Article History
Received: 19.01.2023
Received in revised form: 26.06.2023
Accepted: 26.06.2023
Available online: 28.06.2023
Article Type: Research Article

ADIYAMAN UNIVERSITY
Journal of Educational Sciences
(AUJES)

https://dergipark.org.tr/tr/pub/adyuebd

Investigation of Multidimensional Scale Transformation Methods Applied to Multidimensional Tests According to Various Conditions

Yaşar Mehmet ZOR[1]
[1]Ministry of Education, Turkey

To cite this article:
Zor, Y. M. (2023). Investigation of multidimensional scale transformation methods applied to multidimensional tests according to various conditions. *Adiyaman Univesity Journal of Educational Sciences*, *13*(1), 41-53.

Adiyaman University Journal of Educational Sciences

Volume 13, Number 1, June 2023, Page 41-53 ISSN: 2149-2727

DOI: http://dx.doi.org/10.17984/adyuebd.1239198

Investigation of Multidimensional Scale Transformation Methods Applied to Multidimensional Tests According to Various Conditions

Yaşar Mehmet Zor[1*]

[1] Ministry of Education, Turkey

Abstract

The purpose of this study was to compare the equating errors of item and ability parameters obtained by performing scale transformation methods to two multidimensional test forms under various conditions. Sample size (1000-2000), common item ratio (20% and 40%), correlation between dimensions (0.1-0.5-0.9) and parameter estimation model (2 parameter logistic model and 3 parameter logistic model) were taken as research conditions. Root Mean Squared Error (RMSE) was used to examine the accuracy of the scale transformation results. It was observed that the RMSE value generally decreased as the sample size and common item ratio increased and the correlation between dimensions decreased. Higher equating errors were obtained when the mean-sigma method was used. In the estimation of the discrimination parameter, lower RMSE values were obtained in 2PLM for all methods. In the estimation of difficulty and ability parameters, lower RMSE values were obtained in 2PLM for Stocking-Lord method and in 3PLM for mean-mean and mean-sigma methods.

Key words: Multidimensionality, Scale transformation, Mean-mean, Mean-sigma, Stocking-lord

Introduction

Tests in the field of education and psychology are used for various purposes such as determining the learning deficiencies of individuals, selection and placement of individuals in an educational institution or a job. The decisions to be taken about individuals can be accurate and fair only if the tests are valid and reliable. In many large-scale test applications applied at national and international level, different test forms are used for purposes such as ensuring test security and estimating the change in test scores by using different items (Öztürk Gübeş, 2014). In these applications, different test forms consisting of equal numbers of items are created so that the content and item format of the item are equivalent to each other (Xu, 2009). The main purpose of using different equivalent test forms is to compare the scores obtained from the tests and to use the scores obtained interchangeably. Although the test forms are prepared with similar content and psychometric characteristics in order to measure the same trait, the scores obtained from the test forms should be equated by numerical transformation (Braun & Holland, 1982). As a result of the numerical transformation, the scores obtained from the test forms are placed on the same scale and the scores obtained from the tests can be used interchangeably.

Comparing scores obtained from different test forms that are not at the same scale level may lead to inaccurate results. In order for the scores to be comparable, a statistical adjustment is made between the scores obtained from different test forms by test equating, so that the scores are on the same scale (Kolen & Brennan, 2014). Test equating, which is a statistical technique, reveals the relationship between the scores obtained from test forms (Chu & Kamata, 2003). Angoff (1984) defined test equating as equating the unit system of one form to the unit system of another form. As a result of test equating, scores obtained from different test forms can be used interchangeably (Hambleton & Swaminathan, 1985; Kolen & Brennan, 2014).

Test equating is a statistical process that regulates the differences between the scores obtained from two test forms with similar difficulty and content and allows these scores to be used interchangeably (Kolen & Brennan, 2014). One of the conditions required for test equating is that the test forms measure the same structure. In addition, the reliability coefficients of the test forms should be close to each other and features such as equality, symmetricity and group invariance should be met (Dorans and Holland 2000).

The first step of the test equating process is to decide on the equating design. Kolen and Brennan (2014) explained three basic test equating designs as follows; single group design, random groups design and non-equivalent groups anchor test (NEAT) design. The non-equivalent groups anchor test design is also referred

[*] Corresponding Author: *Yaşar Mehmet Zor*, yasarzor@gmail.com

to as the common-item non-equivalent groups (CINEG) design in the literature (Reckase, 2009; Topczewski, Cui, Woodruff, Chen, & Fang, 2013). Since the non-equivalent groups common test design has been widely used in many national and international large-scale tests, the non-equivalent groups common test design was used in this study. In this design, the test forms (Form X and Form Y) have a common set of items and these forms are administered to groups taking different tests. In this design, the group taking Form X is not considered equivalent to the group taking Form Y. Differences between Form X and Form Y averages are considered as a combination of differences between the groups taking the test and differences between the test forms (Kolen & Brennan, 2014). The difference in ability between the groups is controlled by means of common (anchor) items in both forms and the scores obtained from the test forms are equated to each other (He, 2011).

In a test equating process, after the equating design is selected, the equating method is determined and the item and ability parameters of each test form are estimated according to this method. Then, item and ability parameters are transformed into a common scale. Scale transformation should be performed when using a non-equivalent groups anchor test design, this step is not necessary since the parameters estimated in single and equivalent group designs will be on the same scale. After the item and ability parameters are placed on a common scale, the scale on which the test scores will be reported is determined. If the test scores are to be reported on ability parameters, the process is completed, but if reporting is to be done on true scores, true scores should be estimated according to different ability levels and true scores of both forms should be equated (Kabasakal, 2014).

