ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2017; 8(4);421-434

Investigation of the Performance of Multidimensional Equating Procedures for Common-Item Nonequivalent Groups Design*

Çok Boyutlu Eşitleme Yöntemlerinin Eşdeğer Olmayan Gruplarda Ortak Madde Deseni için Performanslarının İncelenmesi

Burcu ATAR ** Gonca YEŞİLTAŞ ***

Abstract
In this study, the performance of the multidimensional extentions of Stocking-Lord, mean/mean, and mean/sigma equating procedures under common-item nonequivalent groups design was investigated. The performance of those three equating procedures was examined under the combination of various conditions including sample size, ability distribution, correlation between two dimensions, and percentage of anchor items in the test. Item parameter recovery was evaluated calculating RMSE (root man squared error) and BIAS values. It was found that Stocking-Lord procedure provided the smaller RMSE and BIAS values for both item discrimination and item difficulty parameter estimates across most conditions.

Keywords: Multidimensional equating, mean/mean, mean sigma, Stocking-Lord

Öz
Bu çalışmada çok boyutlu veri için adapte edilen Stocking-Lord, ortlama/ortlama ve ortalama/sigma eşitleme yöntemlerinin performansları eşdeğer olmayan gruplarda ortak madde deseni göz önüne alınarak incelenmiştir. Bu üç eşitleme yöntemin performansları orneklem büyüklüğünün, yetenek dağılımının, boyutlar arasındaki korelasyon değerlerinin ve testteki ortak madde yüzdelerinin kombinasyonlari altinda araştırılmıştır. Madde parametre kestirimlerinin değerlendirilmesinde RMSE ve yanlılık değerleri kullanılmıştır. Bu çalışmada çoğu koşul icin hem madde ayırt edicilik parametre kestirimlerinde hem de madde güçlük parametre kestirimlerinde Stocking-Lord yönteminin diger iki yönteme gore daha kucuk RMSE ve yanlılık değerleri verdiği bulunmuştur.

Anahtar Kelimeler: Çok boyutlu eşitleme, ortalama/ortalama, ortalama/sigma, Stocking-Lord

INTRODUCTION

The main reason to administer tests under standardized conditions is to assess the abilities of examinees fairly and objectively. Scores obtained from the large-scale standardized achievement tests administered in the field of education are sometimes used in important decisions such as selecting students to place into educational programs or institutes based on their abilities. These tests are administered once or more than once within a year. A new form of the test is generally used in each administration for the security purposes in those high-stakes tests. Even though the forms are developed to measure the same construct, they may exhibit differences in their statistical characteristics such as item difficulties and reliabilities. Those differences may give advantage to examinees who take the easier and more reliable form of the test. When the different forms of a test are used, the scores obtained from one administration of the test need to be converted into the scores

* Bu araştırma 1059B191300504 kodlu TUBITAK projesi tarafından desteklenmiştir.

** Doç. Dr., Hacettepe Üniversitesi, Eğitim Fakültesi, Ankara-Türkiye, burcua@hacettepe.edu.tr, ORCID ID: orcid.org/0000-0003-3527-686X
*** Dr., Cumhuriyet Üniversitesi, Sivas-Türkiye, goncayesiltas@cumhuriyet.edu.tr, ORCID ID: orcid.org/0000-0002-7291-7083

Geliş Tarihi: 18.08.2017
DOI: 10.21031/epod.335284 *Kabul Tarihi: 21.11.2017*

obtained from the previous administration of the test in order to prevent the reflection of the statistical differences among the forms of the test into examinees' scores. Scores obtained from different forms of a test can be compared after a statistical process called equating (Kolen & Brennan, 2004). Otherwise, it is not appropriate to compare the scores obtained from different forms of the test.

There are different test equating procedures for different data collection designs. Common-item nonequivalent groups design is one of the most widely used data collection designs in equating of the different forms of the large-scale standardized achievement tests by testing companies. Among procedures based on the classical test theory, Tucker linear equating procedure (Gulliksen, 1950), Levine linear equating procedure (Levine, 1955), frequency estimation and chained equipercentile equating procedures (Angoff, 1971) can be used with common-item nonequivalent groups design. Procedures based on item response theory (IRT) can also be used with common-item nonequivalent groups design. Item response theory has prominent properties in test equating. One of the advantages of item response theory is the invariance of item and ability parameters when the model fits the data (Lord,1980). Inavariance property of item response theory has an important role in test equating especially under common-item nonequivalent groups design (Skaggs & Lissitz, 1986). The invariance property of item response theory depends on the tenability of the assumptions. One of the assumptions that should be considered in test equating studies based on item response theory is the unidimensionality assumption. However, tests in actual administrations exhibit a multidimensional structure. As a result, unidimensionality assumption is mostly violated in many testing situations (Li & Lissitz, 2000). Depending on the degree of the violation of the assumption and depending on the conditions as sample size, ability distribution, number of anchor item in the test, and so on, the performace of the equating procedure will be effected. Investigating the robustness of those procedures to the violation of unidimensionality assumption is essential. (Camilli, Wang & Fesq, 1995; De Champlain, 1996; Li & Lissitz, 2000; Oshima, Davey & Lee, 2000). When the unidimensionality assumption is not met, procedures based on multidimensional item response theory (MIRT) might be conducted. It is possible with MIRT to model the interaction of items that can discriminate between different ability levels and examinees that have different proficiencies on those levels (Ackerman, 1994). When the forms of a test exhibit multidimensional structure, MIRT equating can be considered. In MIRT equating, the accuracy of parameter estimates after equating process is critical to address (Li & Lissitz, 2000). The performance of procedures based on MIRT should be investigated under different conditions that are similar to actual testing conditions (Yao & Boughton, 2009). By this way, for different test conditions, a practical equating procedure that provides accurate estimates and minimun equating error can be determined.

There are several publications on multidimensional equating/linking (Hirsch, 1989; De Champlain, 1996; Bolt 1999; Li & Lissitz, 2000; Yao & Boughton, 2009; Yao, 2011; Eser & Gelbal, 2015). In some of those studies real test data were used and in the others simulated data was considered. De Champlain (1996) examined the equating results of unidimensional IRT true-score equating procedure on different subgroups of a two-dimensional real test data. Data used in De Champlain's analyses is the one obtained from the administration of two forms of the Law School Admission Test (LSAT). When the equating functions obtained from subgroups are compared with the ones obtained from the whole group, it was seen that the differences along the scale are small although those differences increases toward the lower end of the scale. Bolt (1999) simulated two-dimensional data using the parameter estimates obtained from two forms of the LSAT data. He compared the performance of unidimensional IRT true-score equating procedure with the performance of unidimensional linear and equipercentile equating procedures under different levels of correlation between dimensions. As a result of that study, it was found that IRT true-score equating procedure performs as well as other 2 conventional procedures when the correlation between dimensions is higher. It was also found that IRT true-score equating procedure performs better than the others when the correlation between dimensions is lower. Yao and Boughton (2009) examined the linking accuracy of test response function procedure under multidimensional perspective for tests including both dichotomously and polytomously scored items. In their study, they used simulated two-dimensional data under different conditons of population distribution, anchor set length, and item structure. They found that parameter recovery was good across all conditons with a well chosen anchor set. Yao (2011) investigated the linking accuracy of multidimensional test response function linking procedure for dichotomously scored items using a simulated data with five-dimension. Sample size, population distribution, and anchor set length were varying conditions in that study. It was found that overall score and domain score recovery was good even for condition with the smallest anchor set. In their study, Eser and Gelbal (2015) examined the performance of multidimesional item response theory model parameter estimates of two and three dimensional tests under the combination of different levels of sample size and test length conditions. They found that the parameters were estimated more accurately as the sample size and test length increased.

