Eurasian Journal of Educational Research 79 (2019) 69-86

Eurasian Journal of Educational Research
www.ejer.com.tr

Investigation of the Orthogonality Assumption in the Bifactor Item Response Theory*

Fulya BARIS PEKMEZCI[1], H. Deniz GULLEROGLU[2]

ARTICLE INFO

Article History:
Received: 21 Jan. 2017
Received in revised form: 17 Aug. 2017
Accepted: 23 Oct. 2017
DOI: 10.14689/ejer.2019.79.4

Keywords
Multidimensional item response theory, Bifactor item response theory, Orthogonality assumption, confidence, Bias of parameter estimation, Factor analysis.

ABSTRACT

Purpose: This study aims to investigate the orthogonality assumption, which restricts the use of Bifactor item response theory under different conditions.

Method: Data of the study have been obtained in accordance with the Bifactor model. It has been produced in accordance with two different models (Model 1 and Model 2) in a simulated way.

Results: As a result of the research, it was found out that the case that two factors were correlated (Model 1) and that all factors were correlated (Model 2) had the same effect on the accuracy of both person and item parameter estimations. While estimating the discrimination parameters, as the orthogonality violation increased, it was concluded that the bias increased, too. As the test length increased, the accuracy of estimations of discrimination and difficulty parameters, namely the reliability decreased. Increasing the number of items increased the accuracy of person parameters, which was the reliability.

Implication for Research and Practice: As test length increases, the Bifactor theory can better tolerate the orthogonality violation in estimation of person parameters. The practitioners who want to use this theory are recommended to work with large item pools. At all correlation levels, the accuracy of the parameter estimations was approximately the same. New studies can be repeated with intermediate correlation levels. Among all the parameters, the parameters whose estimation reliability is the lowest were found to be person parameters.

* This study is based on a summary of the doctoral dissertation entitled "Investigation of Orthogonality Assumption in Bifactor Model Under Different Conditions" prepared in the Educational Measurement and Evaluation Program, Ankara University.

[1] Yozgat Bozok University, TURKEY. e-mail: fulyabaris@gmail.com, ORCID: https://orcid.org/0000-0001-6989-512X

[2] Ankara University, TURKEY. e-mail: denizgulleroglu@yahoo.com, ORCID: https://orcid.org/0000-0001-6995-8223

Introduction

Bifactor item response theory model was developed by Holzinger and Swineford (1937) as an extension of Spearman's Bifactor theory, as can be understood from his name. Bifactor theory assumes that there are more than one specific factor and a general factor explained by these factors, and that these specific effects also have an effect on the general factor (Spearman, 1904). As in all item response theory models, the Bifactor Model has its own assumptions. One of the assumptions of the Bifactor Model is that the data include both general and specific factors. The other assumption that the factors are orthogonal is not possible to be met in practice. In other words, test developers should write only the primary factor and also the items that measure a subdomain. The main problem is that writing such items in practice is very difficult.

According to Canivez (2016), the main advantages of the Bifactor Model are generally these: (a) the effect of the overall factor on each item and groups of items can be easily interpreted. This is not achievable with second-order models, correlated trait models, and uni-dimensional models (Chen, West & Sousa, 2006; Immekus & Imbrie, 2008); (b) the effects of both general and specific factors on the items can be estimated simultaneously (Reise, 2012; Reise, Moore & Haviland, 2010); (c) the psychometric properties that are required to score and interpret general and specific factors are obtainable through the Bifactor Model (DeMars, 2013); (d) the specific effects of general and specific traits in describing other variables are obtained more accurately; and (e) the Bifactor Model provides more accurate and reliable estimations than testlet-effect model in estimating item and person parameters.

Bifactor Model is very common in scaling the psychological properties, and differentiates the specific contributions of the facets on the general factor very well. Therefore, the Bifactor Model is quite suitable for scale development. While developing or evaluating a new multifaceted scale that aims to assess the general structure and specific facets, the power of factor loadings at general and specific factors will be a guide in choosing and evaluating items. The items will ideally have a higher loading at the general factor or at least a greater loading than the specific factor. If the items have a higher loading than the facets in the general structure, these items will be selected, however, if specific factors have larger loadings than the general factor, these items will be removed from the scale. The reason for this is that these items do not contribute significantly to the general structure. Moreover, the Bifactor Model is also used to create a uni-dimensional scale or a short uni-dimensional scale from a multidimensional scale (Stucky & Edelen, 2014; Stucky, Edelen, Vaughan, Tucker & Butler, 2014; Stucky, Thissen & Edelen, 2013). The applications of the Bifactor Model in education indicate that this model is useful in terms of scoring the subscales and assessing the reliability when subscale scores need to be used (Cucina & Byle, 2017; DeMars ,2013; Golay & Lecerf, 2011; Watkins & Beaujea, 2014).

In addition to these advantages, the Bifactor Model has also some limitations. The biggest limitation is the difficulty of meeting the orthogonality assumption of the Bifactor model (Chen, West & Sousa, 2006; Simms, Grös, Watson & O'Hara., 2008). As in the structural equation model, the Bifactor model needs a considerably larger

sample when compared to the total score and individual score approach. Additionally, the Bifactor model interpretations become quite complicated when correlations are allowed among specific factors (Rindskopf & Rose, 1988), and the model can often not be identified. In addition to these, it also gives a weak model adaptation in weak or small factor loadings as in other factorial models (Jennrich & Bentler, 2012; MacCallum, Widaman, Zhang & Hong, 1999).

When the literature about Bifactor Models is reviewed, it has been seen that the focus has always been on the determination of dimensionality and the examination of item performance in the field of education and psychology, algorithm of the Bifactor Model, and comparison of different item response theory models with Bifactor model (Brouwer, Meijer, Weekers & Baneke, 2008; Brown, Finney & France, 2011; Chen, West & Sousa, 2006; Chen, Hayes, Carver, Laurenceau & Zhang, 2012; Demars, 2006; Fukuhara, 2009; Garn, 2017; Gibbons et al., 2007; Hyland, Boduszek, Dhingra, Shevlin & Egan, 2014; Lafond, 2014; Martel, Von Eye & Nigg, 2010; Reise, Ventura et.al, 2011; Rijmen, 2009; Rodriguez, Reise & Haviland, 2016; Thomas, 2012;Yang, Song & Xu, 2002).

