{Yj} + B \qquad (3)$$

$$c_{Xj} = c_{Yj}, \qquad (4)$$

As seen in the equation, the parameter c is independent of scale conversion (Kolen&Brennan, 2004).

In this study, mean-sigma was used among separate calibration methods. Mean-sigma method uses the mean and standard deviation of the parameter b (Hambleton, Swaminathan and Rogers, 1991). The equating coefficients, A and B, are calculated with the following equations.

$$A = \frac{bs_x}{bs_y}$$

$$B = \overline{b_y} - \propto \overline{b_x} \qquad (5)$$

bs_x and bs_y : standard deviation of parameter b obtained from anchor items in forms X and Y

$\overline{b_y}$ and $\overline{b_x}$: mean of parameter b belonging to anchor items in form X and Y

In IRT; Rasch, 1, 2 and 3 PLM (Rasch, 1966; Birnbaum, 1957,1958, 1968) are available for dichotomous scoring data, whereas nominal response model (Bock, 1922; Samejima, 1972), gradual response model (Samejima, 1969), and generalized partial credit model (Muraki, 1992; 1993) are available for multiple category scoring data (Embretson and Reise, 2000; Hambleton and Swaminathan, 1985; Kolen and Brennan, 2004). Since Rasch model was used in this study, information is given about this model. It is the most commonly used model among IRT models (Hambleton et al, 1991). In this model, while item discrimination index (parameter a) is the same for all items, item difficulty parameter (parameter b) is varied. Parameters b determines the position of the item characteristics curve on the ability scale. If the item is difficult, it is located to the right of the ability scale, but to the left to the same scale if it is easy. It is assumed that parameter b often takes a value between -3 and +3. Difficulty index of an item is defined as θ value of the point where the probability of correctly answering the item is 0.5. In this model, the only parameter affecting the performance of the individual is b parameter (Hambleton et al, 1991).

After the appropriate equating method is selected and test forms are equated, the resulting equated scores need evaluation. In their study evaluating different equating criteria, Harris and Course (1993) proposed that any single equating criterion would not be suitable for all equating methods, so multiple criteria should be used. In this study, Bias and RMSE were used as evaluation criteria.

Bias: This criterion, which is used for evaluating the systematic error in equating, is obtained by dividing the total difference between the estimated item parameter value and true item parameter value by number of replication. This value was obtained from the equation below.

$$BIAS(\tau_j) = \frac{1}{R}\sum_{r=1}^{R}(\hat{\tau}_{jr} - \tau_j) \qquad (6)$$

Error (RMSE - Root Mean Square Error): It gives the total error by taking into consideration both systematic and random errors. RMSE is obtained by taking the square root of the ratio of total squares of the differences between the true parameter value and estimated parameter value to the number of replications. The equation for RMSE is given below.

$$RMSE(\tau_j) = \sqrt{\frac{\sum_{r=1}^{R}(\hat{\tau}_{jr} - \tau_j)^2}{R}} \qquad (7)$$

R= Number of replication

$\hat{\tau}_{jr}$= estimated value of parameter j, τ_j= true value of parameter j

