International Journal of Assessment Tools in Education
2019, Vol. 6, No. 1, 25–35

DOI: 10.21449/ijate.442542

Published at http://www.ijate.net http://dergipark.gov.tr/ijate *Research Article*

Can Factor Scores be Used Instead of Total Score and Ability Estimation?

Abdullah Faruk Kilic [1,*]

[1] Department of Educational Measurement and Evaluation, Hacettepe University, Ankara, Turkey

ARTICLE HISTORY

Received: 11 July 2018
Revised: 20 October 2018
Accepted: 07 December 2018

KEYWORDS

Factor Score,
Ability Estimation,
Classical Test Theory,
Item Response Theory,
Total Score

Abstract: The purpose of this study is to investigate whether factor scores can be used instead of ability estimation and total score. For this purpose, the relationships among total score, ability estimation, and factor scores were investigated. In the research, Turkish subtest data from the Transition from Primary to Secondary Education (TEOG) exam applied in April 2014 were used. Total scores in this study were calculated from the total number of correct answers given by individuals to each item. Ability estimations were obtained from a three-parameter logistic model chosen from among item response theory (IRT) models. The Bartlett method was used for factor score estimation. Thus, the ability estimation, sum, and factor scores of each individual were obtained. When the relationship between these variables was investigated, it was observed that there was a high-level, positive, and statistically significant relationship. In the result section of this study, as variables have a high-level relationship, it was suggested that since variables could be used interchangeably, factor scores should be used. Although the total scores of individuals were equal, there were differences in terms of factor score and ability estimations. Therefore, it was suggested that item response theory assumptions were not met, or factor scores should be used when the sample size is small.

1. INTRODUCTION

Scales or achievement tests are commonly used to measure psychological traits of individuals. Based on the data obtained from such measurement tools, scores (ability scores) are obtained for individuals with different methods. There are various scoring methods for determining the level of individuals being measured. Classical test theory (CTT), item response theory (IRT), and factor scores are among these methods.

In classical test theory (CTT), the total number of correct answers given by individuals to items is generally preferred as a scoring method. Typically, the observed scores of individuals is referred to as the total number of correct answers to items (de Ayala, 2009; Price, 2017). Additionally, in CTT, options and items can be weighted, and different scoring methods can be

CONTACT: Abdullah Faruk Kılıç ✉ afarukkilic@windowslive.com Department of Educational Measurement and Evaluation, Hacettepe University, Ankara, Turkey

ISSN-e: 2148-7456 /© IJATE 2019

used. However, it is known that the contribution of these methods in terms of reliability and validity is not high, and efforts are higher than contributions (Gulliksen, 1950).

In IRT, unlike CTT, a non-linear relationship is formed between the answers of individuals to items and their abilities (DeMars, 2010; Hambleton & Swaminathan, 1985). In IRT, item discrimination and item difficulty can affect the estimation of abilities of individuals depending on the selected IRT model. In unidimensional IRT models, there are assumptions such as unidimensional, local independence, and the S-shape of item characteristic function (de Ayala, 2009). By investigating which IRT model (Rasch, one-, two-, and three-parameter logistic models) fits the data, abilities are estimated via that IRT model.

Another scoring method is factor score estimation. Factor score estimation can be divided into two main sections: 1) nonrefined methods and 2) refined methods. Among nonrefined methods, sum methods calculated based on CTT are included (DiStefano, Zhu, & Mîndrilă, 2009). The easiest way to obtain factor scores is to sum raw scores of items relating to item loadings (Comrey & Lee, 1992). An important point to be considered here is to subtract item scores when factor loadings are negative (DiStefano et al., 2009). Another method that is classified as nonrefined and used for estimating factor scores is to sum certain items with factor loadings above a certain threshold. Another method is to use the sum of the standardized scores. Additionally, the sum can be calculated with weighted factor loadings of items. However, in this case, the measurement tool has to be unidimensional (DiStefano et al., 2009).

Refined methods applied in obtaining factor scores can be listed as the regression method, the Bartlett method, and the Anderson-Rubin method. In the regression method, the least squares method is used to obtain the factor score for each individual regarding factor or component. Factor scores are used as dependent variables in regression equations. In the Bartlett method, only common factors influence factor scores. In this method, squares of error variance of variables are minimized. It has been stated that the Bartlett method is unbiased for estimating real factor scores (Hershberger, 2005). The Anderson-Rubin method (Anderson & Rubin, 1956) is derived from the Bartlett method. In this method, factor scores are obtained as unrelated to both other factors and to each other. This method involves more complex calculation processes than the Bartlett method where the factor score is orthogonal, the average is 0, and the standard deviation is 1 (DiStefano et al., 2009).

