KURAM VE UYGULAMADA EĞİTİM BİLİMLERİ EDUCATIONAL SCIENCES: THEORY & PRACTICE

Received: August 6, 2015
Revision received: January 25, 2015
Accepted: February 8, 2015
OnlineFirst: February 15, 2016

ISSN 1303-0485 **eISSN** 2148-7561
www.estp.com.tr
DOI 10.12738/estp.2016.1.0220 • February 2016 • 16(1) • 153-171

Research Article

The Effect of Sample Size on Parametric and Nonparametric Factor Analytical Methods*

Ömür Kaya Kalkan[1]
Hacettepe University

Hülya Kelecioğlu[2]
Hacettepe University

Abstract

Linear factor analysis models used to examine constructs underlying the responses are not very suitable for dichotomous or polytomous response formats. The associated problems cannot be eliminated by polychoric or tetrachoric correlations in place of the Pearson correlation. Therefore, we considered parameters obtained from the NOHARM and FACTOR programs (which use parametric methods) and from the DETECT and DIMTEST programs (which use nonparametric methods) for different sample sizes of a real large dataset (50, 80, 100, 160, 200, 300, 500, 1000, 3000, 5000). A parallel analysis (PA) based on the tetrachoric correlation with the FACTOR program produced inconsistent results among the sampling sizes. However, the analyses based on the Pearson correlation could not adequately determine the dimension numbers. Although DETECT and NOHARM determined the multidimensionality at acceptable level for the 50 sample size, they yielded the most consistent results at sample sizes of 1000 and above.

Keywords

Test dimensionality • Sample size effect • Nonlinear item factor analysis • Exploratory nonparametric dimension assessment • Conditional item pair covariance

* This paper was revised after being presented at the Second International Eurasian Educational Research Congress, Ankara, Turkey, June 08-10, 2015.

1 **Correspondence to:** Ömür Kaya Kalkan (PhD), Hacettepe ASO 1.OSB Vocational School, Hacettepe University, 1. Organize Sanayi Bölgesi, ASORA Ticaret Merkezi, Ayaş yolu 25. Km. Ankara 06935 Turkey. Email: kayakalkan@hacettepe.edu.tr

2 Department of Educational Measurement and Evaluation, Faculty of Educational Sciences, Hacettepe University, Ankara Turkey. Email: hulyaebb@hacettepe.edu.tr

Citation: Kalkan, Ö. K., & Kelecioğlu, H. (2016). The effect of sample size on parametric and nonparametric factor analytical methods. *Educational Sciences: Theory & Practice, 16*, 153-171. http://dx.doi.org/10.12738/estp.2016.1.0220

Educational and psychological tests are extensively used and effectively contribute to many fields. For example, in clinical applications, they can detect severe emotional disorders and behavioural problems, assess teaching programs, determine learning deficiencies, classify students by their capabilities and select eligible recipients of diplomas. Psychological tests are also used to select industrial personnel by classifying and determining a potential employee's professional skills. Traditionally, psychological tests measure differences among individuals or the responses of the same individuals under different conditions. Well-structured tests can provide an accurate measurement of these individual differences (Anastasi & Urbina, 1997).

Investigating the constructs underlying the responses is one of the most important stages of assessing test structures as well as developing, evaluating and continuing large-scale tests. Such an assessment offers empirical evidence for the cognitive processes and content aspects of the test validity (as cited in Tate, 2003). Determining the dimensionality of a group of variables is important when constructing a psychological theory and developing a scale (Timmerman & Lorenzo-Seva, 2011).

Dimensionality is also applied in hypotheses testing of "homogeneity" in classical test theory (CTT), "unidimensionality" in item response theory (IRT). The former provides logical justification for behaviours related to psychological constructs. From the CTT perspective, the items on a psychometric homogeneous test measure only one attribute of a common factor. This type of item set can be defined as "unidimensional," because it indicates variation of respondents on a single dimension (McDonald, 1999). The items of CTT models have been hypothesized to measure the same dominant dimension (Nandakumar & Stout, 1993). If evidence for unidimensionality is obtained, Cronbach alpha coefficient (which determines the reliability in CTT) could be calculated for these measures. In this case, the obtained value is close to the real reliability (Cotton, Campbell, & Malone, 1957; Yang & Green, 2011). The "unidimensionality" of basic assumption of IRT directly affects on the IRT models, the obtained items and ability parameters, the test equating and test scaling parameters and the model-data fit indices (Embretson & Reise, 2000; Hambleton & Swaminathan, 1985; Kolen & Brennan, 2004). Therefore, it is important to assess the dimensionality from the CTT and IRT perspectives.

Factor analysis was developed as a tool for determining psychological attributes and is particularly related to the construct validity process. Construct validity is important when considering the nature and number of dimensions underlying the responses of an aptitude test or an attitude scale (Anastasi & Urbina, 1997; Embretson & Reise, 2000; McDonald, 1999). The dimensionality of a test can be assessed from the test specifications, (which envelope the achievement domain and determine a representative sample of items from this domain), content analysis (performed by a

test development specialist) and psychological analysis methods (which formulate a hypothesized item structure from a psychological perspective) (Ackerman, Gierl, & Walker, 2003). The factor structure can be analyzed by principal component analysis (PCA), principal axis factoring (PAF), unweighted least squares (ULS), generalized least squares (GLS), maximum likelihood (ML), alpha factoring (AF), image factoring (IF) techniques. The factor number is then determined by the Kaiser criterion, scree plot, eigenvalues and parallel analysis, of which parallel analysis is the most recommended method (Timmerman & Lorenzo-Seva, 2011). However, linear factor analysis models used to examine the dimensionality of a test are not suitable for item-level data, because most achievement tests are dichotomous and because personality and behavior scale items have dichotomous or polytomous response formats. As has been stated in many sources, traditional factor analysis methods are based on continuous rating and normality assumptions (Embretson & Reise, 2000). These hypotheses are violated when there are few categories and when the frequency of a category use is unsystematic. Such violations can lead to underestimates of factor loadings and/or overestimates of the number of latent dimensions (Gibbons et al., 2007; Nandakumar & Stout, 1993). Furthermore, it is almost impossible to measure the continuum of traits at all intervals with equal reliability or sensitivity (Waller, Tellegen, McDonald, & Lykken, 1996). To avoid the limitations of linear factor analysis, dichotomous and polytomous data can be analyzed by polyserial, polychoric or tetrachoric correlations instead of the Pearson correlation (Ackerman et al., 2003; Embretson & Reise, 2000; Hambleton & Swaminathan, 1985). Because the matrix of tetrachoric correlation coefficients is generally not positive, the common factor model yields unreliable results. Furthermore, the tetrachoric correlation matrix is not suitable when the θ distribution is not normal. In general, the relationship between item performance and underlying latent ability is nonlinear, causing a misfit between the model and data, and confusion between the dimensionality and item difficulty (which is quantified by the *difficulty factor*). In linear models, the difficulty factor is usually the first factor; however, in nonlinear factor analysis models, the effect of the difficulty factor decreased or eliminated (Ackerman et al., 2003; Bock, Gibbons, & Muraki, 1988; Hattie, Krakowski, Rogers, & Swaminathan, 1996; Nandakumar & Stout, 1993). The covariance among items for candidates at the same ability levels in unidimensional tests should be zero. The covariance among items is typically nonlinear because of which the, linear factor analysis methods cannot sufficiently assess dimensionality. In this case, we should use nonlinear factor analysis methods (Embretson & Reise, 2000;Hambleton & Swaminathan, 1985; McDonald, 1967, 1981).