Aim and Significance of the Study

In the literature the most commonly used equating design is anchor-item test design (Kolen and Brennan, 2004). Anchor items are grouped as internal and external. In internal anchor-item test, scores taken from anchor items are added to the examinee's total score. Conversely, the scores are not added to the total score in external type. Furthermore, while external anchor test is administered to the examinee as a separate form, internal anchor test is an integral part of the main test (Kolen, 1988; Zhu, 1998). Angoff (1971) argued that anchor items should be a parallel miniature of the whole test. Kolen and Brennan (2004) also stated that such items should represent the whole test regarding both difficulty level and content. Many other researchers suggest that the anchor item set must be a smaller version of the total test (Budescu, 1985; Kolen, 2007; Kolen, 1988; Petersen, Kolen, and Hoover, 1989). However, Sinharay and Holland (2007) referred to the lack of evidence regarding the necessity for identical distribution of difficulty level among items in the anchor test and emphasized that such a criterion is quite restrictive. They added that the anchor test with identical scope and average difficulty with equated test but anchor items showing smaller item difficulty distribution than the total test difficulty could yield better equating results. According to Sinharay and Holland (2006), three different anchor tests can be generated from the anchor item difficulty distribution. If the anchor test is similar to the total test in scope and statistics, it is called mini-test. If the anchor test identical to the total test in scope but different in item difficulty distribution all items being at medium level of difficulty, it is called midi-test. However, if the anchor item difficulty distribution falls between mini and midi test, it is called semi-midi test. In studies with smaller anchor item difficulty distribution than in the main test, error was found to be lower (Antal, Proctor and Melican, 2014; Hagge, 2010; Fitzpatrick and Skorupski, 2016; Kim, 2014; Liu, Sinharay, Holland, Feigenbaum &Curley, 2011; Sinharay, Haberman, Holland and Lewis, 2012; Sinharay and Holland, 2006, 2007). Sinharay, Holland and others used midi anchor test as internal anchor test and also used midi anchor test in external anchor test because of difficulty of meeting statistical requirements of the total test. In this study, because of scarcity of examples with internal anchor test (Fitzpatrick and Skorupski, 2016; Kim, 2014), a comparison was made between performance of mini and midi anchor test in internal anchor test. Moreover, review of literature provided no specific example comparing equating results in the Rasch model. In present study, Rasch model was selected from IRT models and mean-sigma method was selected from item calibration methods. This is because parameter b is the only parameter used in both Rasch model and mean-sigma model calibration method, and anchor item difficulty distribution is obtained by changing the parameter b. The aim of the study is to compare performances of midi and mini anchor tests in IRT-based equating.To this end, the study was carried out seeking answer for the following question.

In the case of tests equated using mean-sigma as a seperate calibration method in nonequivalent anchor test (NEAT) design with Rasch-model; how does each of the

- slop coefficient,
- intersept coefficient, and
- bias and total error regarding equated scores differ in relation with sample size, ability distribution, difference between difficulties of test forms and anchor item difficulty?

Method

Generating and Obtaining Study Data

The data generated in this study were equated by using NEAT design. The study also discusses sample size, ability distribution, difference of difficulty between test forms, and anchor item difficulty distribution circumstances. In NEAT design, two different groups are applied two different

tests with anchor items. Equating pattern consists of the old form (Form X), the new form (Form Y) and the anchor items. Item response patterns of respondents taking Form X and Form Y are generated separately. The data generation procedure was conducted according to the Rasch model. In Rasch model, the only item parameter is the difficulty parameter *b*. The item difficulty parameter of Form X was derived from the constant and standard normal distribution *(b~ (0,1))* under all circumstances. The b parameter of Form Y was generated under two circumstances as *(b~ (0.05,1))*and*(b~ (0.2,1))*. Since the purpose of equating is to statistically arrange item difficulties between test forms, distribution of item difficulty parameters was differentiated. Also *b* parameters of the anchor items were studied under two conditions. In the case of mini anchor test with item difficulty distribution identical to the main test, the parameter *b* was chosen with zero mean and one standard deviation *(b~ (0,1))*; in the case of midi anchor test with anchor item difficulty distribution smaller than the main test, it was chosen with zero mean and standard deviation corresponding to 50% of the main test, which is 0.5. Kolen and Brennan (2014) pointed out that equating in Rasch model requires at least 500 subjects in each group. In this study, three different sample sizes were taken for each group as 250 (small), 500 (ideal) and 2000 (large).For the group taking the old form, both mean and standard deviation values were derived from the standard normal distribution (N~(0,1)) and taken as constant under any circumstance. As for the group taking the new form, ability distribution was examined under three conditions. The mean and standard deviation of the ability distribution were derived as (θ~N(0.05,1)), (θ~N(0.5,1)) and (θ~N(0.25,1.25)). The number of test items was designated as 50 considering the number of items in national scale tests held in Turkey. It is suggested that the number of anchor items should be about 20% of the entire test items (Angoff, 1971; Budescu, 1985; Kolen and Brennan, 2004). Hambleton et al (1991) stated that the number of anchor items should be around 20% to 25% of the entire test items. Therefore, the number was fixed at 20% in this study. While the number of anchor items is 5 in a test of 25 items, it was planned as 10 for a test with 50 items.