Although the situation that limits the use of the Bifactor Model is the orthogonality assumption, there has been only one study (Zheng, 2013) carried out in the field about testing of the orthogonality assumption under different conditions. This work (Zheng, 2013) has also been carried out under limited conditions in such a way in every simulation study. Contrary to Zheng's (2013) study, in this study, simulation conditions (test length and correlation levels) were changed. Moreover, item and person parameters were estimated according to Bayesian approach (by Quasi Monte-Carlo estimation). In addition to Zengh's (2013) study, Rindskopf and Rose (1988) found that the interpretation of model parameters gets complicated as correlations among specific factors are allowed in the Bifactor Model. Since cross loadings between factors will also allow correlations between factors, this can be considered as a kind of correlation between factors. Rindskopf and Rose (1988) could not reach any information about level of these cross loadings.

Purpose

The Bifactor Model is a theory that is limited in its use due to the orthogonality assumption that it requires. In addition to this limitation, this model is frequently used in studies of modeling psychological and educational constructs, and developing scales by ignoring the assumption. In cases where the orthogonality assumption is not met, it is not going to be possible to model psychological and educational constructs accurately for the developed scale to reach a correct factorial structure and to have correct parameter estimations. Besides, it is almost impossible to develop measurement instruments in which the correlation between factors in the fields of education and psychology is zero. Forcing the correlated factors to be orthogonal will cause loss of information regarding the measured structure, and will result in unreliable parameter estimations. The precision and the accuracy of parameter estimations, on the other hand, are important in every measurement because parameter estimations are an important element in determining item performance and

respondents' ability level. Resulting from all these reasons, it is necessary to examine the Bifactor model by allowing different correlations among specific factors, in other words, to determine if stable, precise and accurate estimations can be done despite orthogonality violation by which levels of violation are tolerated by the theory itself. It is thought that via this research, the results that are going to be obtained through examining and evaluating the orthogonality assumption that restricts the bifactor model usage under certain criteria will highly contribute to the field.

Method

Research Design

This research is based on the basic research model since it is carried out through the data obtained by Monte Carlo simulation in order to investigate the effect of the violation of the orthogonality assumption at different levels and test lengths on the item and person parameter estimation.

Simulation Study

The data for this study were generated according to two Bifactor two-parameter models with a simulation according to two models (Model 1 and Model 2). Model 1 was the model which showed the violation of orthogonality due to the cross loadings. In this model, the focus was on the effect of orthogonality violation between two specific factors on parameter estimations in all factors. On the other hand, Model 2 showed the correlations among all the specific factors.

The variables that were manipulated in specific models were the correlation levels between factors and the test lengths. The correlation acceptance levels for the models that were set up (Model 1 and Model 2) were as 0.10 (very low), 0.40 (medium), 0.70 (high) (Cohen, 1988). In the framework of this research, it was decided that the minimum test length to be 12 items with reference to the fact that a factor should have at least three items in order to be called as a factor (Kline, 1994). Different from the literature on the field, other test lengths were taken as 40 and 100, taking into consideration that the number of items in each factor was equal.

The variable to be kept constant, namely not to be manipulated, during the research was the sample size. In order to prevent bias that would arise from sample size, the largest sample size (5000) that were used in the current studies was set as a simulation sample. The summary of the research design is given in Table 1.

Table 1

The Summary of the Research Design

	Correlation Levels	Test Length	Sample Size
Model 1			
Model 1.1	$r_{3.4}$=0.10	12-40-100	5000
Model 1.2	$r_{3.4}$=0.40	12-40-100	5000
Model 1.3	$r_{3.4}$=0.70	12-40-100	5000
Model 2			
Model 2.1	$r_{2:4\leq}0.10$	12-40-100	5000
Model 2.2	$0.10 < r_{2:4} \leq 0.40$	12-40-100	5000
Model 2.3	$0.40 < r_{2:4\leq}0.70$	12-40-100	5000

As a result of the literature review, it has been seen that the replication numbers used in the Bifactor Models are generally 100 (Demars, 2006; Zhang, 2008), 200 (Zheng, 2013) and 500 (Cai, Yang & Hansen, 2011). In this study, number of replications was determined as 200 to be practical.

In order to generate the two-parameter Bifactor model data set with the determined number of replications, the distribution of the discrimination parameter (a), the difficulty parameter (b) and the person parameter (θ) should be determined. A simulation model in which the discrimination parameter (a) was uniformly distributed between a range of 0.2 to 2.0 ratios, the difficulty parameter (b) and the person parameter (θ) that were randomly distributed were set. The mathematical expression of the Bifactor two-parameter model was as follows:

$$p(y=1 \mid \theta_g, \theta_s)=\frac{1}{1+ exp\{-(d+a_g\theta_g+a_s\theta_s)\}}$$

The distribution characteristics of the discrimination, difficulty and person parameters were the same for the 18 (2x3x3) condition given in Table 1. A random seed was assigned to the true parameters, which were generated for the first condition and in other conditions, and via this seed, invariance of true parameters between models was provided. The difficulty coefficient that had been produced was transformed into a multidimensional difficulty coefficient by the following formula:

$$d = -b\sqrt{a_g^2 + a_s^2}$$

This study was carried out based on Monte Carlo method using R 3.4.0 GUI software with syntax (Zheng, 2013), which was written to simulate the data according to the determined conditions and to produce Bifactor model parameters. For the accuracy of the generated syntax and the generated data files, the average bias was calculated on a model that did not contain orthogonality violation, and it was observed that the bias average was close to zero.

The Data Analysis of Simulation

Bifactor model predictions were made with the data that were generated in the simulation, and 200 for each condition with a total of 3600 (18x200) data files were obtained. Bifactor model estimations were made with "mirt" (Chalmers, 2016) package in R 3.4.0 GUI software, and descriptive statistics were generated with "psych" (Revelle, 2017) package.