Statistical methods for assessing test dimensionality can be classified as parametric or nonparametric. Parametric methods assume a specific parametric model for the item response function (IRF), whereas nonparametric methods simply assume that the IRFs are monotonic (Zhang & Stout, 1999). Softwares using conditional item

covariances were summarized by Tate (2003) based on parametric methods (Mplus, NOHARM, CHIDIM, TESTFACT and IRTNEW) and nonparametric methods (HCA/CCPROX, DIMTEST and DETECT). Within the scope of this study, the structure of a test comprising dichotomously scoring items was used to empirically evaluate different sampling sizes based on conditional item covariances. The FACTOR and NOHARM programs were used to investigate the parametric models, whereas DIMTEST and DETECT were used to investigate the nonparametric methods.

Parametric Methods

Nonlinear Item Factor Analysis (NOHARM)

It is generally accepted that parametric linear factor analysis does not perform well at an item-level (Ackerman et al., 2003; Embretson & Reise, 2000; Hambleton & Swaminathan, 1985; Waller et al., 1996; Zhang & Stout, 1999). Parametric nonlinear factor analysis (NOHARM) is an alternative approach based on weak local independence (Stout et al., 1996). NOHARM was designed by considering nonlinear regressions between dichotomous items and latent factors; therefore, it is unlikely to generate factors that are psychometrically spurious (Waller et al., 1996). The NOHARM software was written by Fraser (1988). It conforms to the unidimensional and multidimensional normal ogive models of latent trait theory, as presented by McDonald (1967). This program uses the nonlinear factor analytic approach, to estimate the item parameters in both exploratory and confirmatory modes, and it provides reasonable estimates (Ackerman et al., 2003; Van Der Linden & Hambleton, 1997; Zhang & Stout, 1999). The parameter estimates are obtained by minimizing the unweighted least squares (ULS) function in the first and second degree marginal ratios (Maydeu-Olivares, 2001). NOHARM provides the root mean square of residual (RMSR) value, which summarizes the residual covariance matrix and fit of the model for every model estimated (Tate, 2003). NOHARM can estimate up to 50 dimensions and 215 items and is preferable when the data include dichotomous responses (McDonald, 2000; Reckase, 2009).

FACTOR

Univariate and multivariate descriptive statistics and dispersion matrixes can be obtained with the Factor (Version 9.20) software, program developed by Lorenzo-Seva and Ferrando (2013). This software can also perform exploratory common factor analyses in different forms. These procedures include the minimum average partial (MAP) test proposed by Velicer (1976), parallel analysis (PA, classical implementation) based on the Pearson and polychoric correlation proposed by Horn (1965), and optimal parallel analysis (OPA, optimal implementation) proposed

by Timmerman and Lorenzo-Seva (2011). The MAP test proposed by Velicer is based on a partial correlation matrix. In the PA (Horn, 1965) method, the 500 Pearson correlation matrix is obtained assuming a normal distribution. This method is recommended when Pearson correlation matrixes are used to study the principal components. FACTOR software is conduct analyses based on the OPA, Pearson or tetrachoric correlation matrices. The user can determine the number of random correlation matrices and the procedure for constructing these matrices (i.e. normal distribution or permutation of sample values). ULS is used to determine the factor/component number using PA and minimum rank factor analysis (MRFA) is used for OPA. There are no restrictions on the numbers of variables and people (Lorenzo-Seva & Ferrando, 2006, 2013; Zwick & Velicer, 1986).

Nonparametric Methods

Using the nonparametric dimensionality assessment methods of DETECT, DIMTEST and HCA/CCPROX, we can define the multidimensional latent structure of a test using conditional item paired covariances (Zhang & Stout, 1999).

DETECT

DETECT is an exploratory nonparametric dimensionality assessment method that uses the dominant dimensions in the dataset and the magnitude of departure from unidimensionality. Roussos, Reese and Harris (1997) stated that DETECT can also define the dominant dimension measured by each item (as cited in Ackerman et al., 2003). The main aim of DETECT is to define the groups by maximizing the DETECT index. This index shows the magnitude of departure from unidimensionality. The conditioned DETECT index is obtained by calculating all the item covariances (Zhang & Stout, 1999).

DIMTEST

DIMTEST is a nonparametric statistical method that uses a hypothesis test to assess the existence of multidimensionality. Furthermore, DIMTEST can be seen a technique that determines the lack of the fit compared to the local independent unidimensional latent trait model. The DIMTEST method separates the test items into two groups: the assessment subtest (AT1) and partitioning subtest (PT). It then evaluates the conditioned covariance relationship between these two groups. The test statistics (T) calculated by DIMTEST represent the degree of dimensional discrimination of the two-item group (Ackerman et al., 2003; Nandakumar & Stout, 1993; Stout et al., 1996).

NOHARM, DIMTEST and DETECT clearly determine the presence of multidimensional test structure (Tate, 2003). Both DETECT and NOHARM correctly place all the items in the correct clusters and define the correct dimension numbers

(Finch & Habing, 2005). In addition, FACTOR performs parallel analysis (Horn, 1965) or optimal parallel analysis (Timmerman & Lorenzo-Seva, 2011) based on Pearson or tetrachoric correlation. Furthermore, all these software packages are freely available. Because of these advantages, we employ NOHARM, DIMTEST, DETECT and FACTOR software in the present research.

When realizing a factor analysis, there is no shortage of suggestions regarding the appropriate sample size. Gorsuch (1974) and Norman and Streiner (2003), suggested a minimum of 5-fold sample size; Jöreskog and Sörbom (1989) and Vieira (2011) proposed a minimum of 10-fold sample size; Thompson (2004), recommended a flexible sample size between 10- and 20-fold; Comrey and Lee (1992) suggested that a sample size 50 is very weak, 100 is weak, 200 is average, 300 is good, 500 is very good and 1000 is perfect. According to Tabachnick and Fidell (2006), a sample size of 300 is generally sufficient. In contrast, Kline (2011) and Boomsma (as cited in Tanaka 1987) emphasized that a sample size 200 usually sufficient. In summary, the recommended minimum sample size 3 to 20 times the number of variables. Sample size is the one of the most widely addressed issues in all of statistical analysis (Mundfrom, Shaw, & Ke, 2005). Considering these different suggestions, the impact of sample size should also be examined by nonparametric factor analysis.