The study examines the performance of IRT true score equating under mean-sigma as a seperate calibration method applying to 36 conditions for simulations covering 3 different sample sizes, 3 ability distributions, 2 different test form difficulty difference and 2 different anchor item difficulty distributions. Mean-sigma was preferred as a calibration method because it uses item difficulty parameter only. All replications for the test forms were conducted on the software called R. Item and ability parameter estimates were conducted by the *ltm package* in R (Rizopoulos, 2015). Item parameters were estimated by using the Marginal Maximum Likelihood and ability parameters were estimated by using Expected a Posteriori (Rizopoulos, 2015).After item parameter estimations were completed, scale calibrations and equating were performed by using the codes written by researcher. To obtain more stable estimates, 100 replications were performed for each condition.

Results

This section gives results regarding bias and error values obtained from equated scores and equating slope and the intercept coefficients according to the equating operations carried out in NEAT design under simulation conditions.

Results regarding Research Questions

In relation with first research question, bias and error results obtained for slope parameter among equating coefficients under conditions of sample size, difficulty difference between test forms, distribution of ability and distribution of anchor item difficulty are given in Table 1. Slope parameter is calculated based on standard deviation of the parameter *b* obtained from anchor items.

Table 1. Bias and Error Values Obtained from Slope Parameter

Sample Size	Difficulty Difference	Common Item Type	Ability Distribution of the Group Taking the New Form					
			N(0.05;1)		N(0.5;1)		N(0.25;1.25)	
			Bias	RMSE	Bias	RMSE	Bias	RMSE
250	0.05	Mini	-0.001*	0.071	-0.004*	0.107	0.015	0.068
	0.05	Midi	-0.003*	0.108	-0.004*	0.109	0.017	0.110
500	0.05	Mini	0.001	0.052	0.004	0.083	0.016	0.060
	0.05	Midi	0.000	0.097	0.004	0.083	0.012	0.093
2000	0.05	Mini	0.001	0.026	0.003	0.026	0.015	0.031
	0.05	Midi	0.000	0.037	0.001	0.04	0.013	0.043
250	0.2	Mini	-0.003*	0.068	-0.006*	0.067	0.017	0.068
	0.2	Midi	-0.005*	0.110	-0.004*	0.110	0.019	0.110
500	0.2	Mini	0.001	0.052	0.001	0.084	0.016	0.058
	0.2	Midi	0.001	0.097	0.003	0.082	0.012	0.093
2000	0.2	Mini	0.001	0.027	0.002	0.028	0.015	0.031
	0.2	Midi	0.001	0.038	0.001	0.041	0.013	0.043

* negative biased

Table 1 shows that bias values of slope parameters are quite low when ability distribution is (0.05; 1) and (0.5; 1); but negative bias estimates are obtained in samples of 250 examinees. In cases of ability distribution (0.25; 1.25), bias values are seen to increase in comparison to other conditions of distribution. While midi anchor tests yield higher rates of bias in ability distribution at (0.25; 1.25) and sample of 250 examinees, they yield smaller bias or remain unchanged under all other conditions. In addition, it is observed that as sample size increases, bias values remain the same or decrease in all conditions. In addition, it is observed that as sample size increases, bias values remain the same or decrease in all conditions.