The evaluation of the accuracy of parameter estimations throughout the replications was carried out via mean bias, root mean square error (RMSE), and standard error of estimates (SE).

$$\text{Average Bias } (\hat{\beta}) = \frac{\sum_{r=1}^{R}(\hat{\beta}_r - \beta)}{R}$$

$$\text{RMSE}(\hat{\beta}) = \sqrt{\frac{1}{R}\sum_{r=1}^{R}(\hat{\beta}_r - \beta)^2}$$

$$\text{SE}(\widehat{\beta}) = \sqrt{\frac{1}{R}\sum_{r=1}^{R}(\hat{\beta}_r - \frac{\sum_{r=1}^{R}\hat{\beta}_r}{R})^2}$$

Given in the above formulas;

β: true individual parameter or item parameter

β̂: the individual and substance parameters predicted at the rth replication (Li & Rupp, 2011).

Results

Parameter Estimation Bias

The average bias values calculated from the files that were obtained from 200 replications for the models (Model 1 and Model 2), which were set up, are given in Table 2. In order to examine the recovery in the item parameter estimations, the bias was calculated by taking the average of the difference between the true parameters and the estimated parameters.

When the parameter estimation bias given in Table 2 was examined, the pattern seen for Model 1 and Model 2 was the same in all test lengths for discrimination parameters. When the test length increased from 40 items to 100 items, a decrease in the average bias was observed. Contrary to this, as the test length increaseds, the standard deviation of the bias scores got larger and the range widened. To put it in other words, the increase in the number of items led to a decrease in the reliability of the estimations. This can be explained by the increase in the amount of biased items. That is, the more correlated item was added to the model, the greater the variability got. When the models were examined among within themselves, the standard deviation values increased as the correlation between the factors increased with regard that the mean deviation did not change significantly. The average bias of the item and person parameters is given in Table 2.

Table 2

Mean Bias of Items and Person Parameters

			Mean Bias		
			12 items	*40 items*	*100 items*
			$(\overline{X}, \sigma)$	$(\overline{X}, \sigma)$	$(\overline{X}, \sigma)$
Discrimination parameter	Model 1	Model 1.1	0.010(0.050)	-0.040(0.650)	0.004(0.690)
		Model 1.2	0.002(0.120)	-0.030(0.650)	-0.009(0.690)
		Model 1.3	-0.020(0.310)	-0.030(0.650)	0.002 (0.700)
	Model 2	Model 2.1	0.020(0.070)	-0.050(0.650)	0.010(0.690)
		Model 2.2	0.000(0.180)	-0.030(0.650)	0.000(0.690)
		Model 2.3	-0.001(0.370)	-0.100(0.680)	-0.005(0.720)
Difficulty parameter	Model 1	Model 1.1.	0.260(0.330)	0.049(1.780)	-0.008(1.830)
		Model 1.2	0.240(0.340)	0.083(1.780)	-0.066(1.830)
		Model 1.3	0.260(0.330)	0.058(1.780)	-0.027(1.830)
	Model 2	Model 2.1	0.250(0.330)	0.020(1.780)	-0.040(1.830)
		Model 2.2	0.250(0.340)	0.050(1.780)	-0.060(1.830)
		Model 2.3	0.250(0.340)	0.040(1.780)	-0.060(1.830)
Person parameter	Model 1	Model 1.1	-0.010(0.670)	0.000(0.410)	0.010 (0.320)
		Model 1.2	0.000(0.690)	-0.010(0.440)	0.010 (0.340)
		Model 1.3	-0.010(0.710)	0.000(0.470)	-0.010(0.360)
	Model 2	Model 2.1	0.000(0.690)	0.010(0.410)	0.000 (0.280)
		Model 2.2	-0.001(0.720)	0.000(0.460)	0.000 (0.320)
		Model 2.3	0.000(0.800)	0.000(0.560)	0.020 (0.410)

As it can be seen in Table 2, when the estimation bias of the intercept coefficients was examined, as the test length increased for Model 1, the average bias scores decreased. The increase in the test length for Model 1 affected the parameter estimate recovery. This was not observed evidently when the test length for Model 2 was increased from 40 to 100 items. The standard deviation values increased as the test length increased, in other words the variability increased. The fact that the variability increased the reliability of estimations were reduced. At test lengths of 40 and 100 items, the greatest standard deviation values were observed on the difficulty coefficients. When the models were examined within themselves, although there was not much change in the average of bias, the standard deviations were almost the same. When all test lengths (12, 40, 100) for both Model 1 and Model 2 were examined all together, when the estimation of person parameters were examined, the distorted parameters were found to be at the test length of 12 items. Generally, the variability of bias scores was high at all test lengths. When Model 1 and Model 2 were compared, it was observed that the standard deviations were quite similar. It was observed that as the test length increased, the variability decreased for both Model 1 and Model 2. It can be said that the test length has an effect on the recovery in parameter estimations.

Increasing the test length reduced the variability. This finding is consistent with Zheng's (2013) study. Estimation accuracy was higher at the test lengths of 40 and 100 items when compared to the test length of 12 items. Directly proportional to the test length, the fact that the variability decreased indicated that the test length might have an effect on the recovery of parameters. However, the variability was high at all test lengths, and this reduced the reliability of the parameter estimation.

The Accuracy and Stability of Parameter Estimation

Estimation accuracy and stability of discrimination parameters. Table 3 shows the standard error values and average RMSE values of the discrimination parameters for Model 1 and Model 2. These values are first interpreted by model type, and then by the test length.