In Turkey, test dimensionality has rarely been addressed by nonparametric methods (Özbek Baştuğ, 2012; Özer Özkan, 2012; Özer Özkan & Acar Güvendir, 2014). Using DIMTEST, Özer Özkan and Acar Güvendir (2014) determined that all the subtests in Turkish language, mathematics, science and technology, social sciences and English language in the Student Achievement Determination Exam (ÖBBS) of 2002, 2005 and 2008 are multidimensional. Özbek Baştuğ (2012) analyzed the dimensionality of the Social Sciences subtest of the Primary Schools Selection Exam by parametric and nonparametric methods, and concluded that this sub is multidimensional. Using nonparametric methods, Özer Özkan (2012) determined that the 8th-class Turkish language and mathematics subtests of ÖBBS are also multidimensional.

Comparison of parametric and nonparametric methods based on several variables and studies related to test dimensionality analysis have been performed on real and simulation data. For example, Jang and Roussos (2007) studied the dimensionality of the English as a Foreign Language (TOEFL) test by nonparametric methods. Finch and Habing (2005) reported that NOHARM and DIMTEST, performed equally well when determining the test dimensionality and clustering the items. Tate (2003) compared the dimensionality of the tests comprising dichotomous items obtained by parametric and nonparametric, exploratory and confirmatory factor-analytic methods. He found that the nonparametric exploratory factor-analytic methods consistently outperform the parametric methods. Stout et al. (1996) applied HCA/CCPROX,

DIMTEST and DETECT to the Law School Admission Test (LSAT) of 1991, 1992-1 and 1992-2. They examined the multidimensionality of this test by three methods, giving three different perspectives of the problem.

Any method for assessing the dimensionality of a test is, only a tool. Although many studies have reported promising results, they have lacked evaluation on real test data (Ackerman et al., 2003). Considering that dimensionality of large scale exams in Turkey (SBS, ÖBBS) has rarely been investigated by nonparametric methods, we expect that our study will make a valuable contribution to the literature. Therefore, we applied parametric and nonparametric methods on a real dataset with different sample sizes, and compared the obtained parameters.

Method

Datasets

The data consisted of responses to 16 multiple-choice questions from 480,691 candidates on the Science and Technology Test of the Sixth Grade Level Determination Examination (SBS) in 2008, from the Republic of Turkey, Ministry of National Education (T.C. MEB). This subtest measures the educational attainment of 6th- grade students in the Science and Technology curriculum of a given year (MEB, 2011). In SBS-2008 6th-grade Science and Technology subtest, students were assessed based on their knowledge, understanding, problem-solving skills and application of the scientific method. In addition, when the Science and Technology subtest was described with single factor, the KR-20 reliability was 0.76, indicating that the item difficulty varied from 0.14 to 0.69, and the item discrimination varied between 0.09 to 0.80 (as cited in Güzeller, 2012). Descriptive statistics of the Science and Technology lesson of 2008 SBS are given in Table 1.

Table 1
Test Average and Standard Deviations of the Science and Technology lesson of 2008 SBS

	Number of items	Mean	Standard deviation
Science and Technology Test	16	4.78	4.46

As described above, the determination of adequately sample size in exploratory factor analysis methods is widely contended in the literature. Therefore, we created sample sizes of 50, 80 (16 items × 5), 100, 160 (16 items × 10), 200, 300, 500, 1000, 3000 and 5000 persons by random methods.

Data Analysis

First, randomly selected samples of 50, 80, 100, 160, 200, 300, 500, 1000, 3000 and 5000 people from the 480,601 person dataset was obtained using the Statistical Program

for the Social Sciences (SPSS) 20.0, following which the descriptive statistics for all the sample size were obtained. DIMTEST was used to examine the dimensionality of the datasets. We determined the dimensionality of the dataset using the T statistics and p (0.01) values obtained from this software. Considering how well the underlying data structure approximated the simple structure, we evaluated the multidimensionality of the data using the obtained DETECT index, and the r index. The DETECT index of unidimensional and multidimensional data approaches zero and one, respectively. Kim (1994) classified data as unidimensional (DETECT index < 0.1), weakly multidimensional (0.10 < DETECT index < 0.50) or moderately multidimensional (0.51 < DETECT index < 1.0). In addition, r values larger than 0.80 imply that the underlying data structure approximates a simple structures (as cited in Ackerman et al., 2003). These analyses were performed by the nonparametric dimensionality assessment software, DIMPACK 1.0, which contains the DIMTEST V.2.1, DETECT V.2.1 and CCPROX/HAC methods. The numbers of factors and components were determined by the PA and OPA methods implemented in Factor 9.2 software. The PA and OPA methods were separately performed on the Pearson and tetrachoric correlation matrices. The numbers of factors and components were determined using PA; ULS was used as an estimator method and MRFA, was used for the OPA. The varimax factor rotation method was used to maximize the factor simplicity of the PA and OPA methods. For each sample size, we evaluated the data-model fit using the RMSR and Kelley's criterion, and conducted a parametric exploratory factor analysis. The program outputs included advised dimension numbers (ADNs). The nonlinear parametric exploratory factor analysis for each sampling size was performed using NOHARM 4.0. The data-model fit was evaluated using RMSR and the Tanaka goodness-of-fit indices. The obtained numbers of dimensions were compared with those from DETECT. A confirmatory factor analysis was performed using NOHARM 4.0. The same fit indices were used to evaluate the data-model fit. Kaiser–Meyer–Olkin (KMO) test is used to evaluate the adequacy for the exploratory factor analysis. The result of this test should be greater than 0.7; if it is lower than 0.5, the factor analysis should not be continued (Field, 2009; Leech, Barrett, & Morgan, 2005).

Results

Descriptive Statistics of Sample Sizes

The descriptive statistics for the ten sample sizes are shown in Table 2.

Table 2 reveals no excessive deviations from the normal distribution. Regardless of sample size, the skewness and kurtosis coefficients were within the boundaries of −1 and +1 (Çokluk, Şekercioğlu, & Büyüköztürk, 2010).