When total error values of slope parameter are examined, it is seen that anchor item difficulty distribution varies in all conditions. In other words, midi tests yield higher error rates. As sample size increases, error rate seems to decrease. Besides, as item difficulty difference increases, error remained unchanged or change inconsiderably in many conditions.

Table 2 shows bias and error results obtained from the equating intercept parameter, which is one of the equating coefficients under conditions of sample size, difficulty difference between forms, ability distribution and anchor item difficulty distribution. Equating intercept parameter is calculated from mean of the parameter b obtained from anchor items.

Table 2. Bias and Error Values Obtained from Equating Intercept Parameter

Sample Size	Difficulty Difference	Common Item Type	Ability Distribution of the Group Taking the New Form					
			N(0.05;1)		N(0.5;1)		N(0.25;1.25)	
			Bias	RMSE	Bias	RMSE	Bias	RMSE
250	0.05	Mini	-0.012*	0.066	0.008	0.078	-0.012*	0.066
	0.05	Midi	-0.010*	0.062	0.006	0.063	-0.017*	0.066
500	0.05	Mini	-0.007*	0.058	0.003	0.062	-0.002*	0.059
	0.05	Midi	-0.012*	0.055	0.003	0.062	-0.007*	0.055
2000	0.05	Mini	-0.004*	0.023	-0.004*	0.025	0.001	0.023
	0.05	Midi	-0.009*	0.023	-0.013*	0.031	0.006	0.023
250	0.2	Mini	0.015	0.067	-0.013*	0.078	-0.001*	0.063
	0.2	Midi	0.001	0.063	-0.011*	0.077	-0.016*	0.063
500	0.2	Mini	0.001	0.061	-0.006*	0.063	-0.006*	0.061
	0.2	Midi	0.002	0.053	-0.007*	0.063	-0.017*	0.058
2000	0.2	Mini	0.002	0.023	-0.005*	0.025	0.000	0.023
	0.2	Midi	0.002	0.023	-0.013*	0.031	0.004	0.024

* negative bias

It is understood from Table 2 that negative biased estimates are found in all sample sizes when difficulty distribution between tests is similar at ability distribution of (0.05,1); in samples of 2000 examinees when difficulty distribution between tests is similar at ability distribution of (0.05,1); in all sample sizes with increased difference of difficulty between forms; in both differences of difficulty at ability distribution of (0.05,1), and lastly in samples of 250 to 500 . Among negative biased estimates, midi anchor tests result in lower bias except in samples of 250. As for positive biased estimates; similar or smaller estimates are obtained from mini and midi anchor tests at ability distribution of (0.05,1) and (0.5,1), but mini anchor tests give smaller values in samples of 2000 examinees at ability distribution of (0.25,1.25).

As total error values from the equating intercept parameter are examined, it is seen that the highest error rate occurs in the case of increased ability distribution mean (0.5,1), while close mean error rates are obtained under the other two ability distribution conditions. The highest error rate occurs at ability distribution (0.5,1) which has higher meanwhile error rates are found to be close in the case of the other two distributions. In all circumstances, midi anchor test yields smaller error rates than mini test or mini and midi anchor tests result in equal error rates. When difficulty difference between forms increases, error rate remains the same or change little in many conditions. Moreover, larger samples lead to smaller error.

Bias and error values obtained from equated scores under conditions of sample size, difference of difficulty between forms, test length and anchor item difficulty distribution are shown in Table 3. The bias and error rates for this parameter are calculated by taking the mean of the values found for each of the equated scores.