Considering the multidimensional structure of many testing data, it is critical to investigate the performance of different multidimensional equating/linking procedures in depth and observe which procedure performs better under different conditions. Since it is possible to encounter with various types of data under different settings in real testing, it is important to conduct studies with simulated data under conditions that are close to real testing situations. In this study, sample size, ability distibution, correlation between two dimesions, and percentage of anchor items in a test with 40 dichotomously scored items are considered as varying conditions that are close to real testing situations to compare the performance of multidimensional extensions of Stocking-Lord, mean/mean, and mean/sigma equating procedures under common-item nonequivalent groups design.

Purpose of the Study

The purpose of this study was to investigate the performance of three multidimensional equating procedures in the recovery of item and ability parameter estimates under various simulation conditions. Different simulation conditions were formed with the combination of sample size, ability distribution, correlation between dimensions, and the percentage of anchor items in the test. The equating procedures compared in this study under common-item nonequivalent groups design were the extentions of unidimenisonal equating procedures – Stocking-Lord, mean/mean, and mean/sigma.

METHOD

Simulated data under common-item nonequivalent groups design was used in this study. For the purpose of the study, examinee responses to a 40 dichotomously scored items in a two-dimensional test were generated based on the item parameter estimates from the study of Yao & Boughton (2009). As the structure of the data, some of the items loaded on a single dimension (simple structure), some of the items loaded on both dimensions (complex structure).

SimuMIRT program (Yao, 2003) was used to generate response data under various conditions. In their study, Li & Lissitz (2000) found that a sample size of 2000 with 20 anchor items from a 40-item test was adequate for the multidimensional test response function equating procedure. Factors manipulated and factors held constant were adapted considering real testing situations in this study. Sample size and number of anchor items in the test were two of the factors that were manipulated. Two levels of sample size (1000 and 2000) and four levels of percentage of anchor items in the test (15%, 30%, 60%, and 100%) were considered. In addition to sample size and percentage of anchor items, ability distribution and correlation between dimensions were manipulated. Two levels of ability distribution (for a base of comparison, multivariate normal distribution with mean of 0 and standard deviation of 1 in both dimensions was generated; for other level, mean of -0.5 and 0.5 with standard deviation of 1 on both dimensions were generated) and three levels of correlation between two dimensions (0, 0.5 and 0.8) were taken into account. As a result of the combination of those factors, 48 simulation conditions were generated. Simulation conditions for each level of percentage of anchor items in the test were shown in Table 1. Each simulation condition was replicated 20 times. The computer softwares used here for data simulation, parameter estimation, and linking purposes are softwares developed for data simulation, parameter estimation, and linking. Linking software has to be run seperately for each analysis as opposed to other softwares. In addition, each output must be examined individually to obtain the essential values. Because of these reasons, if the number of replications increases, it requires longer time to complete the analysis. The number of replications has been limited by Yao and Boughton (2009) as well as by Yue and Hongyun (2013). Test length and the number of dimensions were the factors that were held constant.

BMIRT program (Yao, 2003) was used to estimate the item and ability parameters of the response data under compensatory multidimensional three-parameter (M-3PL) IRT model. BMIRT was found to produce accurate parameter estimates under the M-3PL model (Yao & Boughton, 2007; Yao & Schwarz, 2006).

Equating was conducted using multidimensional Stocking-Lord, mean/mean, and mean/sigma equating procedures. LinkMIRT program (Yao, 2004) was used for equating in all simulation conditions.

To evaluate the item parameter recovery, root mean square error (RMSE), and bias (BIAS) values were calculated using the following formulas:

$$RMSE(\hat{\tau}) = \sqrt{\frac{\sum_{r}^{R}(\hat{\tau}_r - \tau)^2}{R}} \text{ and}$$

$$Bias(\hat{\tau}) = \frac{\sum_{r}^{R}\hat{\tau}_r}{R} - \tau$$

where τ is the true value of the parameter, $\hat{\tau}_r$ is the estimated value of the parameter at the r^{th} replication, and R is the number of replications.

Table 1. Simulation Conditions

Condition	Sample Size	Mean1	Mean2	Var-Cov Matrix
C1	1000	0	0	1,0,0,1
C2	2000	0	0	1,0,0,2
C3	1000	0	0	1,0.5,0.5,1
C4	2000	0	0	1,0.5,0.5,2
C5	1000	0	0	1,0.8,0.8,1
C6	2000	0	0	1,0.8,0.8,2
C7	1000	0.5	-0.5	1,0,0,1
C8	2000	0.5	-0.5	1,0,0,2
C9	1000	0.5	-0.5	1,0.5,0.5,1
C10	2000	0.5	-0.5	1,0.5,0.5,2
C11	1000	0.5	-0.5	1,0.8,0.8,1
C12	2000	0.5	-0.5	1,0.8,0.8,2

Note. These twelve conditions were repeated for each level of percentage of anchor item

RESULTS

Item Parameter Recovery

RMSE and BIAS values of the first item discrimination parameter related to the first dimension estimates calculated for the three equating procedures across all conditions are given in Table 2 and Table 3, respectively. RMSE and BIAS values of the first item discrimination parameter estimates for the levels of the percentage of anchor items are shown in Figure 1 and Figure 2, respectively for each equating procedure

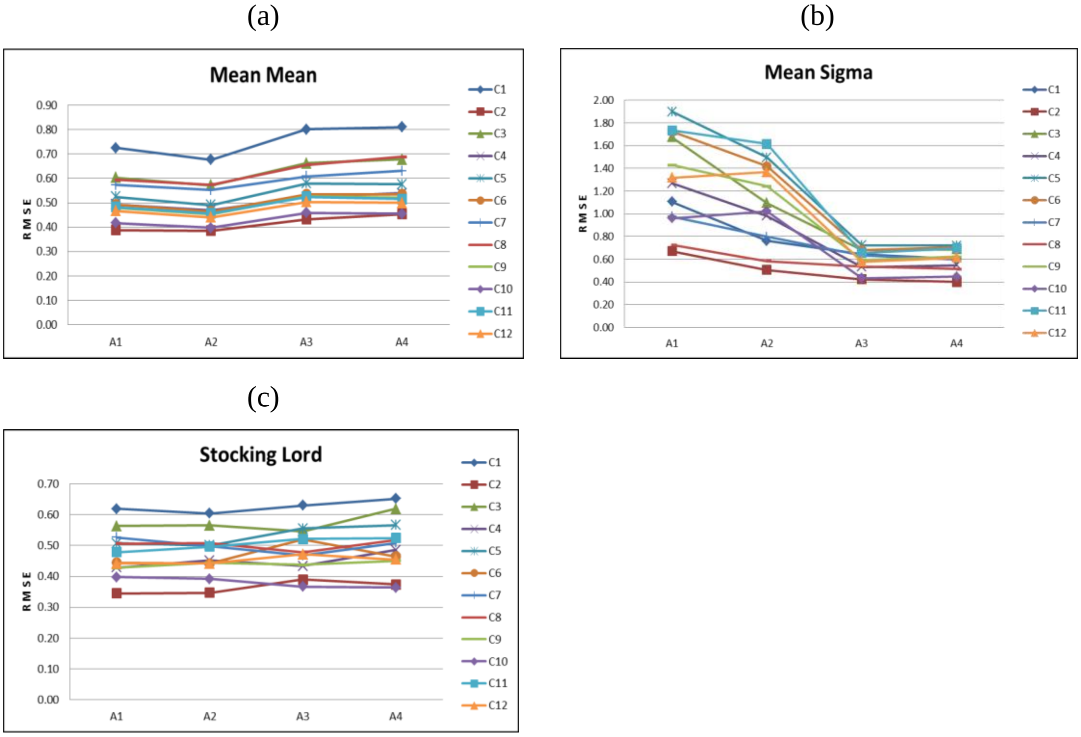
Table 2. RMSE for the First Item Discrimination Parameter