Table 2
Descriptive Statistics for Different Sampling Sizes

–	Number of items	Mean	Standard deviation	Skewness	Kurtosis
50	16	7.84	4.04	0.17	−0.95
80	16	7.34	4.13	0.39	−0.88
100	16	7.20	3.72	0.40	−0.67
160	16	7.25	3.77	0.48	−0.70
200	16	7.16	3.81	0.50	−0.79
300	16	7.42	3.75	0.43	−0.85
500	16	7.24	3.85	0.45	−0.89
1000	16	7.01	3.78	0.54	−0.67
3000	16	7.31	3.62	0.46	−0.69
5000	16	7.23	3.58	0.51	−0.61

Dimensionality of the Dataset

DIMTEST was used to statistically test whether the item groups came from different dimensions. In this hypothesis test, the H0 and the H1 hypotheses state that the items come from different dimensions, respectively (in the H1 hypothesis, the dataset is multidimensional).

On the basis of the results ($t = 2.9636$, $p < .01$), we rejected H0 hypothesis and accepted that the dataset of the 2008 T.C. MEB SBS Sixth Grade Science and Technology course was multidimensional.

Nonparametric Exploratory Factor Analysis Using DETECT

Table 3 contains the dimension numbers obtained for each sampling size, the DETECT and the r index values.

Table 3
DETECT Nonparametric Exploratory Factor Analysis Results

Sample Size	Dimensions with items				DETECT index	r index
	I	II	III	IV		
50	1,7,9,11,13,14,15	2,4,6,8,16	3,5,10,12	-	0.75	0.5965
80	1,2,3,7,11,12,14	4,6,9,10,13,16	5,8,15	-	0.7250	0.6337
100	1,3,8,9,12,13	2,7,14,15	4,5,6,10,11,16	-	0.6833	0.6161
160	1,3,4,6	2,7,9,11,12,13,14,15	5,8,10,16	-	0.7250	0.6162
200	1,8,9,10	2,7,11,12,13,14,15	3,4,5,6,16	-	0.7667	0.7414
300	1,3,4,5,10,16	2,7,8,12,13,14,15	6,9,11	-	0.7250	0.6753
500	1,2,7,10,11,12,13,14,15	3,4,6,9,16	5,8	-	0.7000	0.6870
1000	1,8	2,7,11,12,14,15	3,4,6,13,16	5,9,10	0.7667	0.7074
3000	1,6,16	2,7,8,11,12,13,14,15	3,4,5	9,10	0.7917	0.7842
5000	1,5,8	2,7,9,11,12,13,14,15	3,4,6,16	10	0.8000	0.8141

The DETECT indices for all the sample sizes varied from 0.6833 to 0.80. All indices were between 0.51 and 1.0 implying that the data are moderately multidimensional. The r value of the 5000 sample size (0.80) indicates an approximately simple structure of the data holds. Each of the 1000, 3000 and 5000 sample datasets were 4-dimensional (see Table 3).

Factor 9.2 Analyses Based on the Pearson and Tetrachoric Correlations

The factor numbers were determined in PA and OPA analyses of the Pearson and tetrachoric correlation matrices. The statistical results are listed in Table 4. In order to evaluate its adequacy in the exploratory factor analysis, each sample size was examined by the Kaiser–Meyer–Olkin (KMO) test. The adequacies of the 50 and 100 sample sizes were "good," whereas those of the 80, 160, 200, 300, 500, 1000, 3000 and 5000 sample sizes were "very good." We applied the Bartlett test to determine whether our correlation matrices significantly differed from the identity matrix. As expected from previous literature (Field, 2009; Stevens, 2009; Tabachnick & Fidell, 2006), the chi-squared value in the Bartlett's test was significant (p<.00001) for all samples, implying sufficient correlation for a reasonable factor analysis of the variables (Leech et al., 2005).

Table 4
PA and OPA Results based on the Pearson and Tetrachoric Correlations

			PA (Horn, 1965)						OPA (Timmerman & Lorenzo-Seva, 2011)					
			Pearson Correlation			Tetrachoric Correlation			Pearson Correlation			Tetrachoric Correlation		
Sample Size	KMO Test	Bartlett Test (p)	ADN*	RMSR	Kelley's Criterion	ADN*	RMSR	Kelley's Criterion	ADN*	RMSR	Kelley's Criterion	ADN*	RMSR	Kelley's Criterion
50	0.75213	0.00001	1	0.0676	0.1429	3	-	-	1	0.0767	0.1429	1	0.0847	0.1429
80	0.75565	0.00001	1	0.0498	0.1125	1	0.0653	0.1125	1	0.0603	0.1125	1	0.0781	0.1125
100	0.75803	0.00001	1	0.0548	0.1005	3	0.0783	0.1005	1	0.0620	0.1005	1	0.0938	0.1005
160	0.82844	0.00001	1	0.0388	0.0793	2	0.0627	0.0793	1	0.0459	0.0793	1	0.0755	0.0793
200	0.86494	0.00001	1	0.0326	0.0709	2	0.0498	0.0709	1	0.0388	0.0709	1	0.0580	0.0709
300	0.85955	0.00001	1	0.0332	0.0578	2	0.0535	0.0578	1	0.0393	0.0578	1	0.0633*	0.0578
500	0.89033	0.00001	2	0.0279	0.0448	3	0.0430	0.0448	1	0.0323	0.0448	1	0.0497*	0.0448
1000	0.89622	0.00001	2	0.0214	0.0316	2	0.0346*	0.0316	1	0.0246	0.0316	1	0.0397*	0.0316
3000	0.89887	0.00001	2	0.0146	0.0183	3	0.0223*	0.0183	1	0.0180	0.0183	1	0.0272*	0.0183
5000	0.89598	0.00001	2	0.0141	0.0141	3	0.0221*	0.0141	1	0.0165*	0.0141	1	0.0258	0.0141

Note. ADN*: Advised Dimension Number

Table 4 lists the RMSRs of the data-model fit based on the Pearson correlation matrix, calculated by the method proposed by Horn. The RMR or RMSR defines the square root of the mean of the covariance residuals, which are the differences between the corresponding elements of the observed and predicted covariance matrices (Brown, 2006; Hooper, Coughlan, & Mullen, 2008; Nargundkar, 2008; Westland, 2015). In Table 4, the RMSRs range from .0141 to .0676. An RMSR below 0.08 indicates a sufficient fit; however, for well-fitting models, the RMSR should be less than .05 (Brown, 2006; Hu & Bentler, 1999; Tabachnick & Fidell, 2006). In this assessment, the RMSR values were sufficient for the 50 and 100 (< .08) and good (< .05) for the other 8 sample sizes. Applying the same method to the tetrachoric correlation matrix, the RMSRs range from 0.0221 to 0.0783; however, they could not be calculate for the 50 sample sizes. The RMSRs were sufficient for the 80, 100, 160 and 300 sample sizes (< .08), and good for the 200, 500, 1000, 3000 and 5000 sample sizes (< .05).

