ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2018; 9(4);326-338

An Investigation of the Factors Affecting the Vertical Scaling of Multidimensional Mixed-Format Tests*

Akif AVCU ** Hülya KELECİOĞLU ***

Abstract

This study examined the effect of the structure of a common item set (only dichotomous common items – mixed-format common item sets), parameter estimation methods and scale shrinkage on vertical scaling results when multidimensional datasets were used within the context of Common Item Nonequivalent Group (CINEG) design. Interactions between these variables were also investigated. The study was performed using simulated data. Measurement error and bias indexes were used to evaluate the quality of vertical scaling. All the procedures used in the data analysis were replicated 50 times to increase the generalizability of the results. R program was used for the data generation, calibration of the parameters and vertical scaling procedures. Possible interactions were investigated with factorial analysis of variance by using SPSS. The results showed a consistent effect of the common item format in all conditions. In addition, some interactions between the variables were observed. These findings are discussed and some recommendations are provided.

Keywords: Vertical scaling, Multidimensional tests, Mixed format tests

INTRODUCTION

Test scores are among the primary sources of information that educators and educational institutions use in making important decisions about students. Thus, test scores must provide the accurate information to facilitate appropriate decisions (Kolen and Brennan, 2004). However, different forms of the same test are often used due to the reasons such as test safety and follow up student development. A functional link between these forms needs to be established so that the scores from different test forms are comparable. This process is called test linking. Test linking is the process of establishing a relationship between different test forms. There is no requirement that the content and difficulty levels between the test forms for test binding be the same. Test equating is a special form of linking in which the aim is to use the scores between the different test forms interchangeably. Hence, test forms should be similar in content and difficulty (Kolen and Brennan, 2004). Vertical scaling is similar to the equating because different test forms are linked to each other. However, test forms differ in content and difficulty because they reflect progression between classes or age groups. Therefore, while vertical scaling is used to compare different test forms, the scores at each level can not be used in place of each other. When the scores are put onto a common scale, students' grade-to-grade improvement can be seen. The main aim of vertical scaling is to observe student progress.

Dichotomous items were the most widely used item format in the 20th century (Koretz and Hamilton, 2006). Today, however, the use of mixed-format tests, which contain both dichotomous and polytomously scored items, is rapidly becoming widespread. Mixed-format tests offer many advantages. According to Livingston (2009), multiple-choice questions may be used to measure a test taker's ability with high reliability for a wide range of contents, in a short time and at low cost. On the other hand, open-ended questions measure higher-level cognitive skills more effectively, but tests

* This study is based on the dissertation "Çok boyutlu karma-format testlerin ölçeklenmesini etkileyen faktörlerin incelenmesi" advised by Prof. Dr. Hulya Kelecioglu in July, 2006.

** Phd, Marmara University, Atatürk Faculty of Education, İstanbul-Turkey, e-mail:avcuakif@gmail.com ORCID ID: 0003-1977-7592
*** Professor, Hacettepe University, Ankara-Turkey, e-mail:hulyakelecioglu@gmail.com, ORCID ID: 0002-0741-9934

To cite this article:
Avcu, A., & Kelecioğlu, H. (2018). An investigation of the factors affecting the vertical scaling of multidimensional mixed-format tests. *Journal of Measurement and Evaluation in Education and Psychology, 9*(4), 326-338. DOI: 10.21031/epod.394659

Received: 14.02.2018
Accepted:13.10.2018

made up of these items have narrower content, are costly to assess, and are likely to be subjective. Mixed-format tests eliminate the disadvantages of different formats and increase the psychometric qualities of the instruments.

One of the variables that this study examines is the effect of scale shrinkage which becomes relevant when measurement tools are applied at different time points to detect students' progress. Scale Shrinkage is the extent to which the variance and range of scores decrease in the second application compared to the first (Yen, 1985). As students continue any program, they become more homogenous in terms of their ability compared to beginning. This leads their score variances to shrink at the later test applications. The scale shrinkage corresponds to the homogeneity. So far, there has been a lack of research on how scale shrinkage affect the results of vertical scaling.

Another important variable examined in this study is the structure of the common item set. Mixed-format tests are scaled vertically through the use of only dichotomous items, mixed-format items (including at least two different items in the common item set and only polytomous items). In this study, only the dichotomous common item set and mixed-format common item set conditions were compared. Although, positive outcomes were obtained for the mixed common item set for vertical scaling applications (e.g., Kim and Lee, 2006), it could be valuable to see the results within the context of the current study where a different combination of variables is included.

Most software programs routinely carry out estimations using expectation - maximization (EM) algorithms (Bock and Aitkin, 1981). Another method that has recently started to be used is the Metropolis-Hastings Robbins-Monro (MHRM) method developed by Chai (2010). The performance of EM and MHRM have been compared in estimating multidimensional dichotomous models (Han and Paek, 2010) and multidimensional polytomous models (Kuo and Sheng, 2016). However, no study was found comparing their performances in the context of multidimension mixed-format scaling. A comparison of these estimation methods could contribute important insights to the literature. Thus, we also varied the estimation method across the study conditions.

Vertical Scaling of Mixed-Format Tests

Dichotomous Item Response (IRT) Models:

Dichotomous response models are based on a three-parameter logistic model developed by Birnbaum (1968). This model is expressed in formula 1 (Lord and Novick, 1968).

$$p_i(\theta) = c_i + \frac{1 - c_i}{1 + e^{-a_i(\theta - b_i)}} \quad (1)$$

Here θ corresponds to the level of the individual's ability, a to the distinction parameter, b to the difficulty parameter, and c to the so-called chance parameter of the item. When this model was first introduced, it was used for one-dimensional tests, but since the 1980-s it has been used for multidimensional models as well. The generalization of Birnbaum' (1968) model for multidimensional tests is given in the following section.