As it can be seen in Table 3, when the standard errors on the model basis were examined, it was observed that the table values (average value and standard deviation) were the same for both models (except the 12 item) ($\overline{X}_{SE\ \&Model\ 1}$= 0.083 , $\overline{X}_{SE\&Model\ 2}$= 0.086). When the RMSE averages were examined, it was observed that the table values for Model 1 and Model 2 were very close ($\overline{X}_{RMSE\ \&Model\ 1}$= 0.410 , $\overline{X}_{RMSE\ \&Model\ 2}$= 0.435)

Table 3

Discrimination Coefficients, Standard Error and Mean RMSE Values for Model 1 and Model 2

		Test Length		
	Model	12 items	40 items	100 items
SE	Model 1.1	0.120(0.090)	0.060(0.020)	0.050(0.010)
	Model 1.2	0.140(0.120)	0.060(0.020)	0.050(0.010)
	Model 1.3	0.160(0.150)	0.060(0.020)	0.050(0.010)
	Model 2.1	0.130(0.110)	0.060(0.020)	0.050(0.010)
	Model 2.2	0.140(0.130)	0.060(0.020)	0.050(0.010)
	Model 2.3	0.180(0.160)	0.060(0.020)	0.050(0.010)
	Model	12 items	40 items	100 items
RMSE	Model 1.1	0.140(0.120)	0.480(0.440)	0.540(0.430)
	Model 1.2	0.180(0.120)	0.480(0.440)	0.540(0.430)
	Model 1.3	0.290(0.230)	0.490(0.450)	0.550(0.430)
	Model 2.1	0.150(0.110)	0.480(0.440)	0.540(0.430)
	Model 2.2	0.220(0.140)	0.490(0.430)	0.540(0.440)
	Model 2.3	0.390(0.180)	0.540(0.430)	0.570(0.440)

According to these findings; it can be said that the fact that the two factors were correlated (Model-1) and all factors were correlated (Model-2) had almost the same effect in estimating the discrimination parameters. Consequently, there was no difference in the accuracy of parameter estimations for both models (Model 1 and Model 2). It can be concluded from this that the model parameter dis not have an influence on the accuracy of the parameter estimation.

When the RMSE values were examined according to the test length, it was considered that the test length might have an influence on the accuracy of the parameter estimation. The average RMSE values increased as the test length increased ($\overline{X}_{RMSE\&12}$= 0.228, $\overline{X}_{RMSE\&40}$= 0.493, $\overline{X}_{RMSE\&100}$= 0.546). Namely, as the number of items increased, the accuracy of the discrimination parameters decreased. When the standard errors were examined, it was observed that the standard error decreased as the test length increased ($\overline{X}_{SE\&12}$= 0.145, $\overline{X}_{SE\&40}$= 0.060, $\overline{X}_{SE\&100}$= 0.050). The standard error is the standard deviation of the simulation samples, in other words, a distance measure. Because of this, the standard error is actually a measure of precision (Walther &Moore, 2005). In this case, it can be said that as the test length increased, the estimations of the discrimination parameters were more reliable, that is, the test lengths might influence the estimation accuracy of the discrimination parameters.

Estimation accuracy and stability of difficulty parameters. Table 4 shows the standard error averages and the average RMSE values for Model 1 and Model 2 of the difficulty parameters in an order. When the standard errors of the models were analyzed, it was observed that there was not much difference between the table values (average and standard deviation) ($\overline{X}_{SE\&Model\ 1}$= 0.046 , $\overline{X}_{SE\&Model\ 2}$= 0.047). When the RMSE averages were studied, it was concluded that the condition for the standard error was also observed here. Table values were the same for Model 1 and Model 2 ($\overline{X}_{RMSE\&Model\ 1}$= 1.056 , $\overline{X}_{RMSE\&Model\ 2}$= 1.056).

According to these findings, it can be said that in the estimation of the difficulty parameter, the fact that two specific factors were related and that all specific factors were related had almost the same influence. To put it in other words, it can be said that model type did not affect the difficulty of parameter estimation. Table 4 presents the standard error and the average RMSE values with difficulty coefficients for Model 1 and Model 2.

As it can be seen in Table 4, it was observed that the standard error averages of Model 1.1, Model 1.2, and Model 1.3 did not differ too much when the models were examined within themselves (according to the degree of the orthogonality violation) ($\overline{X}_{SE\&Model1.1}$= 0.043 , $\overline{X}_{SE\&Model\ 1.2}$= 0.046, $\overline{X}_{SE\&Model1.3}$= 0.050). The same was observed for RMSE averages, too ($\overline{X}_{RMSE\&Model1.1}$= 1.053 , $\overline{X}_{RMSE\&Model\ 1.2}$= 1.060, $\overline{X}_{RMSE\&Model1.3}$= 1.056). When the model was examined for sub-models, it was observed that the standard error averages of Model 2.1, Model 2.2, Model 2.3 did not vary much ($\overline{X}_{SE\&Model\ 2.1}$= 0.046 , $\overline{X}_{SE\&Model\ 2.2}$= 0.046, $\overline{X}_{SE\&Model\ 2.3}$= 0.050). The same was observed for RMSE averages, too ($\overline{X}_{RMSE\&Model\ 2.1}$= 1.053 , $\overline{X}_{RMS\&EModel\ 2.2}$= 1.060, $\overline{X}_{RMSE\&Model\ 2.3}$= 1.056). The level of the orthogonality violation did not affect the estimation of the difficulty parameters. This finding overlaps with the study of Zheng (2013).

Table 4
Difficulty Coefficients Standard Error and Mean RMSE Values for Model 1 and Model 2

		Test Lengths		
	Model	12 items	40 items	100 items
SE	Model 1.1	0.050(0.030)	0.040(0.010)	0.040(0.010)
	Model 1.2	0.060(0.040)	0.040(0.010)	0.040(0.010)
	Model 1.3	0.070(0.050)	0.040(0.010)	0.040(0.010)
	Model 2.1	0.060(0.040)	0.040(0.010)	0.040(0.010)
	Model 2.2	0.060(0.050)	0.040(0.010)	0.040(0.010)
	Model 2.3	0.070(0.050)	0.040(0.010)	0.040(0.010)
	Model	12 items	40 items	100 items
RMSE	Model 1.1	0.330(0.260)	1.400(1.060)	1.430(1.130)
	Model 1.2	0.330(0.260)	1.420(1.060)	1.430(1.140)
	Model 1.3	0.330(0.270)	1.410(1.060)	1.430(1.140)
	Model 2.1	0.330(0.250)	1.400(1.070)	1.430(1.130)
	Model 2.2	0.340(0.260)	1.410(1.060)	1.430(1.130)
	Model 2.3	0.330(0.260)	1.410(1.060)	1.430(1.140)

When the RMSE values were examined according to the test length, it was observed that when the test length increased from 12 to 40, the estimation accuracy decreased, but when it increased from 40 to 100, this situation did not vary much ($\overline{X}_{RMSE\&12}$= 0.616, $\overline{X}_{RMSE\&40}$= 1.408, $\overline{X}_{RMSE\&100}$= 1.430). Contrary to study of Zheng (2013), when the accuracy of the estimations among the parameters in the framework of this study was taken into consideration, the difficulty parameters were the lowest parameters in terms of the test length and the model type. As Jennrich and Bentler (2012) pointed out in their research, when the correlation between factors was allowed, the results couldn't be interpreted. As the test length increased, the standard error values increased ($\overline{X}_{SE\&12}$= 0.310, $\overline{X}_{SE\&40}$= 0.493, $\overline{X}_{RMSE\&100}$= 0.546). The increase in the standard error indicated that the estimation accuracy decreased as the test length increased.