	15%			30%			60%			100%		
	MM	**MS**	**SL**	**MM**	**MS**	**SL**	**MM**	**MS**	**SL**	**MM**	**MS**	**SL**
C1	0.73	1.11	0.62	0.68	0.76	0.60	0.80	0.65	0.63	0.81	0.60	0.65
C2	0.39	0.67	0.35	0.39	0.50	0.35	0.43	0.42	0.39	0.45	0.40	0.37
C3	0.60	1.67	0.56	0.57	1.10	0.57	0.66	0.68	0.55	0.68	0.69	0.62
C4	0.49	1.27	0.43	0.47	0.98	0.45	0.52	0.53	0.43	0.54	0.55	0.49
C5	0.52	1.90	0.51	0.49	1.49	0.50	0.58	0.72	0.55	0.58	0.72	0.57
C6	0.49	1.72	0.44	0.46	1.42	0.44	0.54	0.68	0.52	0.54	0.70	0.46
C7	0.57	0.98	0.53	0.55	0.80	0.50	0.61	0.63	0.47	0.63	0.59	0.51
C8	0.59	0.72	0.51	0.57	0.58	0.51	0.66	0.53	0.48	0.69	0.51	0.52
C9	0.48	1.43	0.43	0.45	1.24	0.44	0.53	0.59	0.44	0.52	0.62	0.45
C10	0.42	0.96	0.40	0.40	1.02	0.39	0.46	0.43	0.37	0.46	0.45	0.36
C11	0.48	1.73	0.48	0.50	1.61	0.50	0.52	0.65	0.52	0.52	0.70	0.52
C12	0.47	1.32	0.44	0.44	1.37	0.44	0.50	0.58	0.47	0.50	0.61	0.45

Table 3. BIAS for the First Item Discrimination Parameter

	15%			30%			60%			100%		
	MM	**MS**	**SL**	**MM**	**MS**	**SL**	**MM**	**MS**	**SL**	**MM**	**MS**	**SL**
C1	0.53	0.76	0.42	0.50	0.44	0.40	0.59	0.44	0.39	0.59	0.43	0.40
C2	0.25	0.48	0.20	0.25	0.31	0.16	0.26	0.27	0.19	0.27	0.25	0.16
C3	0.50	1.38	0.41	0.47	0.88	0.40	0.55	0.48	0.37	0.56	0.50	0.40
C4	0.40	1.10	0.29	0.38	0.81	0.28	0.43	0.37	0.27	0.44	0.38	0.27
C5	0.45	1.72	0.40	0.42	1.35	0.38	0.51	0.47	0.39	0.51	0.50	0.38
C6	0.44	1.61	0.37	0.42	1.33	0.37	0.49	0.45	0.36	0.49	0.49	0.34
C7	0.39	0.66	0.30	0.38	0.47	0.26	0.42	0.36	0.26	0.43	0.34	0.26
C8	0.41	0.46	0.31	0.40	0.35	0.29	0.45	0.37	0.31	0.47	0.36	0.32
C9	0.40	1.21	0.33	0.38	1.06	0.31	0.44	0.38	0.32	0.44	0.41	0.30
C10	0.34	0.83	0.27	0.33	0.89	0.23	0.37	0.31	0.24	0.37	0.32	0.22
C11	0.41	1.56	0.38	0.39	1.42	0.35	0.46	0.48	0.37	0.45	0.52	0.36
C12	0.42	1.19	0.37	0.39	1.25	0.35	0.46	0.43	0.35	0.46	0.47	0.34

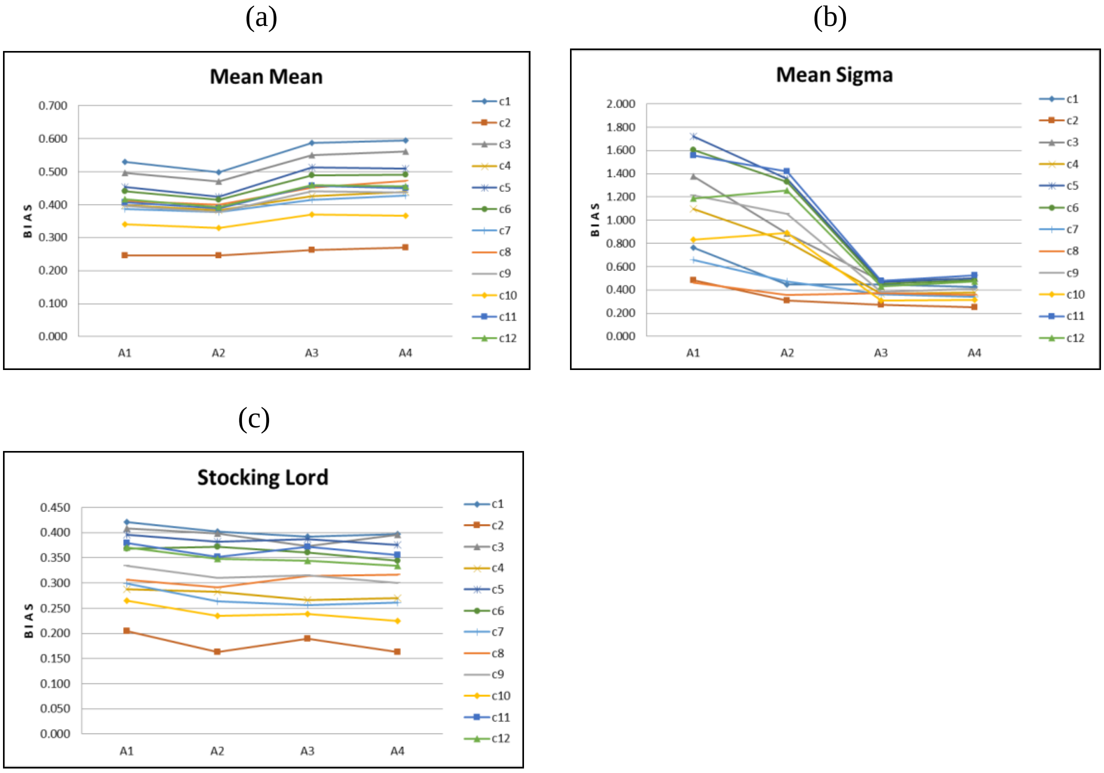
Based on the RMSE values in Table 2 and the BIAS values in Table 3 for the first discrimination parameter, it can be said that Stocking-Lord procedure produced smaller RMSE and BIAS values than mean/mean and mean/sigma procedures under all conditions. Percentage of anchor items in the test factor affected the results of mean/sigma procedure the most. The RMSE and BIAS values calculated for mean/sigma procedure were much higher than the ones calculated for mean/mean and Stocking-Lord procedures when the percentage of anchor items were 15% and 30%. The RMSE values were decreased for mean/sigma procedure as the percentage of anchor items were incerased. Under the 15% and 30% anchor items conditions, the RMSE values for mean/sigma procedure were the smallest when the sample size was 2000 and the correlation between dimensions was 0.