Table 3. Bias and Error Values Obtained from Equated Scores

Sample Size	Difficulty Difference	Common Item Type	Ability Distribution of the Group Taking the New Form					
			N(0.05;1)		N(0.5;1)		N(0.25;1.25)	
			Bias	RMSE	Bias	RMSE	Bias	RMSE
250	0.05	Mini	-0.025	0.476	-0.017	0.451	-0.019	0.468
	0.05	Midi	-0.021	0.451	-0.019	0.420	-0.034	0.450
500	0.05	Mini	0.013	0.397	0.015	0.354	0.017	0.411
	0.05	Midi	0.017	0.370	0.017	0.324	0.019	0.365
2000	0.05	Mini	0.008	0.169	0.014	0.176	0.015	0.174
	0.05	Midi	0.009	0.159	0.013	0.162	0.015	0.165
250	0.2	Mini	-0.024	0.484	-0.013	0.502	-0.006	0.48
	0.2	Midi	-0.008	0.482	-0.015	0.496	-0.019	0.492
500	0.2	Mini	0.015	0.405	0.016	0.423	0.019	0.415
	0.2	Midi	0.018	0.400	0.018	0.394	0.020	0.372
2000	0.2	Mini	0.011	0.171	0.015	0.178	0.017	0.175
	0.2	Midi	0.011	0.170	0.015	0.174	0.018	0.171

*negative bias

Table 3 shows that equated scores generate negative bias in samples of 250 examinees and positive bias in the other samples. As for samples of 250 examinees, smaller bias are obtained from mini anchor test at ability distribution of N(0.05,1), while bias are smaller in midi anchor test as a result of the other two distributions. Mini anchor test gives smaller bias in samples of 500, while midi anchor test yields smaller or equal bias in samples of 2000. Total error rates for equated scores reveal that midi anchor test results lower rates than mini test in all conditions except for the sample of 250 examinees when distribution of ability is at (0.25,1.25) and difference of difficulty between forms increases. Also larger samples lead to smaller error rates. However, increased difference of difficulty among forms leads to increased error rates in all conditions. Of all, the highest error rate is recorded against increased difference of difficulty among tests (0.2) and ability distribution of (0.5,1).

Apart from that, ANOVA analysis was performed to follow up variance of bias and errors under conditions provided in this study. As a result of ANOVA analysis, bias and total error rates of equated scores with equating slope and intercept parameter are demonstrated in Table 4. The table reveals F and eta square (***η2)*** values of significant effects.

Table 4. Significant ANOVA Results of Bias and Equating Error Values for Equated Scores

			Results		
	Criteria	***Effects***	***df***	***F***	***η2***
Equated Scores	**Error**	SS	2	131,001*	0,99
		ADD	1	107,245*	0,859
		SS*AB*ADD	4	14,754*	0,937
		SS*GF*ADD	2	19,831*	0,908
	Bias	SS*AB*ADD	4	33,406*	0,971
		SS*AB*DDF	4	10,837*	0,916
A	**Bias**	AB	2	21,005*	0,949
		SS *ADD	2	11,795	0,901
B	**Error**	SS	2	69,604	0,964
		SS*ADD	4	7,007	0,873

*p<0.002 (SS-sample size; AB-ability distribution; DDF-difference of difficulty between forms, ADD-anchor item difficulty distribution)

The ANOVA results in Table 4 reveal that the interaction effect of sample size, ability distribution, anchor item difficulty distribution and sample size, ability distribution, and difference of difficulty between forms are statistically significant. Moreover, main effect of ability distribution is at significant level in slope parameter equating. Of all interaction effects; the interaction of sample size and anchor item difficulty distribution is found significant. As for effect sizes for bias, main effect and interaction effects seem to be large (Cohen, 1988). Equating errors as a result of ANOVA in equated scores Table 4 indicate that of all study variables, sample size and anchor item difficulty distribution's main effects seem to be at significant level. As for interaction effects; sample size, ability distribution and anchor item difficulty distribution interaction and sample size, difficulty difference and anchor item difficulty distribution interaction is statistically significant. According to results of equating intercept parameter, the main effect of sample size seems to be at significant level; so does the interaction between sample size and anchor item difficulty distribution. Lastly, effect size for errors reveals that main effect and interaction effects generate larger effects (Cohen, 1988).