Tests used in education and psychology are inherently multidimensional to some extent due to the many sources of multidimensionality involved in scoring (Ackerman, Gierl & Walker, 2003). The sources of multidimensionality may be that the test consists of more than one content area or that the test has more than one item format (mixed format tests). In such cases, it is very difficult to meet the unidimensionality assumption (Kim, Lee, & Kolen, 2020). Therefore, the relationship between items in tests used in education and psychology is not as simple as described in unidimensional models. It can be interpreted that the predictions made using unidimensional models have lower errors when there is a dominant dimension in which the items in the test are collected and the dimensionality source of the test is mostly explained by this dimension. In cases where the data structure of the test measures more than one latent trait, unidimensional models make predictions on the axis of the strongest dimension in the data structure, and ability and item parameters estimated using unidimensional models can be highly biased (Gibbons, Immekus, & Bock, 2007; Reckase, 1985). If more than one psychological trait affects the responses to the items in the test and the test consists of more than one content area or more than one item format, it can be interpreted that the unidimensionality assumption is violated and in this case, unidimensional models should not be used (Kim & Lee, 2022; Zhang, 2009). Instead, models under the multidimensional item response theory (MIRT) can be used.

In MIRT, item discrimination parameter is calculated differently from unidimensional item response theory. According to this theory, since the test consists of more than one dimension, the items in the test have a separate discrimination parameter for each dimension. Item discrimination is represented by multidimensional discrimination index (MDISC). MDISC is used in conjunction with the parameter a in unidimensional models. The length of the vector related to the MDISC index is calculated as given in Equation 1, where a_{ik} is the discrimination parameter of item i for each dimension (Reckase, 2009).

$$MDISC = \sqrt{\sum_{n=1}^{k} a_{ik}{}^{2}} \tag{1}$$

While the items in the test have a separate discrimination parameter for each dimension, they have a single difficulty parameter. The multidimensional difficulty index (MDIFF) is a parameter corresponding to the b parameter in the unidimensional item response theory. The d_i in Equation 2 is the intercept parameter related to item difficulty and calculated by the interaction of a and b parameters. MDIFF index is calculated as given in Equation 2 (Reckase, 2009).

$$MDIFF = \frac{-d_i}{MDISC} \tag{2}$$

It has been explained in the previous sections that as the number of latent traits measured by a test increases, the test measures more than one ability. While items in some tests measure more than one latent trait, each of the items in some tests may be combined under a separate latent trait. Tests with items that are related to more than one ability are called complex structured tests, while tests in which items are gathered under different abilities and are related only to the level of ability they measure are called simple structured tests (Zhang, 2012). While the items in a simple structured test load on only one dimension, in a complex structured test the items load on more than one dimension (Ackerman, Gierl, & Walker, 2003). Simple structure may never occur in real data structures, however, simple structure assumes that the secondary loadings of items for data fluctuate around zero (Kim & Lee, 2022). In this study, two-dimensional simple structured data sets were generated and analyses were conducted on these data sets.

In the test equating study using the NEAT design, the parameters estimated from different forms may not be on the same scale due to the differences in ability between the groups. Therefore, a linear transformation is required to place the parameters estimated from the test forms on the same scale. Separate calibration and concurrent calibration methods can be used to transform the item parameters of the test forms applied to different groups to the same scale (Kolen & Brennan, 2014). Since separate calibration has been found to give more accurate and reliable results in multidimensional data structures (Kim & Kolen, 2006; Kolen & Brennan, 2014), separate calibration method was used in this study.

Separate calibration refers to the situation where item and ability parameters for each form are estimated separately and then an additional linking procedure is applied to place the two sets of parameters on the same scale (Kim, 2018). In NEAT design, item parameters of common items are used for scale transformation. With the scale transformation process, it is aimed to transform the item and ability parameters estimated from the new test form into the scale of the item and ability parameters of the old test form. For this purpose, it is necessary to obtain the slope (A) and intercept (B) constants. With these constants, the equivalent of the ability parameter value in one form can be found in the other form. Considering that there are I and J forms of a test, the ability (θ) parameter of person i in Test I and its equivalent in Test J can be calculated as given in Equation 3 (Kolen & Brennan, 2014).

$$\theta_{Ji} = A\theta_{Ii} + B \tag{3}$$

The transformation of item parameters from Test I to Test J is given in Equation 4. Since the guess parameter (parameter c) is on the probability scale, there is no need for transformation.

$$\begin{aligned} a_{Jj} &= \frac{a_{Ij}}{A} \\ b_{Jj} &= Ab_{Ij} + B \\ c_{Jj} &= c_{Ij} \end{aligned} \tag{4}$$

a_{Ij}, b_{Ij}, c_{Ij} : Item parameters for test I of item j

a_{Jj}, b_{Jj}, c_{Jj} : Rescaled item parameters of item j for test J

The methods used for separate calibration in scale transformation can be defined as moment methods and characteristic curve methods. Moment methods include mean-mean and mean-sigma methods, while characteristic curve methods include Stocking Lord and Haebara methods. In this study, mean-mean and mean-sigma methods from moment methods and Stocking Lord from characteristic curve methods were used.

In mean-mean (MM) method, slope (A) and intercept (B) constants are obtained by using the means of discrimination and difficulty parameters for common items (Loyd & Hoover, 1980). The calculation methods of slope and intercept constants are given in Equation 5.

$$\begin{aligned} A &= \frac{\mu(a_i)}{\mu(a_j)} \\ B &= \mu(b_j) - A\mu(b_i) \end{aligned} \tag{5}$$

$\mu(a_i),\mu(a_j)$: Mean of the discrimination parameters estimated from common items in i and j scales, respectively

$\mu(b_i),\mu(b_j)$: Mean of the difficulty parameters estimated from common items in i and j scales, respectively

Means and standard deviations of the difficulty parameters of common items are used to obtain the slope (A) and intercept (B) constants in Mean-Sigma (MS) method, (Marco, 1977). The mathematical expression for the calculation of the constants A and B according to the mean-sigma method is given in Equation 6.

$$A=\frac{\sigma(b_j)}{\sigma(b_i)} \qquad B=\mu(b_j)-A\mu(b_i) \tag{6}$$

$\mu(b_i),\mu(b_j)$: The mean of the difficulty parameters estimated from common items in scales i and j, respectively

$\sigma(b_j),\sigma(b_i)$: Standard deviation of difficulty parameters estimated from common items in scales j and i, respectively