Furthermore, in order to evaluate the model-data fit, we compared the RMSR values with Kelley's criterion. Kelley (1935) and Harman (1962) proposed that if the RMSR is much larger than Kelley's criterion, the model cannot fit the data adequately (as cited in Lorenzo-Seva & Ferrando, 2013). In the PA results based on the Pearson correlation, the RMSR values were smaller than Kelley's criterion (and equal to Kelley's criterion for the 5000 sample size). In the PA results based on the tetrachoric correlation matrix, the model-data fit weakened as the sample size increased, and very weak for sample size of 1000, 300 and 5000. The RMSRs for these three sample sizes exceeded Kelley's criterion.

The Pearson and tetrachoric correlation matrices were also evaluated by OPA based on the MRFA proposed by Timmerman and Lorenzo-Seva (2011). In the analysis results based on the Pearson correlation matrix, the RMSRs varied from 0.0165 to 0.0767. The model-data fit was sufficient (< .08) for the 50, 80 and 100 sample sizes, and good (< .05) for the other 7 sample sizes. Furthermore, the RMSR decreased as the sample size increased. However, even the minimum RMSR (0165 for the 5000 sample) exceeded Kelley's criterion (.0141).

In the same OPA analysis based on the tetrachoric correlation matrix, the RMSR ranged from .02580 to 0.0938. The model-data fit was poor at sample sizes of 50 and 100 (RMSR > .08), and sufficient (RMSR < .08) at sample sizes of 80, 160, 200 and 300. The 500, 1000, 3000 and 5000 sample sizes yielded good fits. The RMSR and Kelley's criterion were mismatched. In particular, when the RMSR values exceed 0.08 (indicating an insufficient, data-model fit), Kelley's criterion implied a good model-data fit; in contrast, when the RMSR values were below .05 (suggesting a good fit), Kelley's criterion indicated a very weak data-model fit.

Exploratory and Confirmatory Factor Analysis Using NOHARM

Table 5 lists the results of the nonlinear exploratory factor analysis for the different sample sizes. The dimensions were determined using NOHARM.

Table 5
Results of NOHARM Nonlinear Exploratory Factor Analysis

Sample	Dimensions with items				RMSR	Tanaka
Size	I	II	III	IV		GFI
50	1,2,8,9,10,11,12,14	4,5,6	3,7,13,15,16	-	0,0152183	0,9729747
80	4,5,6	2,3,7,12,13,15,16	1,8,9,10,11,14	-	0,0113527	0,9847796
100	1,2,3,9,10,11,12,13,14,15,16	5,6	4,7,8	-	0,0124933	0,9757302
160	1,2,3,7,10,11,13,14,15	5,6,16	4,8,9,12	-	0,0086983	0,9879185
200	1,2,7,8,9,10,11,15	3,4,5,12,13,16	6,14	-	0,0073259	0,9915035
300	1,3,4,5	2,7,8,9,10,11,12,13,14,15,16	6	-	0,0074040	0,9910113
500	1,5	2,3,7,8,9,10,11,12,13,14,15,16	4,6	-	0,0061799	0,9942925
1000	1,3,8,10,11	2,7,9,12,14,15,16	4,5,6	13	0,0037050	0,9977536
3000	1,4,5,10	2,7,8,9,11,12,13,14,15	3	6,16	0,0024581	0,9989556
5000	1,5	2,7,8,9,10,11,12,13,14,15	3	4,6,16	0,0023488	0,9990267

According to the exploratory factor results from NOHARM, the test structure appeared to be 3-dimensional for the 50, 80, 100, 160, 200, 300 and 500 sample sizes, and 4-dimensional for the 1000, 3000 and 5000 sample sizes. In order to determine the factor model that best fitted the dichotomous dataset, we conducted a NOHARM factor analysis for the 2, 3, 4 and 5 dimensions. In this analysis, the differently-sized datasets were best fitted to a 3- or 4-dimensional structure. The most suitable model-data fit indices are given in Table 6.

Table 6
Results of NOHARM Confirmatory Factor Analysis

		Confirmatory factor analysis			
Sample Size	Dimension Number	3 dimension		4 dimension	
		RMSR	Tanaka GFI	RMSR	Tanaka GFI
50	3	0.0187989	0.9587617	-	-
80	3	0.0152204	0.9726424	-	-
100	3	0.0184759	0.9469213	-	-
160	3	0.0138418	0.9694063	-	-
200	3	0.0108708	0.9812912	-	-
300	3	0.0111485	0.9796207	-	-
500	3	0.0100489	0.9849087	-	-
1000	4	-	-	0.0073488	0.9911622
3000	4	-	-	0.0068369	0.9919204
5000	4	-	-	0.0073456	0.9904810

Table 6 reveals a good fit (RMSR < .05), and the goodness-of-fit indices were greater than 0.95 (Hooper et al., 2008). In the confirmatory factor analysis, the RMSR values were significantly improved (by 10%) in the new model relative to the previous model (Tate, 2003). Considering these criteria, we identified the 50, 80, 100, 160, 200, 300 and 500 sample sizes as 3-dimensional, and the 1000, 3000 and 5000 sample sizes as 4-dimensional.

Discussion

In this study, we examined the effect of sample size on the dimensionality of a test. The dimensionality was computed by parametric and nonparametric exploratory factor analysis methods. The results of DIMTEST implied that the real dataset was multidimensional. Comparative studies of DIMTEST by Hattie et al. (1996) confirmed that DIMTEST consistently evaluates the dimensionality of comparison data.

The factor/component numbers were determined from the RMSR values obtained by the PA based on Pearson correlation. This analysis was implemented by the Factor software. The results completely agreed with those of Kelley's criterion. However, in the analyses based on tetrachoric correlations, the RMSR and Kelley's criterion were weakly matched for sample size of 1000, 3000 and 5000. The RMSR values

for these sample sizes were below 0.03, implying good fits. In the analysis based on the Pearson correlation matrix, the RMSR decreased as the sample size increased. The analyses based on tetrachoric correlations yielded larger RMSRs than those based on Pearson correlation, regardless of the sample sizes. However, the RMSRs evaluated in these analyses were less sensitive to sample size than those derived from Pearson's coefficients. Because sample size exerts the largest effect on RMSR, increasing the sample size increases the estimation accuracy (Hooper et al., 2008; Thomas, 2003). Weng and Cheng (2005) stated that tetrachoric correlations in small samples introduces errors in large samples. Eight sample sizes yielded good fits (RMSR < .05) when based on the Pearson correlation, decreasing to 5 when using tetrachoric correlation. The ADN values based on the Pearson correlation appeared to be consistent, and the predictions underestimated the real number of dimensions. The ADN values based on tetrachoric correlation were negatively affected by the sample size. For example, the ADN value was 3 for the 50 sample size and 1 for the 80 sample size, recovering to 3 for the 100 sample size.