Conclusion and Discussion

This study was carried out to investigate the effect of sample size, difficulty between forms, ability distribution and anchor item difficulty distribution on bias and total error rate through the use of IRT true-score (Rasch model) equating model. For this purpose, data was generated taking into account the conditions applying to real data applications.

In relation with equating slope parameter, average of ability distribution and standard deviation (0.25, 1.25) between groups result in increased bias values compared to the other conditions of ability distribution. The finding suggests that differential distribution of ability between groups leads to increased systematic error. Except for several conditions, the bias values obtained from midi anchor test and mini anchor test seem unchanged or the former yields lower bias. Considering total error rates, it is seen that midi anchor test brings about higher equating errors than mini anchor test in all conditions. The reason behind this could be the fact that equating slope parameter is obtained from difficulty parameter of anchor items'standard deviation ratio in test forms.

In the case of equating intercept parameter; midi anchor test gives smaller negative biased estimates except for a few conditions. As for positive biased estimates, midi and mini test yield close bias and sometimes the former gives smaller bias under two conditions when ability distribution standard deviation is identical. Nevertheless, it varies for the sample of 2000 examinees when mean and standard deviation of ability distribution differ. Then, mini anchor test generates lower levels of bias. Apart from that, the highest equating error is recorded when mean of ability distribution is differential (0.5,1). In the other two ability distribution cases, equating errors are found to be close. For equating intercept parameter, equating errors are found to be equal or smaller in midi anchor test. Equating intercept parameter varies depending on average of the parameter *b* and equating slope parameter. Since equating intercept parameter is dependant on equating slope parameter, error and bias are seen to be affected from it even if average of the parameter *b* remains unchanged in the case of midi anchor test. In relation with slope parameter; while midi test gives higher error rates, midi anchor test is seen to bring smaller error rates in the equating intercept parameter indirectly affected from the parameter *b*.

For equated scores, bias values are found to vary depending on sample size. In samples of 500, mini anchor test results smaller bias; but in samples of 2000, midi anchor test supplies smaller or equal bias to mini test. In another sample, 250, varying distribution of ability is seen to affect bias values. When distribution of ability varies between groups, midi anchor test is found to have better results. In relation with equating error, midi anchor test often lets smaller error rates than mini anchor test except under several conditions.

According to ANOVA results, difference of difficulty between forms is found to generate a significant effect only in relation with interaction effects in equated scores. In relation with equating

coefficients; it is noted that increased difference of difficulty between forms hardly affects equating error, and leads to increased equating error in the case of equated scores. In addition, sample size is found to have a significant effect for equated scores and equating intercept parameter. As sample size increases, error decreases. This finding seems to be in conformity with Cui (2006),Hanson and Beguin (2002), Kim and Cohen (2002),Norman-Dvorak (2009), Speron (2009) and Kim (2014).

In general, midi anchor test is potent of giving as good results as mini anchor test and even the former gives better results in some circumstances. It seems to support findings by Sinharay and Holland (2007), Liu et al. (2011), Kim (2014) and Fitzpatrick and Skorupski (2016). In their study comparing the chain and post-stratification equipercentile equating methods by using data generated on the basis of 2PLM and multidimensional IRT, Sinharay and Holland (2007) found out that midi and semi-midi anchor test yields as good results as mini anchor test in internal anchor test. Likewise, Fitzpatrick and Skorupski (2016) reported in their study with 3PLM-compatible data and IRT true-score equating methods that midi anchor test supplies as good results as mini anchor test in internal anchor test. Also in their comparative study on chain equating methods, Tucker and frequency estimate equating methods by using true data, Liu et al (2011) found out that midi anchor test provides better results than mini anchor test considering standard and total error in internal anchor test. In a study by Kim (2014) on 3PLM-compatible data using Kernel equating, frequency estimate, modified frequency estimate, equipercentile equating and chain equating methods, midi anchor test noted better results in relation with bias and equating error. It can be inferred from the mentioned studies that smaller variability of anchor item difficulty distribution than total test leads to smaller error rates (bias and equating). The reason is that midi anchor test contains easier items, which causes more examinees to answer the test correctly and thus equating results increase accuracy and precision.