For example, let us say that, i = 1,……, N are different participants, j = 1,……, and n are test items. Also, suppose m is a latent factor. θ = (θ_{i1},….,θ_{iN}). The slope parameters associated with the dimensions are a_j = (a_1,……, a_j). The likelihood of responding to a dichotomous item for multidimensional 3PLMs becomes as presented in formula 2:

$$\phi\,(x_{ij} = 1 \mid \theta_i, \alpha_j, d_j, \gamma_j) = \gamma_j + \frac{(1-\gamma_j)}{1+\exp[-D\left(\alpha_{j,}^{\mathrm{T}}\ \theta_i + d_j\right)]} \quad (2)$$

Here, d_j corresponds to the intersection parameter, γ_j corresponds to "*chance*" parameter, and D corresponds to the scaling constant. This value is generally taken as 1.702, and it is used to transform the logistic metric to the traditional normal ogive metric (Reckase, 2009).

Polytomous IRT Models

Although there are different models for polytomous items in the literature, the graded response model (GRM) is preferred in this study. In this model, developed by Samejima (1969, 1972), if we assume that the discrimination parameters are kept constant. $\tilde{P}_{ijk}$ corresponds to the cumulative probability that a person *i* with θ_i ability level can obtain a score beyond the category k of the item j. For the K_j categories, $\tilde{P}_{ijk}$ could be expressed as follows:

$$\tilde{P}_{ijk}=\tilde{P}_{jk}(\theta_i) = \tilde{P}(\theta_i)|\alpha_j, b_{jk}) = \begin{cases} 1 & \\ \dfrac{\exp[Da_j(\theta_i - b_{jk})]}{1+\exp[Da_j(\theta_i - b_{jk})]} & \begin{matrix} k = 1, \\ 2 \le k \le K_j, \\ K > K_j \end{matrix} \\ 0 & \end{cases} \quad (3)$$

Here, a_j corresponds to the discrimination parameter, b_{jk} corresponds to difficulty (or threshold parameter) from the second category to category K, and D corresponds to the scaling constant. The category response function, P_{ijk}, corresponds to the difference between two adjacent cumulative probabilities and is expressed as follows:

$$P_{ijk} = P_{jk}(\theta_i) = \tilde{P}_{ijk} - \tilde{P}_{ij(k+1)} \quad (4)$$

Samejima (1969) and Carlson (1995) used the GRM to generalize this to multidimensional situations.

In the model, the boundaries of the response categories for the C_j categories belonging to item *j* and the $d_j = d_1, \ldots\ldots d_{(c_{(j)}-1)}$ intersections are expressed as follows:

$$\phi(x_{ij} \ge 0|\theta_i, \alpha_j, d_j) = 1$$

$$\phi(x_{ij} \ge 1|\theta_i, \alpha_j, d_j = \frac{1}{1+\exp[-D(\alpha_j^T, \theta_i, d_j]}$$

$$\phi(x_{ij} \ge 2|\theta_i, \alpha_j, d_j = \frac{1}{1+\exp[-D(\alpha_j^T, \theta_i, d_j]}$$

.....

$$\phi(x_{ij} \ge 0|\theta_i, \alpha_j, d_j) = 0$$

(5)

Vertical Scaling

In the literature, moment and characteristic methods are the most commonly preferred methods used to apply vertical scaling. The moment methods, namely, mean / sigma (Marco 1977) and mean/mean (Loyd and Hoover, 1980), are the simplest methods, and only the parameter estimates need to be known in order to estimate linking constants. Alternative methods to the moment methods are the characteristic curve methods developed by Haebara (1980) and Stocking and Lord (1983). These methods based on minimization of the differences between characteristics curves of items. Comprehensive analysis and comparisons of these methods were provided by Kolen and Brennan (2004). These methods were extended to link mixed format tests. A detailed information can be found in Kim and Lee (2006)'s study. This study uses the Haebara method.

Purpose of the Study

Based on the literature presented above, the aim of the current study is to investgate the effect of common item structure, scale shrinkage, and estimation methods on the vertical scaling of multidimensional mixed-format tests.

METHOD

Data Simulation

Simulated datasets were used in the study. In addition, population parameters were also simulated considering that the values can be observable in real testing conditions. The dimensionality structure was prepared considering the two-tier model proposed by Cai (2010). In two-tier models, main dimensions and special dimensions are used as the source of the dimensionality. The terms "*main*" and "*special*" do not imply that main dimensions are theoretically more important or the variance/covariance structure between the dimensions is different. There is no theoretical relationship between main and special dimensions in a two-tier model. In addition, special factors are mutually orthogonal, and items have loading from only one dimension. On the other hand, the main factors may be related to each other. In the context of this study, content and item format were regarded as two sources of dimensionality in the data simulation. Accordingly, "*content*" was regarded as a special dimension source and "*item format*" as a main dimension source. Dual effect of content and item format on test dimensionality was investigated by Zhang (2016), but no study was found that take both factors into consideration when conducting scaling studies using simulated data. Figure 1 shows the model used for the data simulation in the current study. As seen, the three dimensions based on content and the two dimensions based on item format are intertwined in the model in compatible with two-tier models. The variance-covariance matrix used for the data simulation was established based on this model. The matrix is shown in Fügure 2. As seen in the figure, the relationship between the general factors is set to be 0.75 Among the special factors, this value is 0. Likewise, covariance values are assumed to be 0, showing no relationship between general and special items.