When determining the factor/component numbers, the OPA based a on the Pearson correlation (implemented in the Factor software) obtained a good fit between Kelley's criterion and the RMSR values. The RMSE exceeded Kelley's criterion (.0165 > .0141) only for the 5000 sample size. However, in the analyses based on tetrachoric correlation, the RMSRs exceeded the Kelly criterion for 4 sample sizes (300, 500, 1000 and 3000). Compared to the Pearson correlation, the number of sample sizes with good-fit values decreased from 7 to 4. In the OPA analyses based on Pearson correlation matrix, the RMSR decreased with increasing sample size. The analyses based on tetrachoric correlation yielded higher RMSRs than those based on the Pearson correlation, regardless of the sample size. Timmerman and Lorenzo-Seva (2011) stated that Pearson correlation is preferable to tetrachoric correlation when defining common factor numbers. For all sample sizes, OPA based on both the Pearson and the tetrachoric correlations, returned an ADN value of 1.

The ordered polytomous items of the OPA method produce better dimensionality results than the PA method (Timmerman & Lorenzo-Seva, 2011). However, in the present study, the exploratory OPA analyses based on the Pearson and tetrachoric correlations did not distinguish the multidimensionality of the dichotomous dataset. The ADN value was 1 for all sample sizes, and the type of correlation matrix exerted no significant effect. Furthermore, the dimensionality assessment using by this method was insensitive to sample size. Linear factor models generally define the first factor as the "difficulty factor." If the first factor dominates, the effects of the other factors may be undervalued (Ackerman et al., 2003; Bock et al., 1988; Hattie et al., 1996; Nandakumar & Stout, 1993). The PA results were slightly superior to the OPA results, were nonetheless unsatisfactory. Similarly, Horn (1965) reported that the

PA produced consistently determines the threshold values of important components (Beaducel, 2001; Franklin, Gibson, Robertson, Pohlmann, & Fralish, 1995; Weng & Cheng, 2005). This finding is also consistent with the parallel analysis results of data obtained from a 5-point Likert scale (Kalkan, 2014). However, when based on Pearson correlations, some of analytical predictions were underestimates. Conversely, the assumption of continuous ranking and normality are thought to be violated for small category numbers. Thus, underestimating the factor loadings and/or overestimating the number of latent variables (Gibbons et al., 2007; Nandakumar & Stout, 1993). This situation might arise from the partially skewed (0.17–0.54) data in our study. According to Weng and Cheng (2005), Horn's PA method reasonably determine unidimensionality, but is weakened when an items are skewed. Tabachnick and Fidell (2006) stated that transforming a discrete variable into a series of dichotomous variables would degrade the linear relationship between the dichotomous variables and the other variables. The multidimensionality of the PA based on tetrachoric correlation, approached the actual multidimensionality, but the results varied among the sample sizes. Replacing the tetrachoric correlation matrix with the tetrachoric correlation matrix did not sufficiently eliminate some contradictory situations. Similarly, replacing the Pearson correlation coefficient matrix with polyserial, polychoric or tetrachoric correlations cannot eliminated the problems caused by applying linear factor analysis to dichotomous or polytomous (Ackerman et al., 2003; Embretson & Reise, 2000; Hambleton & Swaminathan, 1985).

The dominant dimensions of each item measured by DETECT best matched the dimensions obtained in the NOHARM exploratory factor analysis. Finch and Habing (2005) reported that both DETECT and NOHARM correctly cluster all items and identify the correct number of dimensions. However, although both methods determine the same number of dimensions, they cluster items into different dimensions. The successful performance of both methods may originate from the high correlation between their dimensions (Finch & Habing, 2005). In simulated data, 15 values were correlated between the methods. Among these correlations, the minimum was 0.74, 5 ranged from 0.74 to 0.90, and 9 exceeded 0.90 (Finch & Habing, 2005). In our study, the interdimensional correlations varied from 0.37 to 0.5. The differences probably derive from these weak correlations.

The NOHARM confirmatory factor analyses for each sample were conducted and best model-data fits were obtained in 3 and 4 dimensions structure. The test structure were consistent in the NOHARM confirmatory factor analysis and the DETECT results (nonparametric method). To determine the best fit model, we searched for an RMSR improvement of 10% over a previous model (Tate, 2003).

Various methods and proposals for determining the sufficient sampling size for exploratory factor analysis have been reported in the literature. Among these are 'rules

of thumb' methods which taking into account only the sample size), another methods that simultaneously consider the sampling size and factor loadings, and the Kaiser–Meyer–Olkin (KMO) sampling sufficiency measure, and the Satorra–Saris, Monte Carlo methods etc. These methods determine the sample size for which the parameter estimations achieve sufficient statistical power. Parameters based on correlation coefficients estimated from small samples are less reliable. Consequently, the sample must be sufficiently large to reliably estimate the correlation coefficients. The ratios of the sample sizes provide additional information, and a more reliable model reflects the population statistics (Brown, 2006; Floyd & Widaman, 1995; Tabachnick & Fidell, 2006; Tanaka, 1987). Our analytical results depend on whether the exploratory or confirmatory factor analyses are parametric or nonparametric, and whether they are linear or nonlinear. The parametric approach to OPA could not sufficiently determine the multidimensionality of the dichotomous dataset. The results of PA based on tetrachoric correlations were more successful but varied with sample size.

NOHARM, which employs a nonlinear parametric method, determined the multidimensionality of the dichotomous dataset for sample sizes greater than 50. Sample size of 1000 and higher yielded the most stable factor structures and reveals a 4-dimensional structure. Mundfrom et al. (2005) showed that factor numbers between 3 and 6, require a minimum sampling size of 1200.

Analytical factor studies require at least 500 observations. If the number of observations exceeds 2000, the factor solutions are stable (Comrey & Lee, 1992; MacCallum, Widaman, Zhang, & Hong, 1999). Tate (2003) showed that NOHARM solutions are generally good to perfect. Similar conclusions were reported by Hambleton and Rovinelli (1986), Hattie (1984), Knol and Berger (1991), McDonald (1985) and Nandakumar (1994).

The findings of the nonparametric methods in the present study (DIMTEST t statistic, r index, DETECT, NOHARM) imply that the data of the 2008 SBS Science and Technology subtest are multidimensional. In contrast, Akın (2009) examined dimensionality of the same test by parametric methods and reported a unidimensional structure (as cited in Örs, 2010). Örs (2010) examined the 6th-grade Science and Technology subtest of 2009 SBS and concluded this subtest is unidimensional. However, in the unidimensional assumption, the variance of the first dimension is 38.53%, and eigenvalues of the remaining three factors exceed 1. The unexplained variance in this structure is 61.47%. Örs (2010) states that Akın (2009) were reported similar results. Employing nonparametric methods Özer Özkan (2012) identified the subtest of Turkish language, mathematics, science and technology, social sciences and English language in the ÖBBS of 2002, 2005 and 2008 as multidimensional. According to Lee (2007) a test that strictly measures a single latent trait is infeasible in practice, particularly when the test measures a latent construct related to human

cognition. Therefore, use of parametric and nonparametric methods in the test dimensionality studies requires further research.