(a)

(b)

(c)

Figure 1. RMSE Values of the First Item Discrimination Parameter for Three Equating Procedures by the Percentage of Anchor Items Under All Conditions


(a)

(b)

(c)

Figure 2. BIAS Values of the First Item Discrimination Parameter for Three Equating Procedures by the Percentage of Anchor Items Under All Conditions

It can be seen in Figure 1 that the percentage of anchor items factor did not effect the results of mean/mean and Stocking-Lord procedures for the first item discrimination parameter much. The most effective factor for those two procedures was the sample size. Mean/mean and Stocking-Lord procedures produced smaller RMSE values when the sample size was larger. Sample size was also an effective factor for mean/sigma method.

RMSE and BIAS values of the second item discrimination parameter related to the second dimension estimates calculated for the three equating procedures across all conditions are given in Table 4 and Table 5, respectively. RMSE and BIAS values of the second item discrimination parameter estimates for the levels of the percentage of anchor items are shown in Figure 3 and Figure 4, respectively for each equating procedure.

Table 4. RMSE for the Second Item Discrimination Parameter

	15%			30%			60%			100%		
	MM	**MS**	**SL**	**MM**	**MS**	**SL**	**MM**	**MS**	**SL**	**MM**	**MS**	**SL**
C1	0.74	0.94	0.70	0.87	0.76	0.78	0.82	0.80	0.85	0.88	0.77	0.93
C2	0.45	0.40	0.37	0.57	0.54	0.51	0.55	0.68	0.55	0.59	0.64	0.55
C3	0.66	1.27	0.73	0.78	0.85	0.81	0.74	0.74	0.79	0.81	0.72	0.91
C4	0.53	0.88	0.55	0.60	0.68	0.61	0.58	0.70	0.63	0.63	0.66	0.72
C5	0.61	1.35	0.65	0.67	1.01	0.65	0.63	0.82	0.69	0.67	0.77	0.74
C6	0.54	1.03	0.57	0.60	0.83	0.56	0.57	0.75	0.67	0.60	0.72	0.64
C7	0.69	0.91	0.73	0.83	0.76	0.75	0.80	0.81	0.61	0.88	0.77	0.83
C8	0.73	0.77	0.74	0.89	0.68	0.77	0.87	0.80	0.68	0.95	0.77	0.88
C9	0.62	1.10	0.64	0.71	0.82	0.70	0.65	0.73	0.59	0.72	0.71	0.66
C10	0.52	0.73	0.54	0.61	0.69	0.57	0.57	0.70	0.47	0.62	0.67	0.51
C11	0.62	1.20	0.66	0.71	0.98	0.74	0.66	0.77	0.72	0.72	0.74	0.79
C12	0.60	0.91	0.59	0.66	0.84	0.67	0.63	0.73	0.62	0.68	0.70	0.63

Table 5. BIAS for the Second Item Discrimination Parameter

	15%			30%			60%			100%		
	MM	**MS**	**SL**	**MM**	**MS**	**SL**	**MM**	**MS**	**SL**	**MM**	**MS**	**SL**
C1	0.48	0.52	0.50	0.57	0.36	0.51	0.53	0.67	0.58	0.58	0.65	0.55
C2	0.31	0.36	0.32	0.34	0.27	0.31	0.32	0.58	0.34	0.34	0.53	0.98
C3	0.51	0.90	0.54	0.57	0.58	0.54	0.54	0.55	0.54	0.60	0.53	0.54
C4	0.39	0.63	0.38	0.43	0.44	0.38	0.41	0.61	0.39	0.44	0.55	0.36
C5	0.49	1.10	0.50	0.53	0.81	0.47	0.50	0.57	0.45	0.54	0.54	0.41
C6	0.45	0.85	0.43	0.49	0.68	0.43	0.47	0.62	0.40	0.51	0.57	0.37
C7	0.47	0.53	0.47	0.54	0.40	0.47	0.51	0.65	0.46	0.57	0.64	0.49
C8	0.51	0.46	0.54	0.60	0.43	0.51	0.58	0.71	0.51	0.64	0.69	0.56
C9	0.48	0.80	0.46	0.54	0.58	0.47	0.50	0.60	0.44	0.56	0.60	0.42
C10	0.40	0.56	0.37	0.46	0.52	0.38	0.43	0.65	0.37	0.47	0.61	0.33
C11	0.52	1.01	0.52	0.58	0.78	0.52	0.54	0.57	0.47	0.60	0.59	0.48
C12	0.52	0.73	0.48	0.55	0.69	0.49	0.53	0.65	0.45	0.58	0.62	0.41

Based on the RMSE values in Table 4 and the BIAS values in Table 5 for the second discrimination parameter,it can be said that mean/mean, mean/sigma, and Stocking-Lord procedures generally produced close RMSE and BIAS values under manipulated simulation conditions. When the sample size is smaller and the correlation between two dimension is higher, mean/sigma procedure had higher RMSE and BIAS values than other two procedures.

As seen in Figure 3 and Figure 4, the RMSE and the BIAS values tend to decrease as the percentage of anchor items increases. However, the larger sample size and no correlation between dimensions condition produced the smallest RMSE and BIAS values for mean/sigma procedure. The RMSE values for mean/mean and Stocking-Lord procedures tend to increase as the percentage of anchor items increases. The BIAS values for mean/mean and Stocking-Lord procedures are more stable across the levels of percentage of anchor items. As was the case for the first discrimination parameter, the larger sample size conditions produced smaller RMSE and BIAS values for the second discrimination parameter for three methods.