Simulation of Person and Item Parameters

In order to obtain accurate and stable parameter estimations in Multidimensional Item Response Theory (MIRT), as a sample size of 3000 was recomended (Yao and Boughton, 2009). Thus, in this study, the sample size consisted of 3000 simulated examinees. Theta scores were simulated from a normal distribution. The mean for the lower ability group was set to 0 and that for higher ability group was set to 1 with different scale shrinkage levels. One point ability difference between the groups was acceptable value that can be seen in real testing conditions (e. g., Kim, 2007). The theta vectors were simulated for each specific factor. Thus, final θ matrices with 3×3000 size were obtained as the population parameters for each group. In addition, variances of the population ability parameters were controlled. For the cases of scale shrinkage, variances were set to a shrinkage of 65% for the higher ability group. This amount of shrinkage was selected based on the study literature review provided by Yen (2005). This level of shrinkage was the case for half of the datasets, while for the rest of the datasets, variances were kept the same for both datasets used in the vertical scaling.

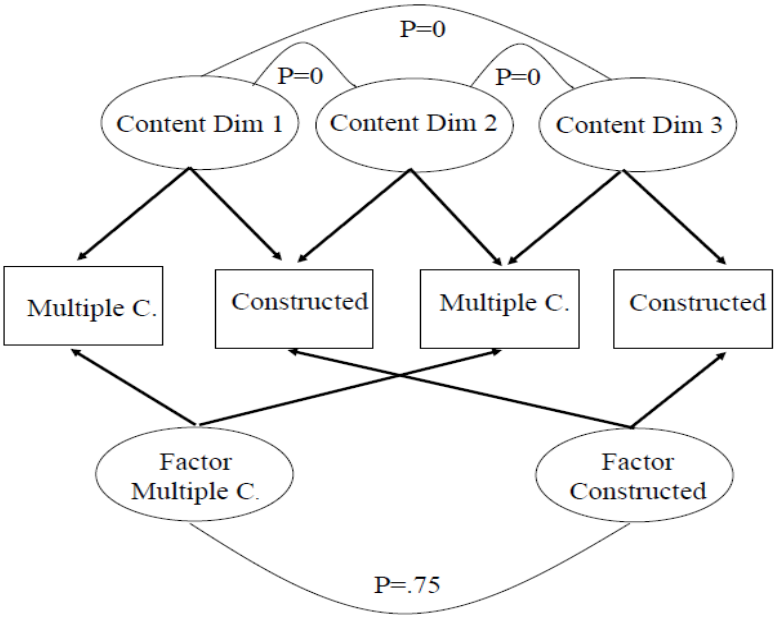

Figure 1. Two Tier Model

In addition, the datasets were created to be composed of 108 items (90 dichitomous and 18 polytomously scored items). In this scenario, there were 54 items (45 dichotomous and 9 polytomous) in each main factor and 36 items (30 dichotomous and 6 polytomous) in each factor.

Population *a* parameters for the generation of the data matrices were generated for each dimension (for each of the main and dimensions dimensions). Thus, a final matrix with a size of 5 × 108 was obtained for each dataset. For the main factors, if the item belonged to a dimension, the mean *a* value was determined as high discrimination power and fixed at 1 and the standard deviation at 0.15. If an item did not belong to a dimension, the mean value was fixed at 0.2 and the standard deviation at 0.03, because these items were not expected to have high level of discrimination. For special factors, if the item is included in that dimension, the mean value was fixed at 1 and the standard deviation at 0.15, while if the item was not included in that dimension, the all the *a* values were fixed to be 0 because of simple structure of spesific factors. All the simulated discrimination parameters were selected from the standard normal distribution.

$$\begin{bmatrix} 1 & 0.75 & 0 & 0 & 0 \\ 0.75 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

Figure 2. Variance-Covariances for Three Dimensional Data

The difficulty parameters (*b*) were produced as 1 × 108 vectors for dichotomous items. For the lower ability group, the mean was set to be 0 with a standard deviation of 1. For the higher ability group, the mean was set to be 1 with a standard deviation of 1. Polytomous items were configured as having a 5-point scoring format. For this reason, four intercept parameters for each item were simulated. The threshold values for the lower ability group are simulated with means that ranged from -1.5 to 1.5 with

a 1-point increase for every adjacent thresholds. For the higher ability group, same procedure was repeated except the range was set to -1 to 1. The distribution of the difficulty parameters was selected from a normal distribution with a standard deviation of 0.1. Data matrices were simulated as described above using parameter estimates and matrices produced for the calibration process.

Parameter Estimation

In this study, 3PLM and GRM were used to calibrate the mixed-format tests. This combination is preferred in many studies. Rosa, Swygert, Nelson, and Thissen (2001) pointed out that 3PLM is preferred for calibration because more parameters on the items are taken into account and the model therefore gives more information. In the literature, GRM and partial credit models are preferred in the calibration of polytomously response models (Kim and Cohen, 2002, Bastari, 2000, Tate, 2000). Dodd (1984) concluded that the two model types produce similar the results despite being conceptually and mathematically different. Cao, Yin and Gao (2007) also found that the two models yield similar results.

For theta estimation the MAP was preferred in this study. Each data set was calibrated separately so that the scaling process could be performed. In the analysis of each data set for EM cycles, the convergence value and the number of iterations were taken as 0.001 and 500, respectively. For the MH-RM estimation technique, the convergence value was set to 0.0001 and the number of iterations to 2000.