In conclusion, at least 50 samples required when determining the multidimensionality of dichotomous datasets by nonparametric or nonlinear methods. Increasing the sample size reduces the RMSR and improves the fit. However, same analysis the parametric PA method based on Pearson or tetrachoric correlations requires a minimum 500 samples. According to the measured KMO sampling sufficiency, good fits obtained at a 10-fold sample size (160).

Ackerman et al. (2003) stated that all methods for evaluating dimensionality are only tools. Comparision studies have yielded promising results; however, they have been limited to simulated data. Furthermore, when analyzing the dimensionality of extensive real databases, these studies propose cross-validity studies of different samples from same population. Such cross-validity studies would improve the reliability of the evaluations. De Champlain and Gessaroli (1991) stated that if the test contains fewer than 25 items and the sample size is below 500, the power of DIMTEST is reduced. Similar conclusions were reached by Hattie et al. (1996) and Nandakumar (1994). The accuracy of DIMTEST's T statistics is affected by the sample sizes and the test length. We expect that the results of real datasets will also depend on these factors, and that the proposed analyses will significantly contribute to related studies.

References

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

Anastasi, A., & Urbina, S. (1997). *Psychological testing*. New Jersey, NJ: Prentice Hall.

Beauducel, A. (2001). Problems with parallel analysis in data sets with oblique simple structure. *Methods of Psychological Research Online, 6*(2), 141–157.

Bock, R. D., Gibbons, R., & Muraki, E. (1988). Full-information item factor analysis. *Applied Psychological Measurement*, *12*(3), 261–280.

Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. New York, NY: The Guilford Press.

Çokluk, Ö., Şekercioğlu, G., & Büyüköztürk, Ş. (2010). *Sosyal bilimler için* çok *değişkenli istatistik SPSS ve Lisrel uygulamaları* [*Multivariate statistics for the social sciences SPSS and LISREL applications*]. Ankara, Turkey: Pegem Akademi.

Comrey, A. L., & Lee, H. B. (1992). *A first course in factor analysis*. New Jersey, NJ: Hillsdale Erlbaum.

Cotton, J. W., Campbell, D. T., & Malone, R. D. (1957). The relationship between factorial composition of test items and measures of test reliability. *Psychometrika*, *22*(4), 347–357.

De Champlain, A., & Gessaroli, M. E. (1991, April). *Assessing test dimensionality using an index based on nonlinear factor analysis.* Paper presented at the annual meeting of the American Educational Research Association, Chicago.

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

Field, A. (2009). *Discovering statistics using SPSS*. London, UK: Sage.

Finch, H., & Habing, B. (2005). Comparison of NOHARM and DETECT in item cluster recovery: Counting dimensions and allocating items. *Journal of Educational Measurement*, *42*(2), 149–169.

Floyd, F. J., & Widaman, K. F. (1995). Factor analysis in the development and refinement of clinical assessment instruments. *Psychological assessment, 7*(3), 286–299.

Franklin, S. B., Gibson, D. J., Robertson, P. A., Pohlmann, J. T., & Fralish, J. S. (1995). Parallel analysis: A method for determining significant principal components. *Southern Illinois University Carbondale OpenSIUC, 6*(1)*,* 99–106.

Fraser, C. (1998). *NOHARM: A computer program for fitting both unidimensional and multidimensional normal ogive models of latent trait theory.* Armidale, New South Wales, Australia: The University of New England.

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.

Gorsuch, R. L. (1974). *Factor analysis*. Philadelphia, PA: Saunders.

Guzeller, C. O. (2012). The relationship between academic averages of primary school science and technology class and test sub-test scores of placement test of science. *Educational Sciences: Theory and Practice*, *12*, 209–214.

Hambleton, R. K., & Rovinelli, R. J. (1986). Assessing the dimensionality of a set of test items. *Applied Psychological Measurement, 10,* 287-302.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications* (Vol. 7). Boston, MA: Kluwer-Nijhoff Publishing.

Hattie, J. (1984). An empirical study of various indices for determining unidimensionality. *Multivariate Behavioral Research, 9,* 49-78.

Hattie, J., Krakowski, K., Rogers, H. J., & Swaminathan, H. (1996). An assessment of Stout's index of essential unidimensionality. *Applied psychological measurement*, *20*(1), 1–14.

Hooper, D., Coughlan, J., & Mullen, M. (2008). Structural equation modelling: Guidelines for determining model fit. *Electronic Journal of Business Research Methods 6*(1), 53–60.

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, *30*, 179–185. http://dx.doi.org/10.1007/BF02289447

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural equation modeling: a multidisciplinary journal*, *6*(1), 1–55.

Jang, E. E., & Roussos, L. (2007). An investigation into the dimensionality of TOEFL using conditional covariance-based nonparametric approach. *Journal of Educational Measurement*, *44*(1), 1–21.

Jöreskog, K. G., & Sörbom, D. (1989). *LISREL 7: User's reference guide.* Mooresville, IN: Scientific Software.

Kalkan, Ö. K. (2014). Mesleki Eğitime Yönelik Tutum Ölçeği geçerlik ve güvenirlik çalışması [A study of reliability and validity an attitude scale towards vocational education]. *Trakya Üniversitesi Eğitim Fakültesi Dergisi*, *4*(1), 117–128.

Kline, R. B. (2011). *Principles and practice of structural equation modeling* (3rd ed.). New York, NY: The Guilford Press.

Knol, D. L., & Berger, M. P. F. (1991). Empirical comparison between factor analysis and multidimensional item response models. *Multivariate Behavioral Research, 26,* 457-477.

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

Lee, S. H. (2007). *Multidimensional item response theory: A SAS MDIRT macro and empirical study of PIAT MATH Test* (Doctoral dissertation, The University of Oklahoma, Oklahoma). Retrieved from https://shareok.org/bitstream/handle/11244/1156/3255213.PDF?sequence=1&isAllowed=y

Leech, N. L., Barrett, K. C., & Morgan, G. A. (2005). *SPSS for intermediate statistics: Use and interpretation*. Mahwah, NJ: Lawrence Erlbaum Associates.

Lorenzo-Seva, U., & Ferrando, P. J. (2006). FACTOR: A computer program to fit the exploratory factor analysis model. *Behavior Research Methods, Instruments, & Computers, 38,* 88–91.

Lorenzo-Seva, U., & Ferrando, P. J. (2013). *Manual of the Program Factor v.9.20.* Departament de Psicologia, Universitat Rovira i Virgili, Tarragona, Spain.