In Stocking-Lord (SL) method, the constants A and B are calculated by minimising the criterion function defined by the difference between the characteristic curves of the items instead of the parameters of the common items (Stocking and Lord, 1983). In this method, the sum is obtained for each parameter set and the square of the difference of the sums over the items is taken. Equation 7 shows the mathematical expression for this method (Kolen and Brennan, 2014).

$$SLdiff\,(\theta_i)=\left[\sum_{j:V}p_{ij}\left(\theta_{ji};a_{Jj},b_{Jj},c_{Jj}\right)-\sum_{j:V}p_{ij}\left(\theta_{ji};\frac{a_{Ij}}{A},\mathrm{A}b_{Ij}+B,c_{Ij}\right)\right]^2 \tag{7}$$

A : Slope constant

B : Intercept constant

p_{ij} : Item characteristic function for person i and item j

a_{Jj},b_{Jj},c_{Jj} : Item parameters for the jth common item in scale J

a_{Ij},b_{Ij},c_{Ij} : Item parameters for the jth common item in scale I

$j:V$: Indicates that the total formula is calculated on common items

SLdiff is summed over the examinees and the constants A and B are obtained by minimising the criterion given in Equation 8 (Kolen and Brennan, 2014).

$$SL_{crit}=\sum_{i}SLdiff\,(\theta_i) \tag{8}$$

Tests used in education and psychology generally do not consist of a single dimension, items are sometimes related to multiple dimensions, and there are other dimensions measured by test items other than a dominant dimension. When unidimensional equating methods are applied to multidimensional data structures, it can be interpreted that the equating relationships will contain a large amount of error due to the violation of the unidimensionality assumption (Brossman, 2010). Multidimensional equating methods should be used for multidimensional data. In this study, an answer to the question of how the magnitude of the equating errors to be obtained when multidimensional equating methods are applied to multidimensional data structures under various conditions is searched.

It is aimed to apply scale transformation methods under NEAT design using two-parameter logistic model (2PLM) and 3-parameter logistic model (3PLM) to two-dimensional simple structured test data produced under two different common item ratios (20% and 40%) with low (0.1), medium (0.5) and high (0.9) correlation between dimensions in sample sizes of 1000 and 2000 people, and to compare the estimated equating errors (RMSE) under the conditions considered in this study. It can be interpreted that the tests used in education and psychology measure more than one latent trait and are multidimensional due to their structure. In scale transformation studies, it is generally accepted that the data are unidimensional and analysis processes are carried out through unidimensional item response theory. Scale transformation studies conducted on multidimensional data structures are quite few. In this context, it is thought that the results of this study, which aims to compare the equating errors obtained from multidimensional scale transformation over multidimensional data and under the conditions considered, will contribute to the literature. In method section, the variables and conditions considered in the research are presented in a table, and the data generation, data analysis and evaluation criteria for these conditions are explained under separate headings.

Method

Research Data

Simulation data was used in the study. The reason for this may be that it is not possible to fulfil all of the conditions (sample size, correlation between dimensions, common item ratio, parameter estimation model) in real data structures. Within the framework of the conditions in the study, data sets for both forms of the test (X and Y) and item and ability parameters were generated using R software, and dichotomous (1-0) item response data were generated from item and ability parameters using R software.

In scale transformation methods, there are studies examining the effect of sample size on equating error. Skaggs and Lissitz (1986) stated that the sample size for 3PLM should be at least 1000, Gübeş (2019) took the sample size as 1000 and 2000 in her study, and Hanson and Beguin (2002), Gök and Kelecioğlu (2014) and Kumlu (2019) took the sample size as 1000 and 3000 in their studies. In this study, the sample size was taken as 1000 and 2000.

Conditions were created so that the correlation between dimensions was low (0.1), medium (0.5) and high (0.9) in the study. High correlation between the dimensions can be shown as an evidence for the unidimensionality of the test (Zhang, 2009). Beguin and Hanson (2001) observed that an increase in the correlation between dimensions leads to an increase in the total error when multidimensional model parameter estimation is used. Gübeş (2019) applied unidimensional scale transformation methods to a two-dimensional test data and obtained higher equating errors as the correlation between dimensions decreased.

Scale transformation is performed through common items (anchor) in NEAT design. In this design, anchor form is divided into two as internal anchor and external anchor. If common items are included in the total score of the individual, it is called internal anchor, if not, it is called external anchor (Crocker & Algina, 1986; Kolen & Brennan, 2014). In this study, internal anchor test was used. Angoff (1984) and Kolen and Brennan (2014) stated that the number of common items in test forms should not be less than 20 items or 20% of the total number of items. In this study, common item ratio was taken as 20% and 40%.

There are studies examining the effect of scale transformation using 2PLM and 3PLM on equating error (Gök & Kelecioğlu, 2014; Kim & Kolen, 2006; Kim & Lee, 2006). Accordingly, it is stated that the model used has an effect on the scale transformation and test equating process. The data of this study were generated according to 2PLM and 3PLM and in this way, it was aimed to reveal the effect of the guess parameter on the equating error when 3PLM was used.

In the study, the discrimination parameter (a) was generated from a uniform distribution with values ranging between 0.6 and 2 for both forms. The difficulty parameter (b) was generated from a normal distribution with a mean of 0 and a standard deviation of 1 with values between -3 and +3, and the guess parameter (c) was generated from a uniform distribution with values between 0.01 and 0.25. Between-groups ability distribution is not considered as a condition in this study. According to the classification based on the difference between the means of ability distributions between groups, if the difference is between 0.05-0.10, it can be defined as "wide", and when it takes values of 0.25 and higher, it can be defined as "very wide" (Wang at al., 2008). However, since the NEAT design was used in this study and the average ability difference between the groups was not desired to affect the results to be obtained regarding the conditions to be examined, the difference was taken as low as 0.05. When generating the ability parameters of one of the groups, the mean was taken as 0 and the standard deviation as 1, and when generating the ability parameters of the other group, the mean was taken as 0.05 and the standard deviation as 1, and the ability parameters were generated to show a multivariate normal distribution. In Table 1, the variables and conditions within the framework of the study are explained.