Evaluation Criteria

As the evaluation criteria of the results, root mean square error (RMSE) and bias were used in parallel with similar studies. RMSE shows the amount of random error fort he scaling process. The computation of RMSE is given in formula 6:

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(\hat{\theta}_i - \theta_i)^2}{n}} \tag{6}$$

The bias values provide information on the systematic error detected during scaling process and are calculated as described in formula 7:

$$\text{Bias} = \frac{\sum_{i=1}^{n}(\hat{\theta}_i - \theta_i)}{n} \tag{7}$$

Data Analysis

The data analysis was performed using the R statistic program (R Core Team, 2015). Different R packages were loaded and the analyses were carried out. Firstly, the "*truncnorm*" package developed by Trautmann et al. (2014) was used when *d*-matrices were derived. This package is preferred for controlling the upper and lower bounds of derived values of population threshold parameters. In this way, it was ensured that the difference between successive threshold parameters did not fall below 0,3 and model-data mis-fit was prevented. Other population parameters were obtained by using the "*rnorm*" command in the R program.

Later, the "*mirt*" package developed by Chalmers (2012) was used. Response matrices is producesd with the command "*sim*", and calibration was performed with "*mirt*" command. Finally, the ability parameters were estimated using the "*fscores*" command. When the scaling was performed, the "*plink*" package developed by Weeks (2010) is utilized. Each analysis was replicated 50 times. Then, the error and bias values of the parameters obtained from 50 replications were used with analysis of variance (ANOVA) to compare the conditions tested, and in 2 × 2 × 2 factorial ANOVA to see the interactions among the conditions being tested.

RESULTS

This section discusses the amount of error (RMSE), bias (Bias) values, and results of the factorial ANOVA for each research question as the major findings of the study. Common item set was excluded in the calculation of the error and bias values. In addition, given values of the dichotomous and polytomous items were calculated separately. As a last caution to the reader, the error and bias values were calculated separately for each of the three dimensions. The values are presented in Table 1.

Table 1 shows that the common item structure had a significant effect on some estimates of the synchronization structure. The error values in the threshold parameters of the polytomous items for the first dimension were higher in cases where the mixed-format common item sets were used. Under all conditions, the *a* parameters of polytomous items were found to have higher in situations where mixed-format common item sets were used. In addition, for the threshold parameters with scale shrinkage and MHRM, the mixed-format common item structure elicited more errors. In the third dimension, the errors and bias values for mixed-format common items were lower except for in the *a* parameter of the polytomous items.

The scale shrinkage effect was examined as the next. For the first dimension, it was found that for the first dimension, the amount of error and bias obtained for the threshold parameter in the tests using dichotomous items was higher when the scale shrank. With the mixed-format common item structure, the MHRM estimation method and scale shrinkage, and the amount of error was lower for all the item parameters. For the second dimension, the bias amounts for the item parameters in the conditions using EM cycles and the only dichotomous common items were lower with no scale shrinkage. Similarly, the bias values of the ability parameters for datasets using mixed-format common items are also lower when the scale is not shrunk. In the third dimension, when EM cycles and dichotomous items were used, the error and bias amounts of the item parameters were generally lower in the cases of no scale shrinkage.

Regarding the estimation method, the error and bias values were lower with no scale shrinkage for parameters *a* and *b* of dichotomous items in the first dimension. On the other hand, with scale shrinkage, only the error values were lower in the EM estimation method. In addition, the error and bias values obtained from the *a* parameters of the polytomous items for the data in which only dichotomous items were included in the common item set were lower with the EM estimation method. These values for the second dimension showed similar changes to those in first dimension. Unlike for the first dimension, it was seen that, for this dimension, the bias values of the ability parameters were lower when the mixed-format item structure was used and there was no scale.Finally, the findings for the third dimension showed EM cycles produced lower error and bias values for all the item parameters with no scale shrinkage and a dichotomous common item structure was preferred.

Later, the 2 × 2 × 2 factorial ANOVA results were examined to see whether the observed differences in the bias and error values were significant, and whether there was an interaction between the conditions investigated. The results are presented in Table 2.

Regarding the interactions for the first dimension, there was a significant interaction between the CIF and SS conditions for the bias values of the ability parameters ($p < .05$). According to the analyses performed to test whether the levels of interaction of the CIF and EM conditions were meaningful, these two conditions interacted with the bias values of the threshold parameters of polytomous items ($p < .05$).