MacCallum, R. C., Widaman, K. F., Zhang, S., & Hong, S. (1999). Sample size in factor analysis. *Psychological methods*, *4*(1), 84–99.

Maydeu-Olivares, A. (2001). Multidimensional item response theory modeling of binary data: Large sample properties of NOHARM estimates. *Journal of Educational and Behavioral Statistics*, *26*(1), 51–71.

McDonald, R. P. (1967). Numerical methods for polynomial models in nonlinear factor analysis. *Psychometrika*, *32*(1), 77–112.

McDonald, R. P. (1981). The dimensionality of tests and items. *British Journal of Mathematical and Statistical Psychology*, *34*(1), 100–117.

McDonald, R. P. (1985). *Factor analysis and related methods.* Hillsdale, NJ: Lawrence Erlbaum.

McDonald, R. P. (1999). *Test theory: A unified approach*. Mahwah, NJ: Lawrence Earlbaum.

McDonald, R. P. (2000). A basis for multidimensional item response theory. *Applied Psychological Measurement*, *24*(2), 99–114.

Milli Eğitim Bakanlığı. (2008). *2008 SBS (6. ve 7. sınıflar) istatistiki bilgileri* [Statistical information of 2008 SBS (6th and 7th grades)]. Retrieved from http://www.memurlar.net/common/news/documents/116328/1.pdf

Milli Eğitim Bakanlığı. (2011). *64 soruda ortaöğretime geçiş (OGES) sistemi ve seviye belirleme sınavı örnek sorular* [64 questions about passing to secondary education system and placement exam sample questions] Retrieved from http://file.setav.org/Files/Pdf/64-soruda-ortaogretime-gecis-sistemi---meb.pdf

Mundfrom, D. J., Shaw, D. G., & Ke, T. L. (2005). Minimum sample size recommendations for conducting factor analyses. *International Journal of Testing*, *5*(2), 159–168.

Nandakumar, R. (1994). Assessing dimensionality of a set of item responses-comparison of different approaches. *Journal of Educational Measurement*, *31*(1), 17–35.

Nandakumar, R., & Stout, W. (1993). Refinements of Stout's procedure for assessing latent trait unidimensionality. *Journal of Educational and Behavioral Statistics*, *18*(1), 41–68.

Nargundkar, R. (2008). *Marketing research: Text and cases 3E*. New Delhi: Tata McGraw-Hill Education.

Norman, G. R., & Streiner, D. L. (2003). *PDQ* statistics (Vol. 3). Toronto: BC Decker Inc.

Örs, S. (2010). *6., 7. ve 8. sınıf seviye belirleme sınavı fen ve teknoloji alt testlerinin faktör yapılarının belirlenmesi* [Determination of factor structures of science and technology sub tests in level determination exams of 6th, 7th and 8th grades] (Master's thesis, Ankara University, Ankara, Turkey). Retrieved from https://tez.yok.gov.tr/UlusalTezMerkezi/

Özbek Baştuğ, Ö. Y. (2012). Assessment of dimensionality in social science subtest. *Educational Sciences: Theory & Practice*, *12*, 382–385.

Özer Özkan, Y. (2012). *Klasik test kuramı, tek boyutlu ve çok boyutlu madde tepki kuramı modellerinden kestirilen öğrenci başarısı belirleme sınavı (ÖBBS) başarı ölçülerinin karşılaştırılması* [A comparison of estimated achievement scores obtained from student achievement assessment test utilizing classical test theory, unidimensional and multidimensional item response theory models]. (Doctoral dissertation, Ankara University, Ankara, Turkey). Retrieved from https://tez.yok.gov.tr/UlusalTezMerkezi/

Özer Özkan, Y., & Acar Güvendir, M. (2014). Türkiye'de uygulanan geniş ölçekli testlerin çok boyutluluğunun analizi [The analysis of large scale tests applied in Turkey in terms of their multidimensionality]. *Mehmet Akif Ersoy Üniversitesi Eğitim Fakültesi Dergisi*, *1*(29), 31–47.

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

Stevens, J. P. (2009). *Applied multivariate statistics for the social sciences*. New York, NY: Routledge.

Stout, W., Habing, B., Douglas, J., Kim, H. R., Roussos, L., & Zhang, J. (1996). Conditional covariance-based nonparametric multidimensionality assessment. *Applied Psychological Measurement, 20*(4), 331–354.

Tabachnick, B. G., & Fidell, L. S. (2006). *Using multivariate statistics*. Boston, MA: Pearson Education.

Tanaka, J. S. (1987). How big is big enough? Sample size and goodness of fit in structural equation models with latent variables. *Child Development*, *58*, 134–146.

Tate, R. (2003). A comparison of selected empirical methods for assessing the structure of responses to test items. *Applied Psychological Measurement*, *27*(3), 159–203.

Thomas, J.M. (2003). MultiTrait-MultiMethod matrices to study bias in social measurement. In J. Z. Arlsdale (Ed.), *Trends in social psychology* (pp. 138–148). New York, NY: Nova Publishers.

Thompson, B. (2004). *Exploratory and confirmatory factor analysis: Understanding concepts and applications*. Washington, DC: American Psychological Association. http://dx.doi.org/10.1037/10694-000

Timmerman, M. E., & Lorenzo-Seva, U. (2011). Dimensionality assessment of ordered polytomous items with parallel analysis. *Psychological Methods*, *16*(2), 209–220.

Van Der Linden, W. J., & Hambleton, R. K. (1997). *Handbook of modern item response theory*. New York, NY: Springer Science+Business Media.

Velicer, W. F. (1976). The relation between factor score estimates, image scores, and principal component scores. *Educational and Psychological Measurement*, *36*(1), 149–159.

Vieira, A. L. (2011). *Interactive LISREL in practice*. New York, NY: Springer.

Waller, N. G., Tellegen, A., McDonald, R. P., & Lykken, D. T. (1996). Exploring nonlinear models in personality assessment: Development and preliminary validation of a negative emotionality scale. *Journal of Personality*, *64*(3), 545–576.

Weng, L. J., & Cheng, C. P. (2005). Parallel analysis with unidimensional binary data. *Educational and Psychological Measurement*, *65*(5), 697–716.

Westland, J. C. (2015). *Structural equation models*. Switzerland: Springer.

Yang, Y., & Green, S. B. (2011). Coefficient Alpha: A Reliability Coefficient for the 21st Century? *Journal of Psychoeducational Assessment, 29*(4), 377–392.

Zhang, J., & Stout, W. (1999). Conditional covariance structure of generalized compensatory multidimensional items. *Psychometrika*, *64*(2), 129–152.

Zwick, W. R., & Velicer, W. F. (1986). Comparison of five rules for determining the number of components to retain. *Psychological Bulletin*, *99*(3), 432–442.