The estimation accuracy and stability of person parameters. Table 5 shows the standard error and RMSE averages for Model 1 and Model 2 of person parameters in an order. When the standard error values of the models were studied, it was observed that the table values (average and standard deviations) were similar ($\overline{X}_{SE\&Model\ 1}$= 0.464, $\overline{X}_{SE\&Model\ 2}$= 0.400). Table 5 shows the standard error and RMSE values for Model 1 and Model 2 of the person parameters.

Table 5
Standard Error and RMSE Values for Model 1 and Model 2 of The Person Parameters

		Test Length		
	Model	12 items	40 items	100 items
SE	Model 1.1	0.450(0.080)	0.490(0.130)	0.490(0.170)
	Model 1.2	0.450(0.080)	0.480(0.130)	0.500(0.170)
	Model 1.3	0.450(0.080)	0.480(0.130)	0.490(0.170)
	Model 2.1	0.420(0.080)	0.420(0.080)	0.360(0.070)
	Model 2.2	0.410(0.070)	0.420(0.090)	0.370(0.070)
	Model 2.3	0.380(0.070)	0.420(0.100)	0.390(0.100)
	Model	12 items	40 items	100 items
RMSE	Model 1.1	0.750(0.320)	0.620(0.210)	0.580(0.200)
	Model 1.2	0.760(0.340)	0.630(0.220)	0.590(0.200)
	Model 1.3	0.770(0.350)	0.640(0.230)	0.600(0.210)
	Model 2.1	0.740(0.340)	0.560(0.200)	0.450(0.140)
	Model 2.2	0.750(0.360)	0.590(0.220)	0.470(0.160)
	Model 2.3	0.790(0.410)	0.650(0.270)	0.540(0.210)

As it can be seen in Table 5, when the RMSE averages were to be examined, the situation for the standard error also appeared here. Table values for Model 1 and Model 2 were almost the same ($\overline{X}_{RMSE\&Model\ 1}$= 0.660 , $\overline{X}_{RMSE\&Model\ 2}$= 0.615). As a result, it can be said that the fact that two factors were correlated and that all specific factors were correlated had almost the same effect in parameter estimations.

When the models were analyzed within themselves (according to the degree of the orthogonality violation), it was observed that the standard error averages of Model 1.1, Model 1.2, Model 1.3 did not vary much ($\overline{X}_{SE\&Model1.1}$= 0.477 , $\overline{X}_{SE\&Model\ 1.2}$= 0.477, $\overline{X}_{SE\&Model1.3}$= 0.473). The same was also observed for RMSE averages ($\overline{X}_{RMSE\&Model1.1}$= 0.650 , $\overline{X}_{RMSE\&Model\ 1.2}$= 0.660, $\overline{X}_{RMSE\&Model1.3}$= 0.670). When the model 2 was examined for sub-models, it was seen that the standard error averages of Model 2.1, Model 2.2, Model 2.3 did not vary much ($\overline{X}_{SE\&Model\ 2.1}$= 0.400 , $\overline{X}_{SE\&Model\ 2.2}$= 0.400, $\overline{X}_{SE\&Model\ 2.3}$= 0.397). The same was also observed for RMSE averages ($\overline{X}_{RMSE\&Model\ 2.1}$= 0.583 , $\overline{X}_{RMSE\&Model\ 2.2}$= 0.603, $\overline{X}_{RMSE\&Model\ 2.3}$= 0.660).

Discussion, Conclusion and Recommendations

This research aims to analyze the effect of the Bifactor item response theory on the item and person parameter estimation under various conditions of the orthogonality assumption violation. As a result of the analyses made for this purpose, the estimation bias of the discrimination parameters for Model 1 increased as the orthogonality violation increased. The increase in test length caused a decrease in the accuracy of the

discrimination and difficulty parameters, in other words the reliability. This can be explained by the increase in the number of correlated items in specific factors. In the estimations of the discrimination parameters, an improvement in parameter estimations was observed with regard to the test length when two factors were related (Model 1), whereas this improvement was not observed when all specific factors were related (Model 2). The parameters whose estimation accuracy was the lowest were the difficulty parameters. It was observed that the model did not have an effect on the estimation accuracy of discrimination, difficulty, and person parameters. To put it in a different way, the case that two factors were correlated (Model 1) and that all specific factors were correlated (Model 2) had the same effect on the accuracy of both the person and item parameters.

Increasing the number of items increased the reliability of the estimations of person parameters. This situation observed in the person parameters was a consequence of the better explanation of the latent trait of individuals as the number of items increased. In estimations of person parameters, the least reliable parameter estimations were at the smallest test length for both models (Model 1 and Model 2). As the test length increased, the reliability of the estimations increased, too. Despite this, among the other parameters, the person parameters whose estimation reliability was the lowest at the all test lengths and the orthogonality violation levels.

The estimation of item and person parameters is an important factor in psychological and educational evaluations. The use of the Bifactor model in correlated structures will lead to biased parameter estimations, and this bias in parameter estimations will lead to bias in evaluation. The researches in the literature suggest that the Bifactor model is a very robust model that is well adapted even to the correlated structures. However, in this research when the parameter bias was examined, this robust structure could not be seen at all.