Table 1. Error and Bias Values Across the Conditions

		No Shrinkage				Shrinkage			
		EM		MHRM		EM		MHRM	
		Error	Bias	Error	Bias	Error	Bias	Error	Bias
	First Dimension								
Dich. common items	*θ*	0.068	-3.654	0.060	-3.294	0.060	-3.227	0.058	-3.431
	a param. (dich.)	0.074	-0.241	0.086	-0.340	0.069	-0.351	0.072	-0.337
	b param. (dich.)	0.624	2.911	0.641	2.992	0.604	2.817	0.669	3.119
	a param. (poly.)	0.132	-0.025	0.139	-0.036	0.136	-0.045	0.132	0.035
	b param. (poly.)	1.116	0.771	1.066	0.835	1.140	0.774	1.074	0.836
Mixed format common items	*θ*	0.034	-1,738	0.032	-1,736	0.032	-1.744	0.033	-1.964
	a param. (dich.)	0.050	0.158	0.050	0.104	0.045	0.158	0.052	0.101
	b param. (dich.)	0.457	2,226	0.475	2,315	0.433	2.111	0.489	2.384
	a param. (poly.)	0.123	0.034	0.129	0.023	0.121	0.035	0.130	0.019
	b param. (poly.)	1,308	0.512	1,189	0.498	1.246	0.547	1.233	0.496
	Second Dimension								
Dich. common items	*θ*	0.053	-2.858	0.049	-2.543	0.047	-2.568	0.050	-2.449
	a param. (dich.)	0.074	-0.244	0.084	-0.347	0.070	-0.351	0.087	-0.339
	b param. (dich.)	0.556	2.595	0.590	2.754	0.578	2.696	0.590	2.755
	a param. (poly.)	0.128	0.023	0.130	0.022	0.122	0.027	0.122	0.021
	b param. (poly.)	1.206	0.771	1.175	0.835	1.258	0.774	1.171	0.836
Mixed format common items	*θ*	0.030	-1,751	0.033	-1,798	0.034	-1.874	0.032	-1.909
	a param. (dich.)	0.050	0.174	0.052	0.121	0.046	0.172	0.052	0.116
	b param. (dich.)	0.528	2,576	0.543	2,644	0.491	2.392	0.552	2.689
	a param. (poly.)	0.139	-0.074	0.139	-0.052	0.142	-0.073	0.133	-0.053
	b param. (poly.)	1,242	0.512	1,175	0.498	1.194	0.547	1.188	0.496
	Third Dimension								
Dich. common items	*θ*	0.040	-2.344	0.039	-2.269	0.040	-2.168	0.034	-2.177
	a param. (dich.)	0.077	-0.250	0.087	-0.354	0.070	-0.361	0.080	-0.345
	b param. (dich.)	0.670	3.127	0.759	3.537	0.717	3.346	0.731	3.412
	a param. (poly.)	0.142	-0.037	0.150	-0.060	0.145	-0.059	0.140	-0.058
	b param. (poly.)	1.620	0.771	1.643	0.835	1.678	0.774	1.615	0.836
Mixed format common items	*θ*	0.016	-0.762	0.016	0.779	0.016	-0.701	0.016	-0.764
	a param. (dich.)	0.046	0.137	0.049	0.081	0.043	0.137	0.051	0.079
	b param. (dich.)	0.422	2,055	0.459	2,237	0.413	2.015	0.460	2.241
	a param. (poly.)	0.160	0.175	0.157	0.171	0.165	0.170	0.153	0.171
	b param. (poly.)	1,326	0.512	1,243	0.498	1.251	0.547	1.275	0.496

Finally, when the interactions between the three conditions were examined, it was found that there was a meaningful three-way interaction for the error values of the *a* parameters of the dichotomous items ($p < .05$). The second and third dimensions showed similar results. When all the results were considered en bloc, it could be seen that, in addition to a clear effect of the common item format, the estimation method had effect, at least, for some dimensions. Although, some interactions were observed, they did not come close to providing a meaningful picture.

Table 2: Factorial ANOVA Results

			Dichotomous			Polytomous	
	Conditions Being Tested		***θ***	***a***	***b***	***a***	***B***
First Dimension	Common Item Format (CIF)	RMSE	298.756**	99.88**	354.555**	9.947**	64.491**
		Bias	790.483**	312.124**	267.519**	113.801**	150.291**
	Scale Shrinkage (SS)	RMSE	3.963*	7.323**	0.001	0.103	0.039
		Bias	0.265	1.214	0.004	0.959	0.157
	Estimation Method (EM)	RMSE	1.620	2.783	18.726**	2.654	11.586**
		Bias	0.001	3.800	18.940**	1.542	0.412
	CIF*SS	RMSE	1.748	0.828	0.208	0.010	0.515
		Bias	4.150*	1.068	0.211	0.487	0.090
	CIF*EM	RMSE	2.633	1.714	0.056	1.046	0.049
		RMSE	1.953	0.121	0.016	1.170	4.051*
	SS*EM	Bias	0.841	0.054	5.577*	0.419	1.581
		RMSE	13.896**	1.167	5.617*	0.436	0.165
	CIF*SS*EM	Bias	0.443	4.341*	0.076	1.407	2.811
		RMSE	1.567	1.320	0.049	1.097	0.135
Second Dimension	Common Item Format (CIF)	RMSE	114.524**	98.007**	17.307**	11.491**	0.023
		Bias	197.916**	333.092**	4.709*	254.163	150.291**
	Scale Shrinkage (SS)	RMSE	0.955	1.476	0.019	1.244	0.032
		Bias	0.836	1.101	0.025	0.039	0.157
	Estimation Method (EM)	RMSE	3.629	4.941*	6.300*	0.140	6.532
		Bias	5.209*	3.818	6.417*	2.802	0.412
	CIF*SS	RMSE	0.178	0.666	1.066	0.503	1.256
		Bias	6.971**	0.840	1.079	0.034	0.090
	CIF*EM	RMSE	2.369	7.065**	0.345	0.492	0.366
		RMSE	3.020	0.027	0.401	4.888	4.051*
	SS*EM	Bias	3.555	3.778	0.272	0.377	0.007
		RMSE	2.282	1.206	0.315	0.041	0.165
	CIF*SS*EM	Bias	0.074	0.636	1.990	0.391	2.481
		RMSE	0.098	1.334	2.037	0.013	0.135
Third Dimension	Common Item Format (CIF)	RMSE	321.824**	115.478**	584.166**	12.644**	312.86**
		Bias	938.236**	302.728**	488.666**	2667.632**	150.291**
	Scale Shrinkage (SS)	RMSE	7.121**	3.017	0.079	0.179	0.023
		Bias	2.648	1.098	0.067	2.223	0.157
	Estimation Method (EM)	RMSE	6.009*	3.205	16.128**	0.343	1.419
		Bias	0.135	4.129*	16.029**	2.305	0.412
	CIF*SS	RMSE	1.225	0.566	0.360	0.157	0.795
		Bias	2.108	0.998	0.350	0.738	0.090
	CIF*EM	RMSE	2.401	3.281	0.159	1.428	0.050
		RMSE	0.515	0.062	0.096	1.164	4.051*
	SS*EM	Bias	0.049	1.816	1.964	1.742	0.066
		RMSE	0.510	1.410	1.857	2.787	0.165
	CIF*SS*EM	Bias	1.243	2.002	3.193	0.055	5.479*
		RMSE	0.240	1.448	3.074	1.255	0.135