As seen in Table 1, a total of 24 (2x3x2x2) conditions were examined in this study, including sample size (2 conditions), correlation between dimensions (3 conditions), common item ratio (2 conditions) and parameter estimation model (2 conditions). In the literature, it was observed that at least 50 iterations were performed for each data set in order to make the research results consistent and stable (Hanson & Beguin, 2002) and 50 iterations were performed for each data set in this study.

Table 1. Variables and Conditions in the Study

Variables	Conditions	Number of Conditions
Sample size	1000-2000	2
Correlation between dimensions	0.1-0.5-0.9	3
Common item ratio	%20-%40	2
Parameter estimation model	2PLM-3PLM	2

Data Anaysis

In the study, 1200 (24x50) data sets for both forms of the test (X and Y) were generated in R software in order to compare the equating errors obtained from mean-mean (MM), mean-sigma (MS) and Stocking Lord (SL) scale transformation methods. Dichotomous (1-0) item response data were generated from the item and ability parameters produced using the mirt package (Chalmers, 2012) in R software. For both forms of the test, multidimensional and simple structure parameter estimations were performed separately for 2PLM and 3PLM in IRT PRO 4.2 software. Markov Chain Monte Carlo (MC-MC) method was used for parameter estimation. In order to set the parameters obtained from both forms on the same scale, scale transformation was performed by using the mean-mean, mean-sigma and Stocking Lord methods, which are separate calibration methods over the MIRT based on the parameters of the items in the first form and the common item parameters in the second form. Linkmirt software (Yao, 2009) was used for multidimensional scale transformation. The softwares used in the study was run through batch script with R software in order to analyse 50 iterations of the data sets at one time. Slope (A) and intercept (B) constants were obtained by using MM, MS and SL scale transformation methods, and then the Root Mean Squared Error (RMSE) value, which gives the amount of error of the transformed item and ability parameters for each scale transformation method, was calculated. The RMSE values estimated after each iteration for item and ability parameters were averaged separately to obtain a single RMSE value for each parameter. The mathematical expression for the calculation of the RMSE value is given in Equation 9.

$$RMSE\left(\tau_j\right)=\sqrt{\frac{\sum_{r=1}^{R}\left(\tau_{jr}-\tau_j\right)^2}{R}} \tag{9}$$

τ_j : True value of parameter j

τ_{jr} : Estimated value of parameter j for repeated data set (r=1, 2, 3, ..., R)

R : Number of iterations

Results and Discussion

Multidimensional scale transformation was performed over the data obtained from both forms within the framework of the conditions in the research by using 3PLM and 2PLM respectively, and the findings obtained were interpreted under separate headings respectively.

Results on RMSE Values Obtained When Scale Transformation is Performed Using 3PLM

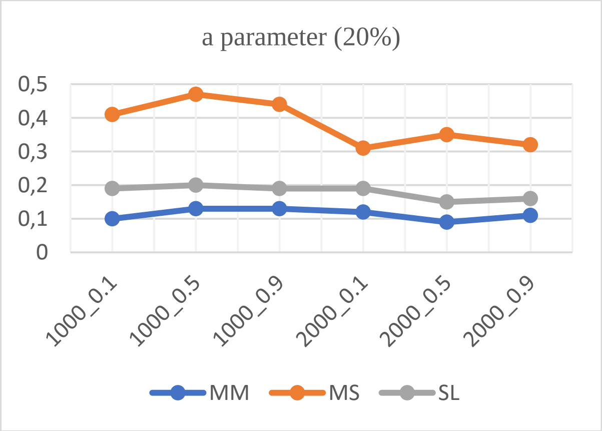
The RMSE values for the item and ability parameters obtained as a result of the multidimensional scale transformation process performed over 3PLM on both forms, including the discrimination parameter a, difficulty parameter b and the ability parameter theta, are given in Table 2.

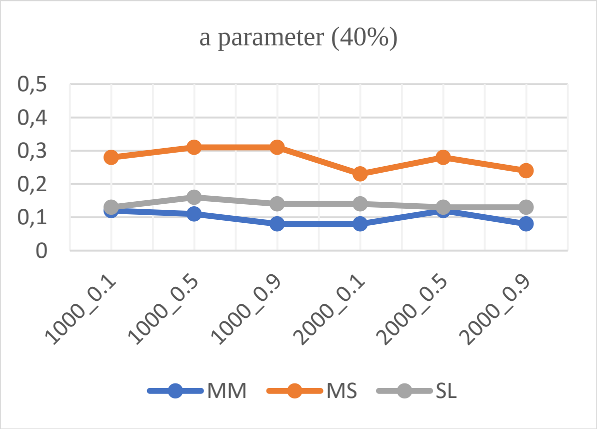
Table 2. RMSE Values of Scale Transformation Methods Using 3PLM

Correlation Between Dimensions	Sample Size	Common Item Ratio	RMSE Value								
			a			b			theta		
			MM	MS	SL	MM	MS	SL	MM	MS	SL
0.1	1000	%20	0.10	0.41	0.19	0.21	0.33	0.25	0.15	0.22	0.20

		%40	0.12	0.28	0.13	0.15	0.20	0.18	0.11	0.15	0.13
	2000	%20	0.12	0.31	0.19	0.15	0.25	0.32	0.11	0.18	0.28
		%40	0.08	0.23	0.14	0.11	0.25	0.28	0.07	0.17	0.22
0.5	1000	%20	0.13	0.47	0.20	0.26	0.39	0.34	0.18	0.25	0.29
		%40	0.11	0.31	0.16	0.14	0.28	0.24	0.10	0.19	0.18
	2000	%20	0.09	0.35	0.15	0.17	0.32	0.23	0.12	0.23	0.17
		%40	0.12	0.28	0.13	0.15	0.20	0.18	0.07	0.10	0.09
0.9	1000	%20	0.13	0.44	0.19	0.21	0.35	0.28	0.14	0.24	0.23
		%40	0.08	0.31	0.14	0.13	0.29	0.26	0.09	0.18	0.20
	2000	%20	0.11	0.32	0.16	0.17	0.33	0.29	0.12	0.23	0.24
		%40	0.08	0.24	0.13	0.10	0.21	0.25	0.07	0.15	0.20