Based on the results of this study, some suggestions can be made for the practitioners or the researchers in the application of the Bifactor model. As test length increases, the Bifactor theory can better tolerate the orthogonality violation in estimation of person parameters. The practitioners who want to use this theory are recommended to work with large item pools. At all correlation levels, the accuracy of the parameter estimations was approximately the same. New studies can be repeated with intermediate correlation levels (0.25, 0.35, etc.). It is stated in the literature that there must be at least 20 items for multidimensional item response theory models. The minimum test length in this study was determined as 12 items. To obtain more unbiased results in the estimation of item parameters, determining the minimum test length as 20 items in future studies can retry the same conditions. Among all the parameters, the parameters whose estimation reliability is the lowest (highest SE averages) were found to be person parameters. Future researches can be tested with different replication numbers and different sample sizes to increase the reliability. This is only a simulation study, and is valid for the specified conditions.

References

Brouwer, D., Meijer, R. R., Weekers, A. M., & Baneke, J. J. (2008). On the dimensionality of the Dispositional Hope Scale. *Psychological Assessment*, *20*(3), 310.

Brown, A. R., Finney, S. J., & France, M. K. (2011). Using the bifactor model to assess the dimensionality of the Hong Psychological Reactance Scale. *Educational and Psychological Measurement*, *71*(1), 170-185.

Cai, L., Yang, J. S., & Hansen, M. (2011). Generalized full-information item bifactor analysis. *Psychological Methods*, 16(3), 221–248.

Canivez, G. L. (2016). Bifactor modeling in construct validation of multifactored tests: Implications for understanding multidimensional constructs and test interpretation. *Principles and Methods of Test Construction: Standards and Recent Advancements*. Gottingen, Germany: Hogrefe Publishers.

Chalmers, P. (2016). Mirt: Multidimensional item response theory. R package version 1.19,URL: https://cran.r-project.org/web/packages/mirt/index.html

Chen, F. F., West, S. G., & Sousa, K. H. (2006). A comparison of bifactor and second-order models of quality of life. *Multivariate Behavioral Research*, *41*(2), 189-225.

Chen, F. F., Hayes, A., Carver, C. S., Laurenceau, J. P., & Zhang, Z. (2012). Modeling general and specific variance in multifaceted constructs: A comparison of the bifactor model to other approaches. *Journal of Personality*, *80*(1), 219-251.

Cohen, J. (1988). *Statistical power analysis for the behavioral science (2nd ed.)*. Hillside, NJ: L. Erlbaum Associates.

Cucina, J., & Byle, K. (2017). The bifactor model fits better than the higher-order model in more than 90% of comparisons for mental abilities test batteries. *Journal of Intelligence*, *5*(3), 27.

DeMars, C. E. (2006). Application of the Bi-Factor multidimensional item response theory model to testlet-based tests. *Journal of Educational Measurement*, *43*(2), 145-168.

DeMars, C. E. (2013). A tutorial on interpreting Bifactor model scores. *International Journal of Testing*, *13*(4), 354-378.

Fukuhara, H. (2009). *A Differential Item Functioning Model for Testlet-Based Items Using A Bi-Factor Multidimensional Item Response Theory Model: A Bayesian Approach*. The Florida State University.

Garn, A. C. (2017). Multidimensional measurement of situational interest in physical education: Application of Bifactor exploratory structural equation modeling. *Journal of Teaching in Physical Education*, *36*(3), 323-339.

Gibbons, R. D., & Hedeker, D. R. (1992). Full-information Item Bi-factor analysis. *Psychometrika*, *57*(3), 423-436.

Gibbons, R. D., Bock, R. D., Hedeker, D., Weiss, D. J., Segawa, E., Bhaumik, D. K., ... & Stover, A. (2007). Full-Information item bifactor analysis of graded response

data. *Applied Psychological Measurement*, *31*(1), 4-19.

Golay, P., & Lecerf, T. (2011). Orthogonal higher order structure and confirmatory factor analysis of the french wechsler adult intelligence scale (WAIS-III). *Psychological Assessment*, *23*(1), 143.

Hyland, P., Boduszek, D., Dhingra, K., Shevlin, M., & Egan, A. (2014). A Bifactor approach to modelling the Rosenberg Self Esteem Scale. *Personality and Individual Differences*, *66*, 188-192.

Holzinger, K. J., & Swineford, F. (1937). The Bi-Factor method. *Psychometrika*, *2*(1), 41-54.

Immekus, J. C., & Imbrie, P. K. (2008). Dimensionality assessment using the full-information item Bifactor analysis for graded response data: An illustration with The State Metacognitive Inventory. *Educational and Psychological Measurement*, *68*(4), 695-709.

Jennrich, R. I., & Bentler, P. M. (2012). Exploratory Bi-Factor analysis: The oblique case. *Psychometrika*, *77*(3), 442-454.

Kline, P. (1994). *An easy guide to factor analysis*. Routledge.

LaFond, L. J. (2014). *Decision consistency and accuracy indices for the Bifactor and Testlet response theory models*. The University of Iowa. UMI Number: 3638391.

Li, Y., & Rupp, A. A. (2011). Performance of The S−X2 Statistic for full-information Bifactor models. *Educational and Psychological Measurement*, *71*(6), 986-1005.

MacCallum, R. C., Widaman, K. F., Zhang, S., & Hong, S. (1999). Sample size in factor analysis. *Psychological methods*, *4*(1), 84.

Martel, M. M., Von Eye, A., & Nigg, J. T. (2010). Revisiting the latent structure of ADHD: is there A 'G' factor?. *Journal of Child Psychology and Psychiatry*, *51*(8), 905-914.

Reise, S. P. (2012). The rediscovery of Bifactor measurement models. *Multivariate Behavioral Research*, *47*(5), 667-696.

Reise, S. P., Moore, T. M., & Haviland, M. G. (2010). Bifactor models and rotations: Exploring the extent to which multidimensional data Yield Univocal Scale Scores. *Journal of personality assessment*, *92*(6), 544-559.

Reise, S. P., Ventura, J., Keefe, R. S., Baade, L. E., Gold, J. M., Green, M. F., ... & Bilder, R. (2011). Bifactor and item response theory analyses of interviewer report scales of cognitive impairment in Schizophrenia. *Psychological Assessment*, *23*(1), 245.