p<0.05; p<0.01***

DISCUSSION

The findings showed that the common item structure significantly affected the amount of errors and bias obtained from the vertical scaling process. Specifically, for mixed-format tests, when the common item set only contained dichotomous items, it caused higher the amount of error, with few exceptions. As stated by Kolen and Brennan (2004), the common set of items needs to be a "mini version" of the total test in terms of content and statistical properties. This means that when polytomous items are placed in the common item set, the common item set becomes more similar to the total and this positively affects the scaling results.

Another finding suggests that the estimation methods provide similar amount of error in almost all conditions. The fact that parameter estimates performed with MHRM and EM cycles have no effect on the results of the scaling in many cases can be explained by the three dimensional data structure preferred in this study. Cai (2008, 2010a) reported that MHRM estimates yield better results for datasets with five or more dimensions, and that these two estimation methods give very similar results for datasets with a lower number of dimensional structures. As stated by Cai (2010a), when EM cycles are used, the number of quadrants required for estimates increases geometrically as the number of dimensions is linearly increased. There is no such requirement for MHRM estimates and it becomes more advantageous to use MHRM for higher dimensional data. As a result, these two estimation methods yielded similar results in the case of the three dimensional data scenario used in this study.

One interesting finding is that the scaling results yielded a high amount of bias that could be explained by the one-point θ difference between the groups which could rarely be observed in real life test applications. An effect of higher ability difference on bias value is reported in the literature (Kim and Lee, 2006). Indeed, many studies have confirmed that bias values gets higher when the differences real and estimated abilities between upper and lower groups increases (Brennan and Kolen, 2008; Kirkpatrick, 2005; Wang, Lee, Wu, Huang, Hu and Harris, 2009; Kim and Lee, 2006). On the other hand, Cao (2008) and Kirkpatrick (2005) stated that the effect of ability differences on measurement results is valid only in the context of the IRT. Furthermore, Hagge (2010) stated that when the difference of abilitiy among the groups is high, the bias value may be relatively low where the correlation between polytomous and dichotomous items is too high.

In the light of these findings, it is suggested that test developers should prefer that common item sets contain mixed-format items when vertical scaling is performed even if this involves some difficulties in practice, such as higher cost and a limited number of available polytomously scored items. Moreover, since this study was conducted by using simulation data, caution should be taken when making generalizations for testing applications. In future studies, it is suggested that the current study be replicated using real data.

REFERENCES

Baker, Frank B. (1992). *Item response theory: parameter estimation techniques*. New York: Marcel Dekker, Inc.

Bastari, B. (2000). *Linking multiple-choice and constructed-response items to a common proficiency scale* (Doctoral dissertation, University of Massachusetts Amherst). Retrieved from https://scholarworks.umass.edu/dissertations/AAI9960735

Birnbaum, A. (1968). Some latent trait models and their use in ınferring an examinee's ability. In FM Lord, MR Novick (eds.), *Statistical theories of mental test scores*, ss. 397-479. Addison-Wesley, Reading, MA.

Bock, R.D. & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, 46, 443–459

Cai, L. (2008). *A Metropolis-Hastings Robbins-Monro algorithm for maximum likelihood nonlinear latent structure analysis with a comprehensive measurement model (Doctoral dissertation*). Retrieved from https://cdr.lib.unc.edu.

Cai, L. (2010). High-dimensional exploratory item factor analysis by a Metropolis-Hastings Robbins-Monro algorithm. *Psychometrika*, 75, 33-57.

Camilli, G., Wang, M., & Fesq, J. (1995). The effects of dimensionality on equating the Law School Admission Test. *Journal of Educational Measurement*, 32, 79-96.

Cao, Y. (2008). *Mixed format test equating: Effects of test dimensionality and common-item sets* (Doctoral dissertation). Retrieved from https://drum.lib.umd.edu.

Cao, Y., Yin, P., & Gao, X. (2007). *Comparison of IRT and classical equating methods for tests consisting of polytomously-scored items*. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

Chalmers, R. P.(2012). mirt: A multidimensional ıtem response theory package for the r environment. *Journal of Statistical Software, 48*(6), 1-29.

Dodd, B. G. (1984). *Attitude scaling: A comparison of the graded response and partial credit latent trait models* (Doctoral dissertation, University of Texas at Austin). Retrieved from https://elibrary.ru/item.asp?id=7426395.