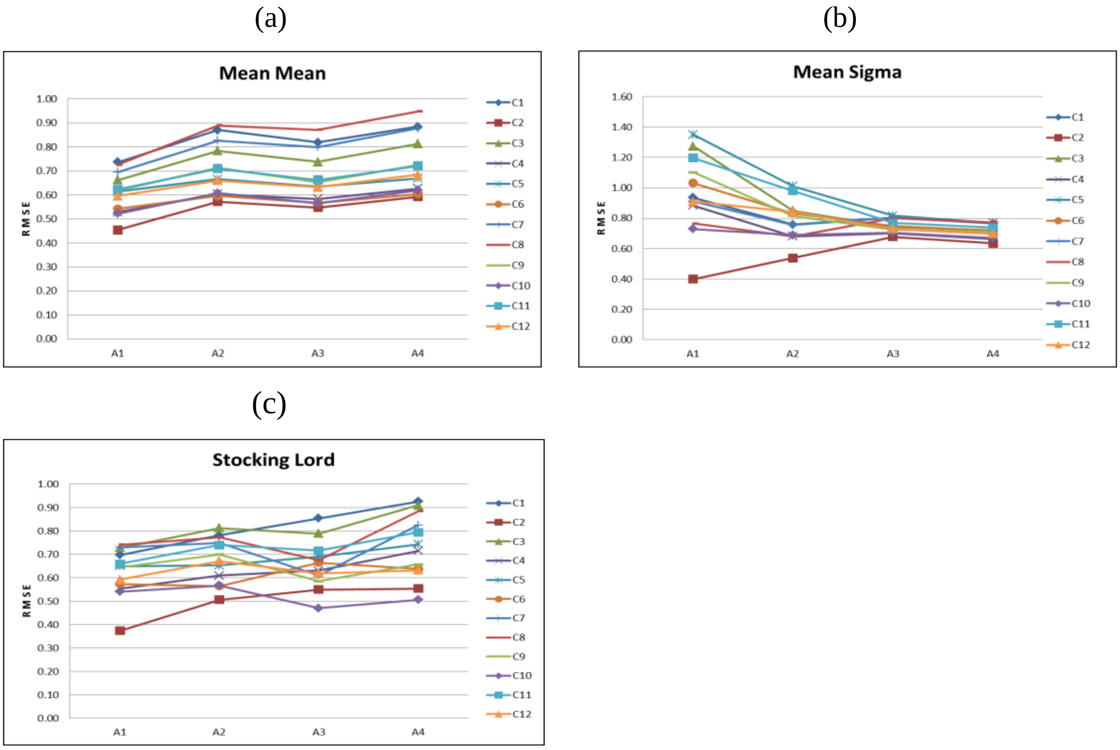


Figure 3. RMSE Values of the Second Item Discrimination Parameter for Three Equating Procedures by the Percentage of Anchor Items Under All Conditions

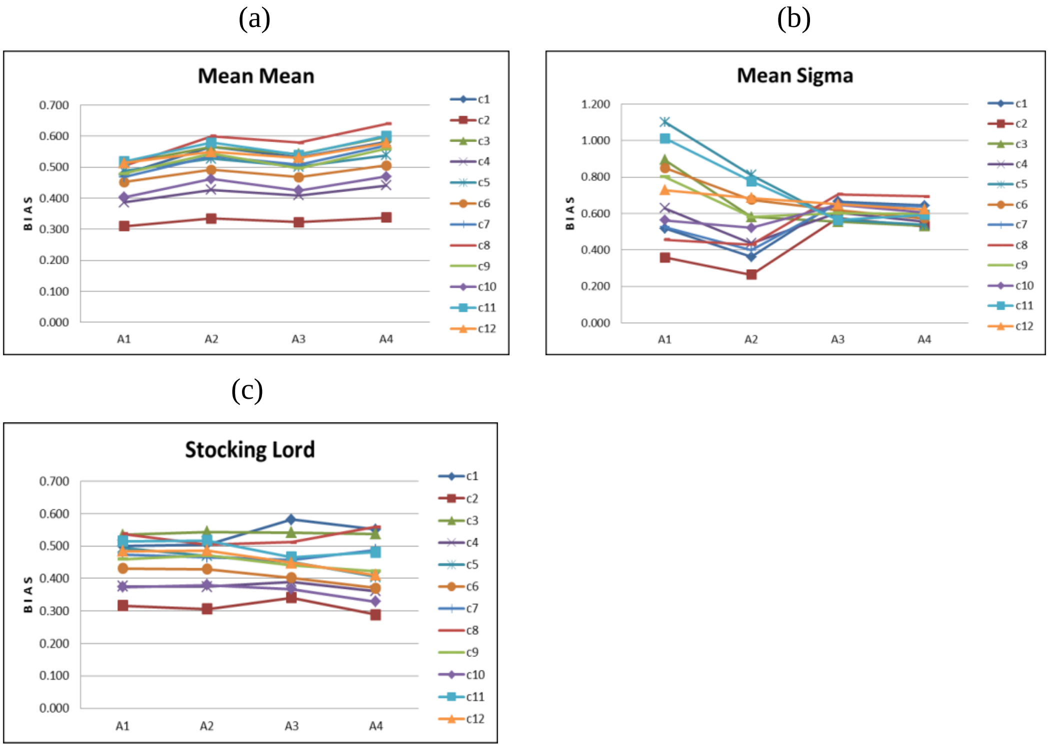


Figure 4. BIAS Values of the Second Item Discrimination Parameter for Three Equating Procedures by the Percentage of Anchor Items Under All Conditions

RMSE and BIAS values of the item difficulty parameter estimates calculated for the three equating procedures across all conditions are given in Table 6 and Table 7, respectively. RMSE and BIAS values of the item difficulty parameter estimates for the levels of the percentage of anchor items are shown in Figure 5 and Figure 6, respectively for each equating procedure.

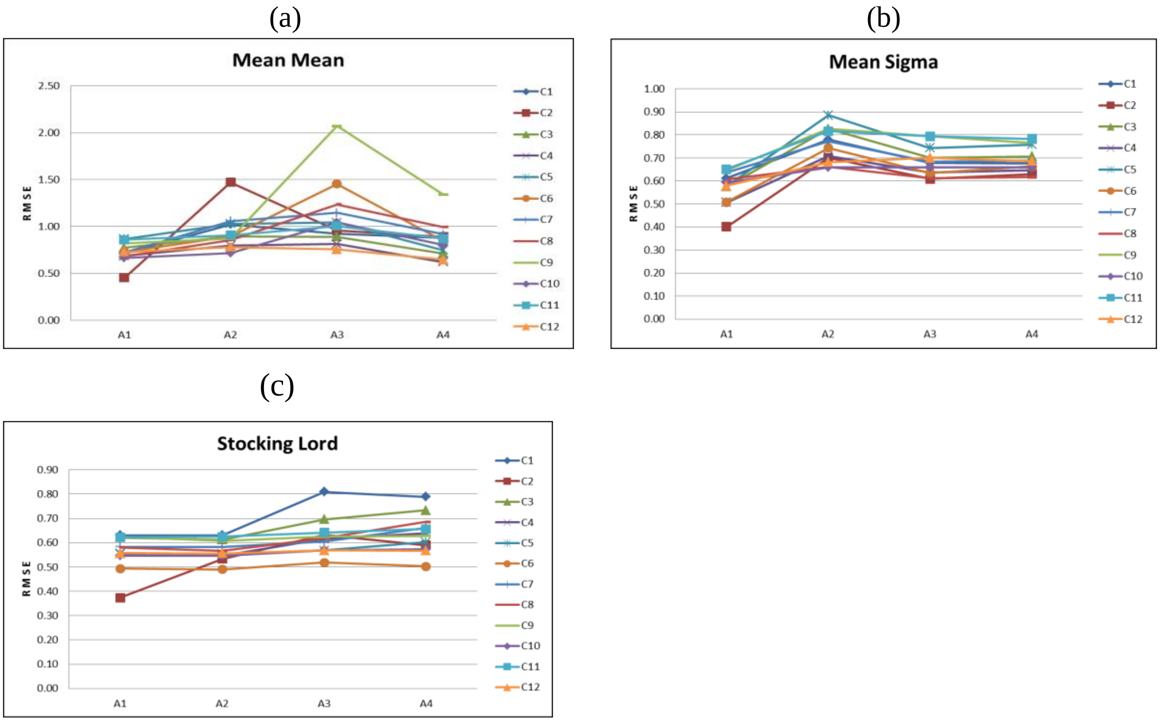
Table 6. RMSE for the Item Difficulty Parameter