Revelle, W. (2017). Psych: Procedures for psychological, psychometric, and personality research. R package version 1.7.5. URL: https://cran.rproject.org/web/packages/psych/index.html

Rijmen, F. (2009). Three multidimensional models for testlet-based tests: Formal relations and an empirical comparison. Research Report. ETS RR-09-37. *Educational Testing Service*.

Rindskopf, D., & Rose, T. (1988). Some theory and applications of confirmatory second-order factor analysis. *Multivariate Behavioral Research*, *23*(1), 51-67.

Rodriguez, A., Reise, S. P., & Haviland, M. G. (2016). Applying Bifactor statistical indices in the evaluation of psychological measures. *Journal of Personality Assessment*, *98*(3), 223-237.

Simms, L. J., Grös, D. F., Watson, D., & O'hara, M. W. (2008). Parsing the general and specific components of depression and anxiety with Bifactor modeling. *Depression and Anxiety*, *25*(7).

Spearman, C. E. (1904). General intelligence objectively determined and measured. *American Journal of Psychology*, 15, 201–293.

Stucky, B. D., Thissen, D., & Orlando Edelen, M. (2013). Using logistic approximations of marginal trace lines to develop short assessments. *Applied Psychological Measurement*, *37*(1), 41-57.

Stucky, B. D., & Edelen, M. O. (2014). Using hierarchical IRT models to create unidimensional measures from multidimensional data. *Handbook of Item Response Theory Modeling: Applications to Typical Performance Assessment*, 183-206.

Stucky, B. D., Edelen, M. O., Vaughan, C. A., Tucker, J. S., & Butler, J. (2014). The psychometric development and initial validation of the DCI-A short form for adolescent therapeutic community treatment process. *Journal of Substance Abuse Treatment*, *46*(4), 516-521.

Thomas, M. L. (2012). Rewards of bridging the divide between measurement and Clinical Theory: Demonstration of a Bifactor model for the brief symptom inventory. *Psychological Assessment*, *24*(1), 101.

Walther, B. A., & Moore, J. L. (2005). The concepts of bias, precision and accuracy, and their use in testing the performance of species richness estimators, with a literature review of estimator performance. *Ecography*, *28*(6), 815-829.

Watkins, M. W., & Beaujean, A. A. (2014). Bifactor structure of the Wechsler Preschool and Primary Scale of Intelligence—Fourth Edition. *School Psychology Quarterly*, *29*(1), 52.

Yang, Y., Song, L., & Xu, T. (2002). Robust estimator for correlated observations based on Bifactor equivalent weights. *Journal of Geodesy*, *76*(6-7), 353-358.

Zhang, B. (2008). Application of unidimensional item response models to tests with items sensitive to secondary dimensions. *The Journal of Experimental Education*, *77*(2), 147-166.

Zheng, C. (2013). *Examination of the parameter estimate bias when violating the orthogonality assumption of the Bifactor model* (Doctoral dissertation). University of Kansas.

İki Faktör Madde Tepki Kuramında Diklik Varsayımının İncelenmesi

Atıf:

Baris Pekmezci, F., & Gulleroglu, D. (2019). Investigation of the orthogonality assumption in the bifactor item response theory. *Eurasian Journal of Educational Research*, *79, 69-86,* DOI: 10.14689/ejer.2019.79.4

Özet

Problem Durumu: İki Faktör Modeli, çok boyutlu madde tepki kuramı (multidimensional item response theory) modellerinden biridir. İki faktör modeline göre birden fazla spesifik (özgül) faktör ve bu faktörler tarafından açıklanan bir genel faktör vardır ve ayrıca bu özgül etkilerin genel faktör üzerinde etkisinin olduğunu varsayılmaktadır. Tüm madde tepki kuramı modellerinde olduğu gibi İki Faktör modelinin de kendine özgü varsayımları vardır. İki Faktör Model'inin en önemli varsayımlarından biri verinin hem genel faktörü hem de spesifik faktörleri içermesidir. Bu varsayım karşılanması zor bir varsayım olmamakla birlikte çok boyutlu veriyi gerektirmektedir. Diğer varsayım olan faktörlerin dik (orthogonal) yani birbirinden bağımsız (ilişkisiz) olması ise pratikte karşılanması çok mümkün olmayan bir varsayımdır. İlişkili faktörleri dik olmaya zorlamak ise ölçülen yapı ile ilgili olarak bilgi kaybına neden olacak ve güvenilir olmayan parametre kestirimleri ile sonuçlanacaktır. Bu çalışma aracılığıyla İki Faktör Modelin kullanımını kısıtlayan varsayımın incelenmesi ve belirli kriterler ışığında değerlendirilmesi ile elde edilecek sonuçların alan yazına hem teorik anlamda hem de modelin daha doğru uygulanabilirliği açısından önemli katkılar sağlayacağı düşünülmektedir.

Araştırmanın Amacı: İki Faktör Kuramı, gerektirdiği varsayımdan (diklik) dolayı kullanımı sınırlanan bir kuramdır. Bu sınırlılığının yanı sıra psikolojik ve eğitsel yapıların modellenmesinde ve ölçek geliştirme çalışmalarında bu varsayım göz ardı edilerek sıklıkla kullanılmaktadır. Diklik varsayımının sağlanmadığı koşullarda psikolojik ve eğitsel yapıların doğru modellenmesi, geliştirilen ölçeğin doğru faktör yapısına ulaşması ve parametre kestirimlerinin doğru olması mümkün olmayacaktır. Bunun yanı sıra eğitim ve psikoloji alanında faktörler arası korelasyonun sıfır olduğu ölçme araçları geliştirmek neredeyse imkansızdır. İlişkili faktörleri dik olmaya zorlamak ise ölçülen yapı ile ilgili olarak bilgi kaybına neden olacak ve güvenilir olmayan parametre kestirimleri ile sonuçlanacaktır. Parametre kestirimlerinin kesinliği ve doğruluğu ise yapılan her ölçme işleminde önemli bir durumdur. Çünkü parametre kestirimleri, madde performansı ve yanıtlayıcı yetenek düzeyinin belirlenmesinde önemli bir unsurdur. Belirtilen bu gerekçelerden kaynaklı, iki faktör kuramının, spesifik faktörler arası farklı ilişki düzeylerine olanak tanıyarak incelenmesi yani hangi diklik ihlal düzeylerinin kuram tarafından tolere edilip, diklik ihlaline rağmen kararlı, kesin ve doğru kestirimler yapılabildiğinin belirlenmesi bu araştırmanın amacıdır.