Haebara, T. (1980). Equating logistic ability scales by a weighted least squares method. *Japanese Psychological Research*, *22*(3), 144–149.
Hagge, S. L. (2010). *The impact of equating method and format representation of common items on the adequacy of mixed-format test equating using nonequivalent groups* (Doctoral dissertation), Retrieved from https://ir.uiowa.edu/etd/680/.
Han, K. T., & Paek I. (2014). A review of commercial software packages for multidimensional IRT modeling. *Appl. Psychol. Meas*. 38, 486–498.
Kim, J. (2007). "*A comparison of calibration methods and proficiency estimators for creating IRT vertical scales*." PhD (Doctor of Philosophy) thesis, University of Iowa.
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. H., & Cohen, A.S. (2002). A comparison of linking and concurrent calibration under the graded response model. *Applied Psychological Measurement*, *26*(1), 25–41.
Kirkpatrick, R. K. (2005). *The effects of item format in common item equating* (Doctoral dissertation, University of Iowa). Available from ProQuest Dissertations and Theses database. (UMI No. 3184724)
Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices*. NY: Springer.
Koretz, D.M., & Hamilton, L.S. (2006). Testing for accountability in K-12. In R.L. Brennan (Ed.), *Educational Measurement* (4th ed., 531-578). Westport, CT: American Council on Education and Praeger Publishers.
Kuo T-C & Sheng Y (2016) A comparison of estimation methods for a multi-unidimensional graded response ırt model. *Front. Psychol*. 7, 880.
Livingston, S. A. (2004). *Equating test scores (without IRT)*. Princeton, NJ: Educational Testing Service.
Marco, G. L., (1977). Item characteristic curve solutions to three intractable testing problems. *Journal of Educational Measurement*, 14, 139-160.
Muraki, E. & Carlson, E. B. (1995). Full-information factor analysis for polytomous item responses. *Applied Psychological Measurement*. 19, 73-90.
R Core Team (2015). *R: A language and environment for statistical computing. R Foundation for Statistical Computing*, Vienna, Austria.
Reckase, Mark D. (2009). *Multidimensional item response theory*. New York: Springer.
Rosa, K., Swygert, K., Nelson, L. & Thissen, D. (2001). Item response theory applied to combinations of multiple-choice and constructed response items: Scale scores for patterns of summed scores. D. Thissen and H. Wainer (Eds.), *Test scoring* (pp. 253-292). Hillsdale, NJ: Lawrence Erlbaum.
Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychomtrika Monograph*, No. 17.
Samejima, F. (1972). A general model for free-response data. *Psychometric Monograph*, No. 18.
Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement*, 7, 201-210.
Tate, R. (2000). Performance of a proposed method for the linking of mixed format tests with constructed response and multiple choice items. *Journal of Educational Measurement*. 37, 329-346.
Wang, T., Lee, W., Brennan, R. L., & 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*(8), 632-651.
Weeks, J. P. (2010) plink: An R package for linking mixed-format tests using IRT-based methods. *Journal of Statistical Software*, *35*(12), 1–33
Yao, L. ve Boughton, K. (2009). Multidimensional linking for tests with mixed ıtem types. *Journal of Educational Measurement*. *46*(2), 177–197.
Yen, W. M. (1985). Increasing item complexity: A possible cause of scale shrinkage for unidimensional item response theory. *Psychometrika*, 50, 399-410.
Zhang, M. (2016). *Exploring dimensionality of scores for mixed-format tests* (Doctoral dissertation). Retrieved from https://ir.uiowa.edu/cgi/viewcontent.cgi?article=6628&context=etd

Çok Boyutlu Karışık Format Testlerinin Dikey Ölçeklemesini Etkileyen Faktörlerin İncelenmesi

Giriş

Testlerden elde edilen puanlar birçok başlık altında alınan önemli kararlar için temel bilgi kaynakları arasındadır. Alınacak önemli kararlardan bağımsız olarak, test puanlarının mümkün olan en kesin

bilgiyi sunması gerekmektedir. Daha kesin bilgi daha iyi kararların alınabilmesi için önemlidir. Bununla birlikte uygulamada test güvenliği ve öğrenci gelişiminin takip edilebilmesi gibi birtakım gerekçeler yüzünden aynı testin farklı formları kullanılmakta veya farklı zamanlarda uygulanan testlerde ortak maddeler kullanılarak testler ölçeklenmektedir. Farklı formlardan elde edilen puanlar daha sonrasında eşitlenmekte ya da ölçeklenmektedir. Bu işlemin hatasız olması gerçekleştirilen sınavların daha adil olması ve öğrencilerin geleceği ile ilgili doğru kararlar verebilmek için önemlidir. Buna göre, puanları önemli kararlar için kullanılan testlere uygulanan dikey ölçekleme yöntemlerinin psikometrik olarak savunulabilir olması önemlidir. Bu sebepten dolayı ölçekleme gerçekleştirilirken uygulayıcıların kararlarını dayandıracakları kuramsal çalışmalar büyük önem taşımaktadır. Bu sebepten dolayı farklı yöntemlerin karşılaştırılması ve farklı durumlar için en az hata veren yöntemlerin belirlenmesi gerekmektedir.