	15%			**30%**			**60%**			**100%**		
	MM	**MS**	**SL**	**MM**	**MS**	**SL**	**MM**	**MS**	**SL**	**MM**	**MS**	**SL**
C1	0.71	0.61	0.63	1.02	0.78	0.63	0.92	0.68	0.81	0.87	0.68	0.79
C2	0.45	0.40	0.37	1.47	0.70	0.53	0.96	0.61	0.63	0.89	0.63	0.59
C3	0.77	0.58	0.62	0.89	0.83	0.61	0.89	0.70	0.70	0.71	0.71	0.73
C4	0.69	0.51	0.56	0.79	0.71	0.55	0.81	0.64	0.62	0.64	0.65	0.64
C5	0.87	0.58	0.56	1.03	0.88	0.55	1.04	0.74	0.57	0.75	0.76	0.60
C6	0.73	0.51	0.49	0.89	0.74	0.49	1.45	0.64	0.52	0.85	0.66	0.50
C7	0.72	0.64	0.58	1.06	0.77	0.58	1.15	0.68	0.61	0.92	0.69	0.66
C8	0.68	0.61	0.58	0.86	0.66	0.57	1.23	0.61	0.62	0.99	0.62	0.69
C9	0.82	0.65	0.62	0.87	0.83	0.61	2.07	0.79	0.63	1.34	0.77	0.63
C10	0.67	0.59	0.55	0.72	0.66	0.55	1.04	0.66	0.57	0.80	0.66	0.57
C11	0.86	0.65	0.62	0.91	0.82	0.63	1.01	0.79	0.64	0.87	0.78	0.66
C12	0.73	0.58	0.56	0.78	0.68	0.56	0.76	0.70	0.57	0.65	0.69	0.57

Table 7. BIAS for the Item Difficulty Parameter

	15%			30%			60%			100%		
	MM	**MS**	**SL**	**MM**	**MS**	**SL**	**MM**	**MS**	**SL**	**MM**	**MS**	**SL**
C1	0.57	0.50	0.54	0.53	0.65	0.53	0.55	0.60	0.59	0.53	0.61	0.58
C2	0.54	0.48	0.50	0.50	0.62	0.45	0.59	0.56	0.47	0.49	0.58	0.46
C3	0.71	0.49	0.52	0.79	0.71	0.50	0.49	0.63	0.52	0.51	0.64	0.54
C4	0.63	0.46	0.49	0.73	0.63	0.46	0.46	0.6	0.46	0.45	0.61	0.47
C5	0.81	0.49	0.47	0.98	0.79	0.46	0.47	0.68	0.46	0.52	0.71	0.46
C6	0.69	0.44	0.44	0.86	0.67	0.43	0.58	0.59	0.42	0.43	0.63	0.42
C7	0.58	0.55	0.47	0.52	0.63	0.48	0.67	0.61	0.50	0.59	0.62	0.49
C8	0.54	0.54	0.46	0.55	0.55	0.47	0.85	0.56	0.47	0.72	0.56	0.47
C9	0.73	0.57	0.52	0.77	0.71	0.51	0.63	0.72	0.54	0.58	0.71	0.54
C10	0.61	0.54	0.46	0.65	0.57	0.47	0.66	0.60	0.51	0.58	0.61	0.51
C11	0.80	0.56	0.53	0.85	0.72	0.52	0.57	0.73	0.55	0.57	0.73	0.55
C12	0.67	0.51	0.49	0.73	0.60	0.48	0.52	0.66	0.50	0.52	0.65	0.50

Based on Table 6 and Table 7, RMSE values for mean/mean procedures are higher than the other two equating procedures across all conditions. Moreover, in terms of BIAS values mean/sigma procedure has higher values than the other two equating procedures under the simulation conditions. RMSE and BIAS values of item difficulty parameter estimates incease as sample size decreases. On the other hand, mean difference and correlation between two dimensions does not have clear effect on the RMSE and BIAS values of item difficulty parameter estimates under the manipulated conditions.


(a)

(b)

(c)

Figure 5. RMSE Values of the Item Difficulty Parameter for Three Equating Procedures by the Percentage of Anchor Items Under All Conditions

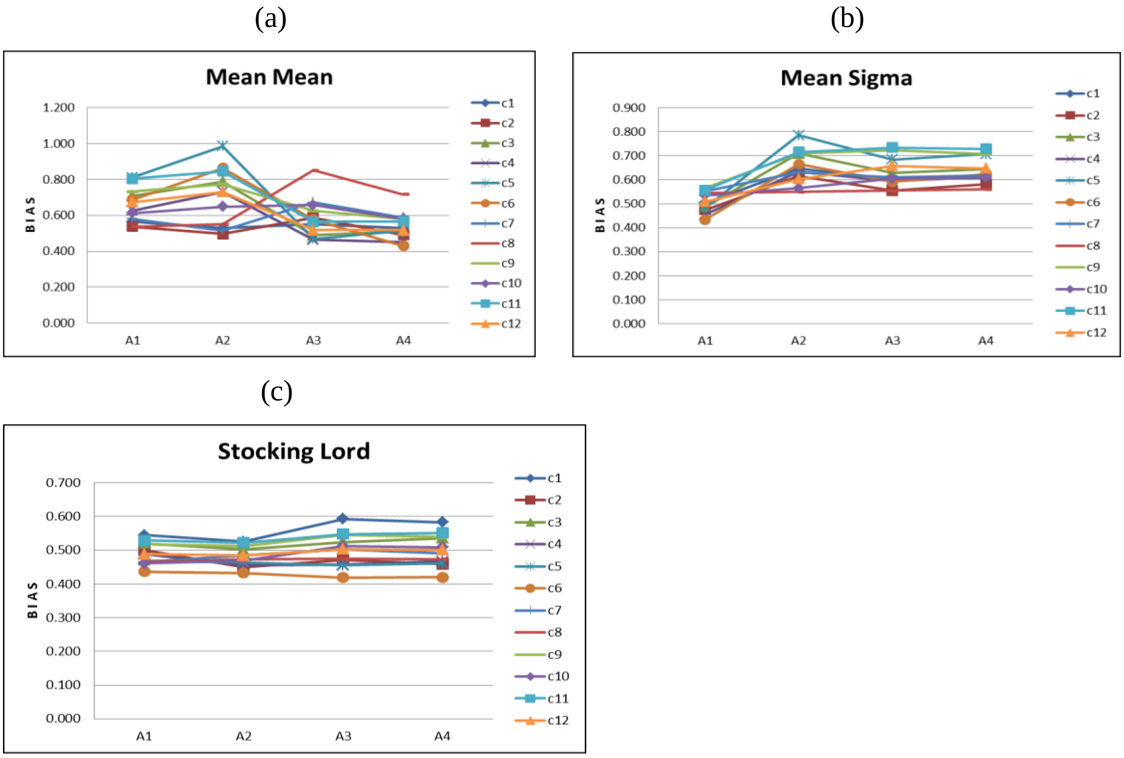


Figure 6. BIAS Values of the Item Difficulty Parameter for Three Equating Procedures by the Percentage of Anchor Items Under All Conditions

CONCLUSION and DISCUSSION

When RMSE and BIAS values of item parameter estimates were examined, it was found that the ability distribution factor did not have considerable effect on the accuracy of item discrimination and item difficulty parameter estimates for the three equating procedures. Those findings are similar to the findings of Yao & Boughton (2009). They found in their study that with a well-chosen anchor set including at least one simple structured item per dimension, parameter estimates are more accurate. On the other hand, sample size factor affected RMSE and BIAS values of item parameter estimates for the three procedures with larger sample size conditions produced smaller RMSE and BIAS values. Those findings are consistent with the findings of Eser and Gelbal (2015). They suggested in their study to run the analyses with a sample size of 2000 examinees for a two-dimensional test. Percentage of anchor items factor affected the results of mean/sigma procedure the most. RMSE and BIAS values of item discrimination parameter estimates tend to decrease as the percentage of anchor items increases for that procedure. RMSE and BIAS values of item discrimination parameter are similar for the levels of the percentage of anchor items factor for mean/mean and Stocking-Lord procedures.