Araştırmanın Yöntemi: Bu araştırma için veriler simülatif yolla iki adet (Model-1 ve Model-2) İki Faktör iki parametreli modele göre üretilmiştir. Model-1 iki spesifik faktör arasında çapraz yüklenmelerden dolayı oluşan diklik ihlalini gösteren modeldir. Burada incelenen nokta iki spesifik faktör arasındaki diklik ihlalinin tüm faktörlerdeki parametre kestirimlerine olan etkisidir. Model-2 ise, tüm spesifik faktörler arasındaki ilişkiyi göstermektedir. Spesifik modellerde manipüle edilen değişkenler faktörler arası korelasyon düzeyleri ve test uzunluklarıdır. Kurulan modeller için korelasyon kabul düzeyleri 0.10 (çok düşük), 0.40 (orta), 0.70 (yüksek) olarak ele alınmıştır. Spesifik faktörlerdeki madde sayılarına karar vermek amacıyla yapılan alan yazın incelemesi sonucunda test uzunlukları 12, 40 ve 100 madde olarak belirlenmiştir. Araştırma boyunca sabit tutulacak (manipüle edilmeyecek) değişken ise örneklem (5000) büyüklüğüdür. Replikasyon sayısı ise 200 olarak belirlenmiştir. Parametre kestirimlerinin replikasyonlar boyunca doğruluğunun değerlendirilmesi; ortalama yanlılık (mean bias), RMSE (hataların kareleri ortalamasının karekökü) ve kestirimlerin standart hatası (Standart Error) ile yapılmıştır.

Araştırmanın Bulguları: Ayırt edicilik parametreleri için tüm test uzunluklarında Model 1 ve Model 2 için görülen örüntü aynı şekildedir. Madde sayısındaki artış ayırt edicilik parametrelerinin kestirim kesinliğinde yani güvenirliğinde düşüşe neden olmuştur. Bu durum yanlı madde miktarındaki artış ile açıklanabilir. Yani modele ne kadar ilişkili madde eklenirse değişkenlik o kadar artmıştır. İki faktörün ilişkili olması durumu (Model-1) ile tüm faktörlerin ilişkili olması durumunun (Model-2), ayırt edicilik parametrelerinin kestiriminde neredeyse aynı etkiye sahip olduğu söylenebilir. Sonuç olarak her iki model için de parametre kestirim doğruluğu arasında farklılık yoktur. Buradan yola çıkarak model türünün parametre kestirim doğruluğuna etkisi olmadığı söylenebilir. Güçlük parametresinin kestiriminde, iki spesifik faktörün ilişkili olma durumu (Model 1) ile tüm spesifik faktörlerin ilişkili olma durumunun (Model 2) neredeyse aynı etkiye sahip olduğu söylenebilir. Yani model türünün güçlük parametre kestirim doğruluğuna etkisi olmadığı söylenebilir. Birey parametreleri incelendiğinde, test uzunluğu ile doğru orantılı şekilde değişkenliğin azalması test uzunluğunun parametre iyileşmesinde etkisi olabileceğine işaret etmektedir. Yine de değişkenlik tüm test uzunluklarında yüksektir. Bu durum parametre kestirim güvenirliklerini düşürmektedir. Birey parametrelerinin kestiriminde, iki spesifik faktörün ilişkili olma durumu ile tüm spesifik faktörlerin ilişkili olma durumunun neredeyse aynı etkiye sahip olduğu söylenebilir.

Araştırmanın Sonuçları ve Önerileri: Kestirim doğruluğu en düşük parametrelerin güçlük parametreleri olduğu görülmüştür. Ayırt edicilik, güçlük ve birey parametrelerinin kestirim doğruluğunda ise modelin öneminin olmadığı görülmüştür. Yani iki spesifik faktörün ilişkili olma durumu (Model 1) ile tüm faktörlerin ilişkili olma durumu (Model 2) hem birey hem de madde parametrelerinin kestirim doğruluğunda aynı etkiye sahiptir. Madde sayısını arttırmak, birey parametrelerinin kestirim kesinliğini yani güvenirliğini arttırmıştır. Birey parametrelerinde gözlenen bu durum, madde sayısı arttıkça bireyin örtük özelliğinin daha iyi açıklandığının bir sonucudur. Birey parametrelerinin kestiriminde, güvenirliği en düşük parametre kestirimleri her iki model için de (Model 1ve Model 2) en küçük test uzunluğundadır.

Test uzunluğu arttıkça kestirim güvenirliği de artmıştır. Buna rağmen tüm test uzunluklarında ve diklik ihlal düzeylerinde kestirim güvenirliği en düşük parametreler birey parametreleridir. Madde ve birey parametrelerinin kestirimi psikolojik ve eğitsel amaçlı değerlendirmelerde önemli bir unsurdur. İki faktör kuramının ilişkili yapılarda kullanılması yanlı parametre kestirimlerine, parametre kestirimlerindeki yanlılık ise değerlendirme sonuçlarında yanlılığı doğuracaktır. Literatürde varolan araştırmalar iki faktör kuramının ilişkili yapılarda bile çok iyi düzeyde uyum verdiği ve robust bir model olduğu belirtmektedir. Bu araştırmada ise parametre bazında yanlılık incelendiğinde bu robust yapı görülememiştir. İki faktör kuramı, birey parametrelerinin kestiriminde test uzunluğu arttıkça diklik varsayımı ihlalini daha iyi tolere edebilmektedir. Bu kuramı kullanmak isteyen uygulayıcıların büyük madde havuzları ile çalışmaları önerilir. Tüm korelasyon düzeylerinde parametre kestirim doğrulukları yaklaşık olarak aynı çıkmıştır. Yeni çalışmalar ara korelasyon (0.25, 0.35 vb.) düzeyleri ile tekrarlanabilir.

Anahtar Kelimeler: Çok boyutlu madde tepki kuramı, İki faktör Madde Tepki Kuramı, diklik varsayımı, parametre kestirim yanlılığı, faktör analizi.