Anchor item difficulty distribution is found to have an effect on ability distribution among groups. In all ability distributions, midi anchor test is found to yield better results than mini anchor test. Yet, insimilar ability distribution between groups, mini and midi anchor tests yield quite close results. On the other hand, the difference between obtained values becomes even larger (except in one condition) as distribution of ability becomes more differential. It can be said that this result is consistent with Fitzpatrick and Skorupski (2016) and Liu et al (2011a, b). However, Liu et al (2011) noted that mini and midi tests deliver equal results in the case of similar/identical ability distribution between groups, while Fitzpatrick and Skorupski (2016) pointed out that mini anchor test yields better results than midi anchor test in few conditions only. Furthermore, in the event of larger difference between groups, results are found to be more biased in our study. It seems to be in compliance with Kim (2014) and Power et al (2011). In both studies, different distribution of ability between groups results in increased errors. Such effect could be, among other reasons, because the group with higher levels of talent could answer correctly the items with medium level of difficulty, while the group with lower levels of talent answered the same items wrong. So if examinee groups show different ability distribution, midi anchor test seems to be preferable.

Sinharay and Holland (2007) stated that developing of anchor tests with restrictive characteristics would prevent developing tests with required statistical characteristics. The main test and the anchor test must have a similar scope and identical statistical characteristics. However, since mini anchor test entails too easy or too difficult items as well, it seems too hard for test developers to meet this requirement. Liu et al (2011) argued that statistically unsatisfactory items are also added to the test for the sake of this necessity. Apart from that, Fitzpatrick and Skorupski (2016) stated that item pools involve less items which are too easy or too difficult, and dispute could rise between content and statistical requirements in developing mini anchor tests. Hence, pressure would be placed to choose easy and difficult items for preparing a mini anchor test, which could have a negative impact on test content. As in previous studies, our study puts forth that anchor tests with smaller variability end up with better results than conventional mini anchor test in a number of conditions. So, it can be suggested that test developers could be more flexible in developing anchor tests.

In this study, mean-sigma method was used as an IRT-based method and data compatible with Rasch model. For future research, different calibration methods could be used to examine the effect of anchor item difficulty distribution. This study was conducted with simulated data. It is suggested to repeat the study with real data in the future to compare results with present study. In addition, bias and total errors were used as evaluation criteria in this study. Future studies could focus on comparing test equating effects of mini and midi anchor tests by using group invariance indices, first-order-second order equity and DTM indices.

References

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.

Antal, J.,Proctor, T. P. &Melican, G.C., (2014). The effect of anchor test construction on scaledrift. *Applied Measurement in Education, 27*: 159–172, 2014.

Budescu, D. (1985). Efficiency of linear equating as a function of the length of the anchor test. *Journal of Educational Measurement, 22*(1), 13-20.

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

Cook, L. L.,&Eignor, D. R. (1991). An NCME instructional module on IRT equating methods. *Educational Measurement: Issues and Practice, 10*(3), 37-45.

Cui, Z. (2006). *Two new alternative smoothing methods in equating: The cubic B-spline presmoothing methodand the direct presmoothing method*. (Unpublished doctoral dissertation). University of Iowa, Iowa City, IA.

Embretson, S. E. & Reise, S. (2000). *Item response theory for psychologists.* New Jersey: Lawrence Erlbaum Associates.

Fitzpatrick, J.,& Skorupski, W. P. (2016). Equating with midi tests using IRT. *Journal of Educational Measurement, 53*(2), 172-189.

Hagge, S. L. (2010). *The impact of equating method and format representation of anchor items on the adequacy of mixed-format test equating using nonequivalent groups* (Unpublished Doctoral Dissertation). University of Iowa. Iowa city.