It can be concluded that Stocking-Lord and mean/mean procedures provided better estimates for the item discrimination parameters while Stocking-Lord and mean/sigma procedures provided better estimates for item difficulty parameter.

When the test forms are in multidimensional structure, it is critical to use appropriate methods in the estimation of item and person parameters and equating/linlking test forms. In that sense, the performance of different methods should be investigated in depth under different conditions that are similar to real testing situations.

In this study, only three of the multidimensional equating procedures were compared for item parameter recovery. In future researches, accuracy of parameter estimates can be compared among unidimensional and multidimensional equating procedures. Conditions may vary in the test length and the number of dimensions considering the real test settings.

REFERENCES

Ackerman, T. A. (1992). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. *Journal of Educational Measurement*, *29*(1), 67-91.

Angoff, W. H. (1971). Scales, norms, and equaivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp.508-600). Washington, DC: American Council on Education. (Reprinted as W. H. Angoff, *Scales, norm, and equivalent scores*. Princeton, NJ: Educational Testing Service, 1984).

Bolt, D. M. (1999). Evaluating the effects of multidimensionality on IRT true score equating. *Applied Measurement in Education, 12*(4), 383-407.

Camilli, G., Wang, M., & Fesq, J. (1995). The effects of dimensionality on equating the law school admission test. *Journal of Educational Measurement*, *32*(1), 79-96.

Davey, T., Oshima, T. C., & Lee, K. (1996). Linking multidimensional item calibrations. *Applied Psychological Measurement*, *11*(3), 221-224.

De Champlain, A. F. (1996). The effect of multidimensionality on IRT true-score equating for subgroups of examinees. *Journal of Educational Measurement, 33*(2), 181-201.

Eser, D. Ç. & Gelbal, S. (2015). Examining parameter estimations of simple and complex structured tests with various dimensionality properties based on multidimensional item response theory. *Journal of Measurement and Evaluation in Education and Psychology*, *6*(2), 331-350.

Gulliksen, H. (1950). *Theory of mental tests*. New York: Wiley.

Hirsch, T. M. (1989). Multidimensional equating. *Journal of Educational Measurement*, *26*(4), 337-349.

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

Levine, R. (1955). Equating the score scales of alternate forms administered to samples of different ability (Research Bulletin, 55-23). Princeton, NJ: Educational Testing Service.

Li, Y. H., & Lissitz, R. W. (2000). An evaluation of the accuracy of multidimensional IRT linking. *Applied Psychological Measurement*, *24*(2), 115-138.

Livingston, S. A., Dorans, N. J., & Wright, N. K. (1990). What combination of sampling and equating methods works best? *Applied Measurement in Education*, *3*(1), 73-95.

Lord, F. M. (1980). Applications of Item Response Theory to practical testing problems. Hillsdale, NJ: Erlbaum.

Skaggs, G., & Lissitz, R. W. (1986). IRT test equating: Relevant issues and a review of recent research. *Review of Educational Research*, *56*(4), 495-529.

Yao, L. (2003). *BMIRT: Bayesian multivariate item response theory* [Computer software and manual]. Monteray, CA: CTB/McGraw Hill.

Yao, L. (2003). *SimuMIRT* [Computer software]. Monteray, CA: DMDC DoD Center.

Yao, L. (2004). *LinkMIRT: Linking of multivariate item response model* [Computer software]. Monteray, CA: DMDC DoD Center.

Yao, L., & Boughton, K. (2007). A multidimensional item response modeling approach for improving subscale proficiency estimation and classification. *Applied Psycholoical Measurement, 31*, 1-23.

Yao, L., & Boughton, K. (2009). Multidimensional linking for tests with mixed item types. *Journal of Educational Measurement*, *46*(2), 177-197.

Yao, L., & Schwarz, R. (2006). A multidimensional partial credit model with associated item and test statistics: An application to mixed format tests. *Applied Psychological Measurement*, *30*, 469-492.

Yao, L. (2011). Multidimensional linking for domain scores and overall scores for nonequivalent groups. *Applied Psychological Measurement*, *35*(1), 48-66.

Yue, L., & Hongyun, L. (2013). Comparison of MIRT linking methods for different common item designs. *Acta Psychologica Sinica*, *45*(4), 466-480.

GENİŞ ÖZET

Giriş

Testlerin standart koşullar altında uygulanmasının başlıca nedeni, adayların yeteneklerini adil ve objektif olarak değerlendirmektir. Adaylar ile ilgili değerlendirmeler önemli kararlarda kullanılır. Güvenlik açısından her bir uygulamada yeni bir test formu kullanılır. Formlar aynı yapıyı ölçmek için geliştirilmiş olsa da, madde güçlükleri ve güvenilirlikleri gibi istatistiksel özelliklerinde farklılıklar gösterebilir. Bunu önlemek için yeni testin puanları önceki uygulamada elde edilen puanlara dönüştürülmelidir. Farklı formlarından elde edilen puanlar eşitleme denilen istatistiksel bir işlemin ardından karşılaştırılabilir (Kolen & Brennan, 2004).

Farklı veri toplama desenleri için farklı test eşitleme yöntemleri vardır. Eşdeğer olmayan gruplarda ortak madde deseni, geniş ölçekli standartlaştırıolmış testlerin farklı formlarının eşitlenmesinde/bağlanmasında en yaygın kullanılan veri toplama desenlerinden birisidir.

Maddeler tepki kuramına MTKdayanan yöntemler, eşdeğer olmayan gruplarda ortak madde deseni kullanılabilir. Madde tepki kuramı eşitlemede belli özelliklere sahiptir. Madde tepki kuramını avantajlarından birisi, model-veri uyumu sağlandığında madde ve yetenek parametrelerinin değişmezliliğidir (Lord, 1980). Madde tepki kuramının değişmezlik özelliği, özellikle eşdeğer olmayan gruplarda ortak madde deseninde test eşitlemede önemli bir role sahiptir (Skaggs & Lissitz, 1986). Madde tepki kuraminin değişmezlik özelliği, varsayımların karşılanmasına bağlıdır. Madde tepki kuramına dayanan eşitleme çalışmalarında dikkate alınması gereken varsayımlardan biri, tek boyutluluk varsayımıdır. Halbuki gerçek uygulamalardaki testler çok boyutlu bir yapı sergilemektedir ve tek boyutluluk varsayımı birçok test durumunda ihlal edilmektedir (Li & Lissitz, 2000). Tek boyutluluk varsayımı karşılanmadığında, çok boyutlu madde tepki kuramına (MIRT) dayanan yöntemler uygulanabilir. Bir testin formları çok boyutlu bir yapı sergilediğinde, MIRT eşitleme düşünülebilir. Eşitleme işlemi sonrasında parametre kestirimlerinin doğruluğu, MIRT eşitleme önemlidir (Li & Lissitz, 2000). MIRT temelli işlemlerin performansı, gerçek test koşullarına benzer farklı koşullar altında araştırılmalıdır (Yao ve Boughton, 2009).