When Table 2 is examined, it is seen that the RMSE value decreases as the sample size increases for all methods in the analyses using 3PLM. Similarly, it can be interpreted that the increase in the common item ratio is effective in obtaining lower RMSE values in general. When the RMSE values obtained for the methods are analysed, it is seen that the RMSE values for the mean-sigma method are higher than the other methods for all parameters. This supports the findings obtained by Atar and Yeşiltaş (2017). In addition, it is seen that the lowest RMSE values are obtained from the mean-mean method. While this is consistent with the findings of Ogasawara (2000) and Gök and Kelecioğlu (2014), which show that the mean-mean method yields better results than the mean-sigma method, it conflicts with the finding reported by Baker and Al-Karni (1991) and Hanson and Beguin (2002) that characteristic curve methods produce lower errors than moment methods. When the correlation between dimensions is 0.1 (low), the sample size is 2000 (high) and the common item ratio is 40% (high), the RMSE values of the mean-mean and mean-sigma methods are the lowest for all parameters. In the Stocking-Lord method, the lowest RMSE values for all parameters were obtained in conditions with correlation between dimensions of 0.5 (medium), sample size of 2000 (high) and common item ratio of 40% (high). Figure 1 shows the RMSE values obtained under all conditions as a result of multidimensional scale transformation using 3PLM for a, b and theta parameters, respectively.





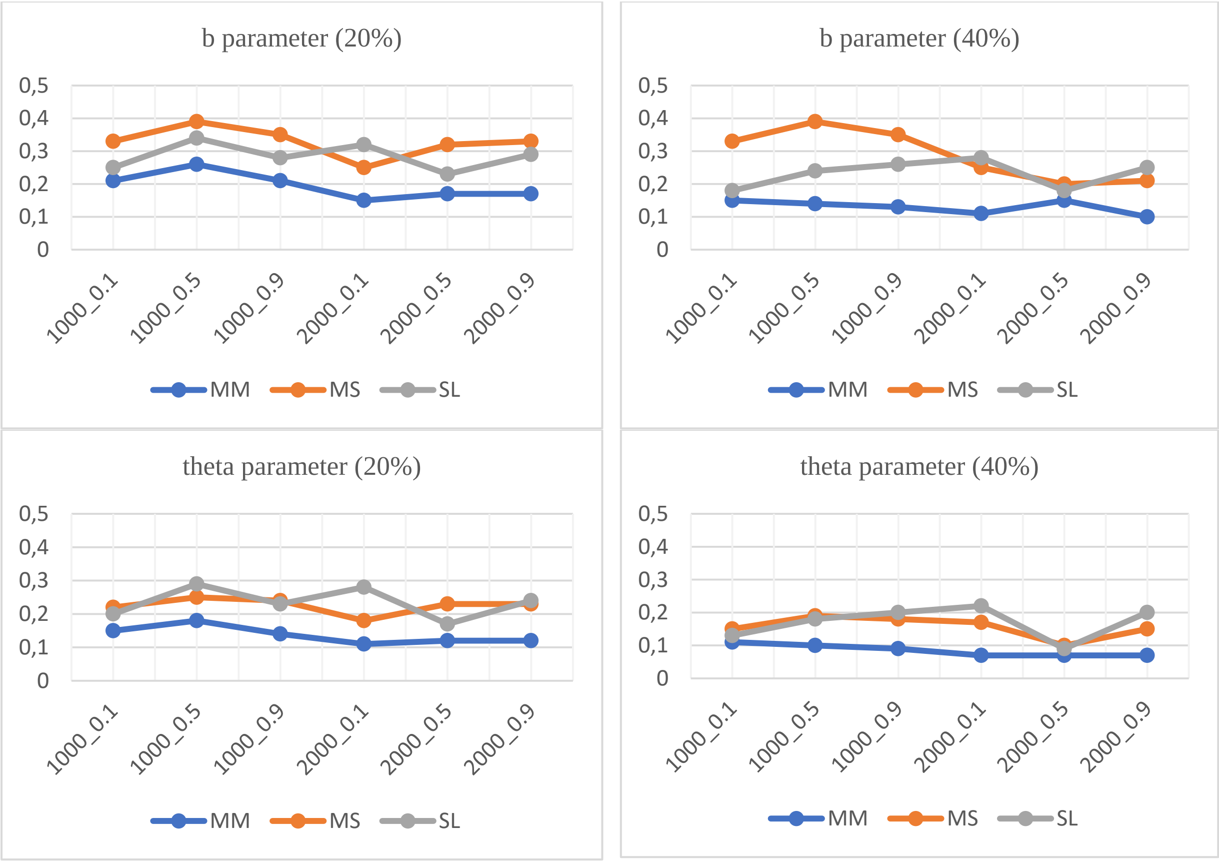


Figure 1. RMSE Values of Scale Transformation Methods Using 3PLM

Figure 1 shows that the lowest RMSE values for a, b and theta parameters, respectively, were obtained from the mean-mean method when the common item ratio was 40% and the sample size was 2000. When the sample size and common item ratio are constant, it can be interpreted that increasing the correlation between dimensions has a relatively decreasing effect on the RMSE values obtained from a and b parameters. The reason for this is that the increase in the correlation between dimensions strengthens the unidimensional characteristic of the test, and this situation is thought to provide more consistent estimation of the parameters of the test items.

Results on RMSE Values Obtained When Scale Transformation is Performed Using 2PLM

RMSE values for item and ability parameters obtained when multidimensional scale transformation was performed on both forms using 2PLM are given in Table 3.