Hambleton, R. K. & Swaminathan, H. (1985). *Item response theory: principles and applications.* Boston: Academic Puslishers Group.

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

Hanson, B. A. &Beguin, A. A. (2002). Obtaining a anchor scale for item response theory item parameters using separate versus concurrent estimation in the anchor-item equating design. *Applied Psychological Measurement, 26*(1), 3–24.

Harris, D. J.,& Crouse, J. D. (1993). A study of criteria used in equating. *Applied Measurement in Education, 6,* 195–240.

Niyazi Karasar. (2007). Bilimsel Araştırma Yöntemi. Nobel Yayın Dağıtım, Ankara.

Kim, S.-H.,&Cohen, A. S. (2002). A comparison of linking and concurrent calibration under the graded response model. *Applied Psychological Measurement, 26*, 25-41.

Kim, H.Y., (2014). *A comparison of smoothing methods for the anchor item nonequivalent groups design*. (Unpublished Doctoral Dissertation). University of Iowa. Iowa city.

Kolen, M. J. (1988). An NCME instructional module on traditional equating methodology. *Educational Measurement: Issues and Practice, 7*, 29-36.

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

Kolen, M. J. (2007). Data collection designs and linking procedures. *Linking and aligning scores and scales* (pp. 31-55). New York: Springer

Liu, J.,Sinharay, S., Holland, P. W., Curley, E., & Feigenbaum, M. (2011a). Test score equating using a mini-version anchor and a midi anchor: A case studyusing SAT data. *Journal of Educational Measurement, 48*, 361–379.

Liu, J.,Sinharay, S., Holland, P., Feigenbaum, M., & Curley, E. (2011b). Observed score equating using a mini-version anchor and an anchor with less spread of difficulty: A comparison study. *Educational and Psychological Measurement, 71*, 346–361.

Livingston, S. A. (2004). *Equating test scores (without IRT)*. Princeton, NJ: Educational Testing Service.

Nozawa, Y. (2008). *Comparison of parametric and nonparametric IRT equating methods under the anchor-item nonequivalent groups design*. (Unpublished doctoral dissertation). University of Iowa. Iowa city.

Norman-Dvorak, R. L. (2009). A comparison of kernel equating to the test characteristic curve method. (Unpublished doctorate thesis), University of Nebraska, Lincoln, Nebraska.

Petersen, N.S.,Kolen, M. J., & Hoover, H. D. (1989). Scaling, norming and equating. In R.L. Linn (Ed.), *Educational Measurement* (3rd ed.) (pp. 221-262). New York: American Council on Education.

Rizopoulos, D. (2015). *Package 'ltm'*. [https://cran.r-project.org/web/packages/ltm/ltm.pdf, Erişim tarihi: Ekim 2016.]

Sinharay, S.,Haberman, S., Holland, P., &Lewis, C. (2012). *A note on the choice of an anchor test in equating.* ETS Research Report RR-12-14. Princeton, NJ: Educational Testing Service.

Sinharay, S.,&Holland, P. (2006). The correlation between the scores of a test and an anchor test. ETS Research Report RR-06-04. Princeton, NJ: Educational Testing Service.

Sinharay, S.,&Holland, P. W. (2007). Is it necessary to make anchor tests mini-versions of

the tests being equated or can some restrictions be relaxed? *Journal of Educational Measurement, 44*, 249–275.

Speron, E. (2009). *A comparison of metric linking procedures in Item Response Theory*. (Unpublished doctorate thesis), University of Illinois, Chicago, Illinois.

Tong, Y.,&Kolen, M. (2005). Assessing equating results on different equating criteria. *Applied Psychological Measurement, 29* (6), 418-432.

Zhu, W. (1998). Test equating: What, why, how?.*Research quarterly for exercise and sport*, *69*(1), 11-23.