Çok boyutlu eşitleme/bağlama ile ilgili birçok yayın bulunmaktadır (Hirsch, 1989; De Champlain, 1996; Bolt 1999; Li & Lissitz, 2000; Yao ve Boughton, 2009; Yao, 2011). Bu araştırmalardan bazılarında gerçek test verileri ve bazilarinda simüle edilen veriler kullanılmış. Bolt (1999) LSAT verilerinin iki formundan elde ettiği parametre kestirimlerini kullanılarak iki boyutlu verileri simüle ettmiştir. Tek boyutlu madde tepki kuramı gerçek puan eşitlemenin performansını, boyutlar arasındaki farklı korelasyon seviyeleri altında tek boyutlu doğrusal ve eşit yüzdelikli eşitleme yöntemlerinin performansıyla karşılaştırmıştır. Bu çalışma sonucunda, boyutlar arasındaki korelasyonun daha yüksek olduğu durumlarda, MTK gerçek puan eşitleme yönteminin performansinin diğer 2 geleneksel yöntemin performansi kadar iyi olduğunu saptamıştır. Ayrıca, MTK gerçek puan eşitleme yönteminin boyutlar arasındaki korelasyonun daha düşük olduğu durumlarda diğerlerinden daha iyi performans gösterdiğini bulmustur. Yao ve Boughton (2009), test yanıt fonksiyonu yönteminin bağlantılı doğruluğunu, hem iki kategorili puanlanan hem de çok kategorilı puanlanan maddeleri içeren testler için çok boyutlu perspektifle incelemistir. Çalışmalarında, popülasyon dağılımı, ortak madde set uzunluğu ve madde yapısı farklı koşullar altında simüle edilmiş iki boyutlu verileri kullanilmisitr. Parametre iyileştirmesinin, iyi seçilmiş bir ortak madde kümesiyle tüm koşullarda iyi olduğunu bulmuşlardır. Yao (2011), beş boyutlu simüle edilmiş verileri kullanarak iki ayrı puanlamalı maddeler için bağlama yöntemini birbirine bağlayan çok boyutlu test yanıt fonksiyonunun bağlantı doğruluğu üzerinde araştırmalar yapmıştır. Bu çalışmada örneklem büyüklüğü, populasyon dağılımı ve ortak madde seti uzunluğu değişik koşullarda simule edilmistir. En küçük ortak madde seti olan koşullar için bile genel puanın ve alan puanı iyileşmesinin iyi olduğu bulunmustur. Pek çok test verisinin çok boyutlu yapısını göz önünde bulundurarak, farklı çok boyutlu eşitleme/bağlama yöntemlerinin performansını derinlemesine araştırmak ve hangi yöntemin farklı koşullar altında daha iyi performans gösterdiğini gözlemlemek çok önemlidir.

Bu çalışmanın amacı çeşitli simülasyon koşulları altında madde parametre kestirimlerinin kararlılığında üç çok boyutlu eşitleme yönteminin performansını araştırmaktır. Farkli orneklem büyüklüğü, yetenek dağılımı, boyutlar arasındaki korelasyon ve testteki ortak maddelerin yüzdesi bir araya getirilerek simülasyon koşulları oluşturulmustur. Bu çalışmada, eşdeğer olmayan gruplarda ortak madde deseni altında çok boyutlu Stocking-Lord, ortalama/ortalama ve ortalama/sigma karşılaştırılmıştır.

Yöntem

Bu çalışmada, eşdeğer olmayan gruplarda ortak madde deseni altında üretilen veriler kullanılmıştır. Çalışmanın amacı doğrultusunda, Yao ve Boughton'un (2009) çalışmasında yer alan madde parametre kestirimlerine dayanarak, iki kategorili puanlanan 40 maddeli iki boyutlu bir teste verilen yanıtlar oluşturulmuştur.

SimuMIRT programı (Yao, 2003) çeşitli koşullar altında yanıt verisi üretmek için kullanılmıştır. Li & Lissitz (2000) çalışmalarında, 20 ortak madde içeren 40 maddelik bir testten için 2000 örneklem büyüklüğünün, çok boyutlu test yanıt fonksiyonu eşitleme yöntemi için yeterli olduğunu bulmuştur. Bu çalışmada, örneklem büyüklüğü (1000 ve 2000), yetenek dağılımı ((0,0), (-0.5,0.5)), ortak madde yuzdesi (% 15,% 30,% 60 ve% 100) ve boyutlar arasindaki korelasyon (0,0.5, 0.8) manipüle edilen faktörlerdir. Her simülasyon koşulu 20 kere tekrarlanmıştır. Test uzunluğu ve boyut sayısı, sabit tutulan faktörlerdir. Parametreleri kestirmek etmek için BMIRT programı (Yao, 2003) kullanılmıştır. Eşitleme, çok boyutlu Stocking-Lord, ortalama/ortalama ve ortalama/sigma eşitleme yöntemleri kullanılarak gerçekleştirilmistir. LinkMIRT programı (Yao, 2004) tüm simülasyon koşullarında eşitleme için kullanılmıştır.

Maddelerin parametre kestirimlerimi değerlendirmek için, RMSE ve yanlılık (BIAS) değerleri hesaplanmıştır.

Sonuç ve Tartışma

Madde parametre kestirimlerinin RMSE ve BIAS değerleri incelendiğinde, yetenek dağılımı faktörünün, üç eşitleme yöntemi için madde ayırt edicilik ve madde güçlük parametreleri kestirimlerine önemli bir etkisi olmadığı bulunmuştur. Diğer yandan, örneklem büyüklüğü faktörü daha büyük örneklem büyüklüğü koşulları altında üç yöntem için de madde parametre kestirimlerinin RMSE ve BIAS değerlerini etkilerken, daha küçük RMSE ve BIAS değerleri üretmiştir. Ortak madde faktörünün yüzdesi en fazla ortalama/sigma yönteminin sonuçlarını etkilemiştir. Bu yöntem için madde ayırt edicilik parametre kestirimlerinin RMSE ve BIAS değerleri, ortak maddelerinin yüzdesi arttıkça azalma eğilimi göstermiştir. Madde ayırt edicilik parametre kestirimlerinin RMSE ve BIAS değerleri, ortalama/ortalama ve Stocking-Lord yöntemleri için ortak madde faktörünün seviyeleri bakımından benzerdir.

Stocking-Lord ve ortalama/ortalama yöntemlerinin madde ayırt edicilik parametresi için daha iyi kestirimler sağladığı, Stocking-Lord ve ortalama/sigma yöntemlerinin madde güçlüğü parametresi için daha iyi kestirimler sağladığı sonucuna varılabilir.

Bu çalışmada çok boyutlu eşitleme yöntemlerinden sadece üçü madde parametrekestirimlerinin kararlılığı bakımından karşılaştırılmıştır. Gelecekteki araştırmalarda parametre kestirimlerinin kararlılığı tek boyutlu ve çok boyutlu eşitleme yöntemleri arasında karşılaştırılabilir. Koşullar gerçek test değerleri göz önünde bulundurularak test uzunluğu ve boyut sayısı çeşitlendirilerek oluşturulabilir.