İki kategorili ve çok kategorili olarak puanlanan maddelerin birlikte yer aldığı karma format testlerin kullanımı gün geçtikçe artmaktadır. Benzer şekilde, büyük ölçekli ve öğrencilerle ilgili önemli kararların alındığı test uygulamalarında birden fazla formunun kullanımı da benzer şekilde yaygınlaşma eğilimindedir. Farklı test formlarından elde edilen puanların karşılaştırılabilir olabilmesi için bu formlar arasında fonksiyonel bir bağ oluşturulması gerekmektedir. Eğer kurulan bu bağ farklı sınıf (ya da test güçlüğü farklılaşan) formlar arasında gerçekleştirilirse, bu işlem dikey ölçekleme olarak adlandırılmaktadır. Dikey ölçeklemede farklı test formları birbirlerine bağlandığı için eşitleme ile benzerdir. Fakat test formları içerik ve güçlük olarak farklıdır çünkü formlar sınıflar arası ya da yaşa bağlı olarak ilerlemeyi yansıtmaktadırlar. Bundan dolayı, dikey ölçekleme farklı test formlarının karşılaştırılması için kullanılmakla birlikte her bir seviyedeki puanlar birbirlerinin yerine kullanılamazlar. Test ölçeklemesinde temel amaç farklı seviyelerdeki puanların karşılaştırılmasıdır. Seviye farklılığı bir öğrencinin bulunduğu sınıf, eğitim öğretim yılının bulunduğu aşama ya da yaştan kaynaklanabilir. Dikey ölçekleme genellikle aynı bireylerin farklı seviyelerde elde ettikleri puanların farklı zamanlara göre karşılaştırılabilmesi için kullanılmaktadır. Bu tür desenler ise DOGOM (Denk Olmayan Gruplarda Ortak Madde) deseni olarak adlandırılmaktadır.

Bu çalışma kapsamında karma format maddelerden oluşan boyutlu testler DOGOM deseni kullanılarak ölçeklendiğinde ortak madde setinin yapısı (yalnızca iki kategorili maddelerden oluşan ortak madde seti - iki ve çok kategorili maddelerin yer aldığı ortak madde seti), yetenek daralması (üst yetenek grubunda yetenek varyansının daralması - varyansın eşit kalması) ve parametre kestirim yöntemlerinin (EM - MHRM) ölçekleme sonuçları üzerindeki etkisi incelenmiştir. Ayrıca bu koşulların etkileşim içinde olup olmadığına bakılmıştır.

Yöntem

Çalışma, türetilmiş veriler kullanılarak gerçekleştirilmiştir. Ölçeklemenin niteliğinin değerlendirilmesinde ölçme hatası ve yanlılık değerleri kullanılmıştır. Veriler türetilirken yanıt matrisleri, içerisinde İKM (iki kategorili madde) ve ÇKM(çok kategorili madde)'ler yer alacak şekilde oluşturulmuştur. İKM'ler için parametre kestirimi 3 parametreli modele (3PLM) göre, ÇKM'ler için ise aşamalı tepki modeline (ATM) göre gerçekleştirilmiştir. Veri türetme ve analizi sürecinde gerçekleştirilen işlem 50 defa tekrarlanmıştır. Ayrıca, araştırmada gerçekleştirilen veri türetme, testlerin kalibrasyonu ve ölçekleme işlemleri için R programı kullanılmıştır. Etkileşimleri incelemek için kullanılan iki ve üç yönlü analizler SPSS ile gerçekleştirilmiştir.

Bulgular ve Tartışma

Araştırmada sonucunda ortak madde yapısının ölçekleme işlemi sonucunda ortaya çıkan hata ve yanlılık miktarını önemli ölçüde etkilediği görülmüştür. Buna göre karma format testlerde ortak madde setinin sadece iki kategorili puanlanan maddelerden oluşması ölçekleme hatasını bazı istisnalar haricinde arttırmaktadır. Elde edilen bu bulgu, diğer koşullardan bağımsız olarak tutarlı bir şekilde gözlenmiştir.

Varyans daralmasının etkisi incelendiğinde yetenek parametresi ve çok kategorili puanlanan maddelerie ait *a* parametreleri için farklılaşmalar olduğu görülmüştür. Gözlenen bu farklılaşmalar yanlılık değerlerine aittir. Çok kategorili puanlanan maddelere ait *a* parametreleri için ise hata değerlerinde farklılaşmalar olduğu bulunmuştur. Her iki parametre için varyansın azaldığı durumda daha iyi sonuçlar elde edildiği görülmüştür.

Kullanılan kestirim yönteminin etkisi incelendiğinde ise bazı boyutlar için yanlılık değerlerinin Metropolis–Hastings Robbins-Monro kestirim yöntemi için daha az olduğu görülmüştür. Ayrıca iki kategorili puanlanan maddelerin *a* ve *b* parametreleri ve çok kategorili puanlanan maddelerin eşik parametreleri için bazı durumlarda kestirim yönteminin hata ve yanlılık değerlerini etkilediği görülmüştür. Çok kategorili puanlanan maddelerin *a* parametresinin ise kestirim yönteminden etkilenmediği görülmüştür.

Son olarak, etkileşimler incelenmiştir. Buna göre, yetenek parametresi bazı koşullara göre yanlılık değerlerinin ikişerli ve üçerli etkileşimler gösterdiği bulunmuştur. İki kategorili maddelere ait *a* ve *b* parametreleri için bakıldığında *b* parametresine ait hata ve yanlılık değerlerinde testin bazı boyutlarında varyans daralması ve kestirim yönteminin etkileşim içinde oldukları görülmüştür. İki kategorili puanlanan maddelere ait *a* parametrelerine ait hata değerleri için birinci boyutunda üç koşulun etkileşim içinde olduğu bulunmuştur. Ayrıca, çok kategorili puanlanan maddelere ait *a* parametreleri ile eşik parametreleri için etkileşim gözlenmemiştir. Üç boyutun tamamı için ortak madde yapısı ve kestirim yöntemi koşulları arasında etkileşim olduğu görülmüştür.

Sonuç olarak, etkisi incelenen koşullar içinde ölçekleme sonuçları üzerinde en fazla etkisi olan koşulun ortak madde yapısı olduğu sonucuna varılmıştır.