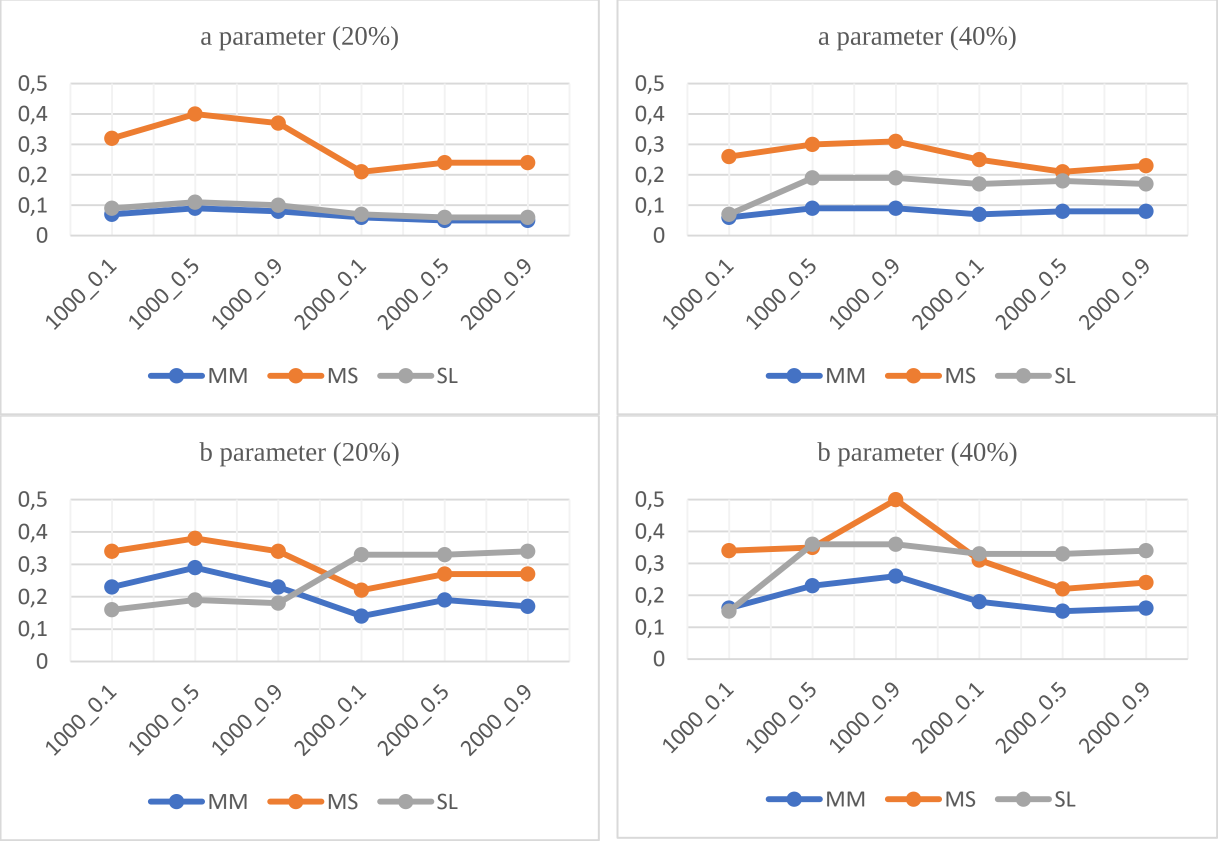
It is seen that the RMSE value decreases as the sample size increases for all scale transformation methods in the analyses using 2PLM as in 3PLM in Table 3. In addition, it can be interpreted that the increase in the common item ratio has a decreasing effect on the RMSE values obtained from the mean-mean and mean-sigma methods and an increasing effect on the RMSE values obtained from the Stocking-Lord method. When the RMSE values obtained for the methods are analysed, it is seen that the values for the mean-sigma method are higher than the other methods. When the correlation between dimensions was 0.1 (low), sample size was 2000 (high) and common item ratio was 20% (low), RMSE values of all methods were the lowest for all parameters.

Table 3. RMSE Values of Scale Transformation Methods Using 2PLM

Correlation Between Dimensions	Sample Size	Common Item Ratio	RMSE Value								
			a			b			theta		
			MM	MS	SL	MM	MS	SL	MM	MS	SL
0.1	1000	%20	0.07	0.32	0.09	0.23	0.34	0.16	0.16	0.25	0.11
		%40	0.06	0.26	0.07	0.16	0.34	0.15	0.11	0.24	0.11
	2000	%20	0.06	0.21	0.07	0.14	0.22	0.14	0.10	0.16	0.10

		%40	0.07	0.25	0.17	0.18	0.31	0.33	0.12	0.21	0.19
0.5	1000	%20	0.09	0.40	0.11	0.29	0.38	0.19	0.19	0.27	0.15
		%40	0.09	0.30	0.19	0.23	0.35	0.36	0.17	0.25	0.21
	2000	%20	0.05	0.24	0.06	0.19	0.27	0.15	0.13	0.21	0.10
		%40	0.08	0.21	0.18	0.15	0.22	0.33	0.09	0.15	0.20
0.9	1000	%20	0.08	0.37	0.10	0.23	0.34	0.18	0.16	0.26	0.15
		%40	0.09	0.31	0.19	0.26	0.53	0.36	0.19	0.34	0.21
	2000	%20	0.05	0.24	0.06	0.17	0.27	0.15	0.11	0.20	0.11
		%40	0.08	0.23	0.17	0.16	0.24	0.34	0.11	0.17	0.20

When RMSE values according to parameter estimation models are analysed, it is seen that RMSE values for discrimination parameter are lower in 2PLM for all methods. When the RMSE values for difficulty and ability parameters are analysed, it is seen that the values for Stocking-Lord method are lower in 2PLM, while lower values are obtained in 3PLM for other methods. In this context, it can be interpreted that generally lower RMSE values are obtained when 2PLM is used. This finding is consistent with the finding of Kaskowitz and De Ayala (2001) that 3PLM estimates parameters a and b with higher error. In cases where 2PLM is used as a parameter estimation model, this may be explained by the fact that more stable item parameter estimates are obtained in large samples (Bökeoğlu at al., 2022). Figure 2 shows the RMSE values obtained under all conditions as a result of multidimensional scale transformation using 2PLM for a, b and theta parameters, respectively.



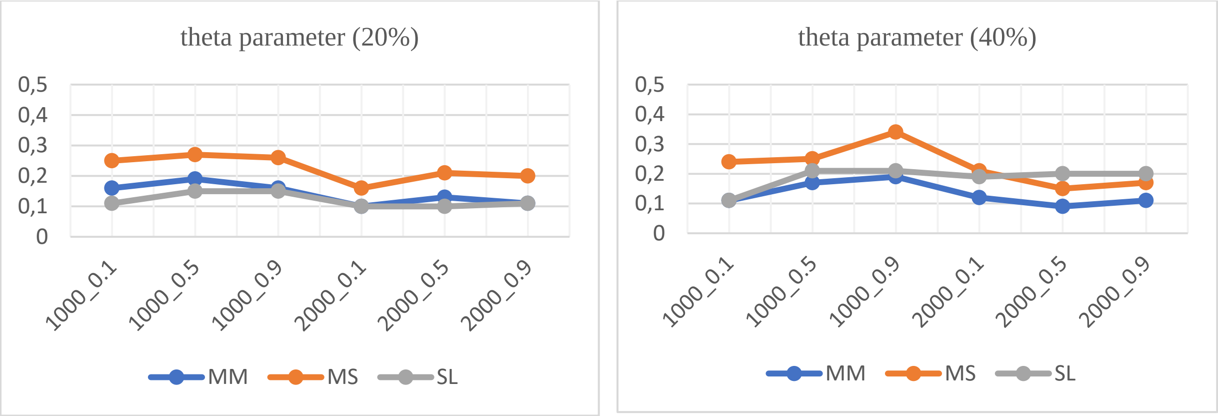


Figure 2. RMSE Values of Scale Transformation Methods Using 2PLM

According to Figure 2, it is seen that the lowest RMSE values for a and b parameters are generally obtained when the sample size is 2000 and the common item ratio is 20%. For the discrimination parameter, the lowest RMSE value was obtained from the mean-mean method when the correlation between dimensions was 0.5 and 0.9, while for the difficulty parameter, the lowest RMSE value was obtained from the mean-mean and Stocking-Lord methods when the correlation between dimensions was 0.1. The lowest RMSE value for the ability parameter was obtained from the mean-mean method when the common item ratio was 40% and the correlation between dimension was 0.5 in a sample size of 2000. When the RMSE values of the scale transformation methods were compared, values obtained from the mean-sigma method were found to be higher.

Conclusion and Recommendations

In this study, it was aimed to compare the equating errors of mean-mean, mean-sigma and Stocking-Lord scale transformation methods under various conditions (sample size, correlation between dimensions, common item ratio and parameter estimation model) using multidimensional item response theory. For this purpose, data were generated according to various conditions by taking into account the conditions considered in previous national and international studies and it was investigated which of these conditions produced the least error.

As a result of the analyses, it was seen that the increase in the sample size and the common item ratio had a decreasing effect on the equating errors, and the low correlation between the dimensions led to low equating errors due to the structure of the data and the methods used. Obtaining low equating errors at high sample sizes is consistent with the findings in the literature (Atar & Yeşiltaş, 2017; Hanson & Beguin, 2002). In addition, the findings obtained from the study are consistent with the finding that the increase in the correlation between dimensions causes an increase in the equating error when multidimensional scale transformation is performed (Beguin & Hanson, 2001; Gübeş, 2019). In addition, in cases where 3PLM was used, the increase in the common item ratio decreased the equating errors.

Estimates with the lowest equating error were obtained in the condition with a sample size of 2000 and the correlation between dimensions was 0.1 when both 2PLM and 3PLM were used. In addition, higher equating errors were obtained when the mean-sigma method was used for both models. In the estimation of the discrimination parameter, lower RMSE values were obtained when 2PLM was used for all methods. In the estimation of difficulty and ability parameters, lower values were obtained in 2PLM for Stocking-Lord method, while lower RMSE values were obtained in 3PLM for mean-mean and mean-sigma methods. It is concluded that the equating errors obtained from the mean-mean and Stocking-Lord methods are lower when the 2PLM is used, and from the mean-mean method when the 3PLM is used. This finding coincides with the findings of Ogasawara (2000), Gök and Kelecioğlu (2014), and conflicts the study by Baker and Al-Karni (1991), who found that the method with the least error is Stocking-Lord. The parameter estimation model caused differences in the equating errors and in determining the scale transformation method with the least error. In this case, it can be interpreted that guess success affects the parameter estimation and also affects the equating error.

In order to be able to interpret and generalise the results of a research correctly, the conditions in the research and their interactions with each other should be taken into consideration. Considering the conditions and the results obtained in this study, it is understood that attention should be paid to the selection of sample size, common item ratio and correlation between dimensions in multidimensional scale transformation. When the results of the research are evaluated in general, it can be interpreted that the scale transformation method with the lowest equating error can be obtained by using the mean-mean method when the sample size is 2000 and the correlation between dimensions is 0.1. Equating errors of all methods decreased due to the fact that the error related to parameter estimation decreased in large samples and more stable estimations were made.

According to the results of the study, the performance of the scale transformation methods differed according to the conditions considered, thus there is no definite conclusion about which method will give the best result. It is important that the findings obtained in scale transformation studies are consistent with the findings of previous studies. Using many methods together and comparing the results will help in choosing the most appropriate method (Hanson & Beguin, 2002).

This study is limited to sample size, common item ratio, correlation between dimensions, parameter estimation method conditions and certain levels of these conditions. A similar study can be conducted by considering the condition of ability distribution between groups. This study was conducted with simulation data. Similar studies can be carried out using real data. By generating simulation data similar to the conditions of real data, the errors obtained from both data types can be compared. In this way, possible differences and errors due to the use of real data or simulation data can be revealed. In addition to the RMSE value used as an evaluation criterion in this study, comparisons can be made using different evaluation criteria such as standard error of equating and bias.

References

Ackerman, T. A., Gierl, M. J., & Walker, C. M. (2003). Using multidimensional item response theory to evaluate educational and psychological tests. *Educational Measurement: Issues and Practice, 22*(3), 37-51.

Angoff, W. H. (1984). *Scales,norms and equivalent scores.* New Jersey: Educational Testing Service.

Atar, B. ve Yeşiltaş, G. (2017). Investigation of the Performance of Multidimensional Equating Procedures for Common-Item Nonequivalent Groups Design. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi, 8*(4), 421-434.

Baker, F. B. & Al-Karni, A. (1991). A comparison of two procedures for computing IRT equating coefficients. *Journal of Educational Measurement, 28*(2), 147–162.

Braun, H. I and Holland, P. W. (1982). Observed-score test equating: A mathematical analysis of some ETS equating procedures. *In P. W. Holland and D.B. Rubin (Ed.), Test equating* (s. 9-49). New York: Academic Press.

Beguin, A. A., & Hanson, B. A. (2001). *Effect of noncompensatory multidimensionality on separate and concurrent estimation in IRT observed score equating*. Paper presented at the The Annual Meeting of the National Council on Measurement in Education, Seattle, WA.

Bökeoğlu, Ö., Uçar, A. ve Balta, E. (2022). Investigation of Scale Transformation Methods in True Score Equating Based on Item Response Theory. *Ankara Üniversitesi Eğitim Bilimleri Fakültesi Dergisi. 55*(1), 1-36.

Brossman, B. G. (2010). *Observed score and true score equating procedures for multidimensional item response theory* (Doctoral dissertation). University of Iowa, Iowa.

Chalmers, R. P. (2012). Mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software, 48*, 1-29.

Chu, K. L. & Kamata, A. (2003). *Test equating with the presence of DIF*. Paper presented at the annual meeting of American Educational Research Association, Chicago.

Crocker, L., and Algina, J. (1986). *Introduction to classical and modern test theory*. New York: Holt, Rinehart & Winston.

Dorans, N. J., & Holland, P. W. (2000). Population invariance and the equatability of tests: basic theory and the linear case. *Journal of Educational Measurement, 37*(4), 281-306.

Gibbons, R. D., Immekus, J., and Bock, R. D. (2007). *Didactic workbook: The added value of multidimensional IRT models*. National Cancer Institute Technical Report.

Gök, B. ve Kelecioğlu, H. (2014). Comparison of IRT Equating Methods Using the Common-Item Nonequivalent Groups Design. *Mersin Üniversitesi Eğitim Fakültesi Dergisi, 10*(1), 120-136

Gübeş, N. Ö. (2019). Effect of Multidimensionality on Concurrent and Separate Calibration in Test Equating. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi, 34*(4), 1061-1074.

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

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

He, Y. (2011). *Evaluating equating properties for mixed-format tests* (Doctoral dissertation). University of Iowa, Iowa.

Kabasakal, K. A. (2014). *The Effect Of Differential Item Functioning on Test Equating* (Doctoral dissertation). Hacettepe Üniversitesi Eğitim Bilimleri Enstitüsü, Ankara.

Kaskowitz, G. S., & De Ayala, R. J. (2001). The effect of error in item parameter estimates on the test response function method of linking. *Applied Psychological Measurement, 25*, 39-52.

Kim, S., & Kolen, M.J. (2006). Robustness to format effects of IRT linking methods for mixed-format tests. *Applied Measurement in Education, 19*(4), 357-381.

Kim, S., & Lee, W. (2006). An extension of four IRT linking methods for mixed- format tests. *Journal of Educational Measurement, 43*(1), 53-76.

Kim, S. Y. (2018). *Simple structure MIRT equating for multidimensional tests* (Doctoral Dissertation). University of Iowa, Iowa.

Kim, S., Lee, W. C. And Kolen, M. J. (2020). Simple-Structure Multidimensional Item Response Theory Equating for Multidimensional Tests. *Educational and Psychological Measurement. 80*(1), 91-125.

Kim, S. & Lee, W. (2022). Several Variations of Simple-Structure MIRT Equating. *Journal of Educational Measurement*. https://doi.org/10.1111/jedm.12341

Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking: Methods and practices (3rd edition)*. New York: Springer.

Kumlu, G. (2019). *An Investigation of Test and Sub-Tests Equating In Terms Of Different Conditions* (Doctoral dissertation). Hacettepe Üniversitesi Eğitim Bilimleri Enstitüsü, Ankara.

Loyd, B. H., & Hoover, H. D. (1980). Vertical equating using the rasch model. *Journal of Educational Measurement, 17*(3), 179-193.

Marco, G. L. (1977). Item characteristic curve solutions to three intractable testing problems. *Journal of Educational Measurement, 14*(2), 139-160.

Ogasawara, H. (2000). Asymptotic standard errors of IRT equating coefficients using moments. *Economic Review (Otaru University of Commerce), 51*(1), 1-23.

Öztürk Gübeş, N. (2014). *The Effects of Test Dimensionality, Common Item Format, Ability Distribution and Scale Transformation Methods on Mixed-Format Test Equating* (Doctoral dissertation). Hacettepe Üniversitesi Eğitim Bilimleri Enstitüsü, Ankara.

Reckase, M. D. (1985). The difficulty of test items that measure more than one ability. *Applied Psychological Measurement, 9*, 401-412.

Reckase, M. D. (2009). *Multidimensional item response theory*. New York: Springer.

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

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 7*, 201-210.

Topczewski, A., Cui, Z., Woodruff, D., Chen, H. and Fang, Y. (2013). A comparison of four linear equating methods for the common-Item nonequivalent groups design using simulation methods. *ACT Research Report Series, 2013*(2).

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

Yao, L. (2009). *LinkMIRT: Linking of Multivariate Item Response Model*. Monterey, CA: Defense Manpower Data Center.

Xu, Y. (2009). *Measuring change in jurisdiction achievement over time: Equating issues in current international assessment programs* (Doctoral dissertation). University of Toronto, Toronto.

Zhang, B. (2009). Application of unidimensional item response models to tests with item sensitive to secondary dimensions. *The Journal of Experimental Education, 77*(2), 147-166.

Zhang, J. (2012). Calibration of response data using MIRT models with simple and mixed structures. *Applied Psychological Measurement, 36*(5), 375-398.