Factor score estimation methods have some advantages. For example, since the correlation between factor score and factors is maximum in the regression method, it has been stated that more valid results are obtained. On the other hand, the Bartlett method can estimate factor score in an unbiased way. In the Anderson-Rubin method, factor scores obtained from two orthogonal factors can be unrelated (DiStefano et al., 2009). It can be said that factor scores obtained from these methods have a high-level relationship (Hershberger, 2005; Horn, 1965). There is also a factor score indeterminacy problem while estimating factor scores. When the total of common and unique variance exceeds the number of items, since the matrix formed from these elements is not a square matrix, the inverse of the matrix cannot be calculated. In this case, factor indeterminacy arises (Grice, 2001).

Generally, total scores are used in scales or achievement tests to decide about individuals. If analysis is conducted based on IRT, the ability parameter is estimated. Estimating factor scores is limited in studies. However, when the total score is considered, item characteristics have no effect on the ability of individuals. The fact that item characteristics influence individuals' ability estimation is seen as an advantage that IRT has over CTT (DeMars, 2010). Therefore, ability estimations are affected by item characteristics. Thus, the effects of items with strong psychometric properties on ability estimates are different. A similar situation is observed while estimating factor scores. Factor scores can be calculated by using factor loadings, unique

variances, regression weights, eigenvalues, and eigenvectors (DiStefano et al., 2009). Therefore, the abilities of individuals can be estimated more accurately.

When the literature is reviewed and item and ability parameters obtained from CTT and IRT are compared, there are studies analyzing parameter invariance (Akyıldız & Şahin, 2017; Bulut, 2018; Çakıcı-Eser, 2013; Cappelleri, Jason Lundy, & Hays, 2014; Çelen & Aybek, 2013; İlhan, 2016; Macdonald & Paunonen, 2002; Stage, 1998a, 1998b; Xu & Stone, 2012). In these studies, generally parameters obtained from IRT and CTT are compared and invariance property is generally obtained in IRT. But it can be said that parameter invariance is hold in CTT with larger samples. However, there are studies considering factor score estimation methods (DiStefano et al., 2009; Green, 1976; Hershberger, 2005; Horn, 1965; Williams, 1978). In these studies, factor score estimations are introduced and, in particular, factor score indeterminacy is emphasized. In addition to these studies, the relationship between factor score and scale scores (Fava & Velicer, 1992) or factor scores obtained from different factor extraction methods were compared with the scale score (Fava & Velicer, 1992; Grice, 2001; Velicer, 1976). These studies were generally conducted as a simulation study. In the current study, the aim is to investigate whether factor scores can be used instead of ability estimation and total score with high-stakes test data. Therefore, the current study investigated whether factor scores can be used instead of ability estimation and total score. Accordingly, in this study, the answer to the question "According to the relationship between total score, ability estimation, and factor score, can factor scores be used instead of ability estimation and total score?" was investigated. Therefore, with the help of the relationship between these variables, suitability of scores for deciding about individuals was discussed.

2. METHOD

In this study, conducted to analyze the relationship between total score, ability estimation, and factor score, the research design was a relational study. In relational studies, relationships and connections are investigated (Büyüköztürk, Kılıç-Çakmak, Akgün, Karadeniz, & Demirel, 2013; Fraenkel, Wallen, & Huyn, 2012). In correlational studies among relational studies, correlations between two or more variables/scores are analyzed (Creswell, 2013). In this study, since the relationship between total score, ability estimation, and factor score was analyzed, the correlational research method was selected from among relational research methods.

2.1. Study Group

In this study, data obtained from the Turkish Test in Transition from Primary to Secondary Education (TEOG) exam applied in April 2014 were used. Accordingly, from among 1,271,284 students, 10,000 students were sampled using simple random sampling. Based on this information, it can be stated that the sampling method of this study was simple random sampling (Büyüköztürk et al., 2013). Data cleaning was applied by analyzing a 10,000-sample data set. Accordingly, data of individuals with repetitive answers or who gave the same answers to each question were deleted and analyses were conducted on 9,773 student data.

2.2. Data Collection Method

Data used in this study were collected from the Ministry of National Education, Measurement, Evaluation, and Exam Services General Directorate. Sampling for the data used in this study was randomly performed by the Measurement, Evaluation, and Exam Services General Directorate and a data set including 10,000 students was given to the researcher. The researcher conducted the data cleaning process.

2.3. Process

In this study, the construct of the data set was analyzed first. For this purpose, the data set was randomly divided into two parts, and while exploratory factor analysis (EFA) was applied to

one half, confirmatory factor analysis (CFA) was applied to the other. For EFA, it was first analyzed whether the data set met EFA assumptions. Accordingly, for multivariate normality, the skewness and kurtosis coefficients of Mardia (1970) were analyzed. Multivariate skewness and kurtosis coefficients showed that the data did not hold the assumption of multivariate normality (p<0.01). Therefore, the principal axis factoring method was adopted, which is stronger in terms of violation of the normality assumption (Costello & Osborne, 2005; Fabrigar, Wegener, MacCallum, & Strahan, 1999). When the adequacy of the sample size was analyzed, it was concluded that a sample of 4,886 was sufficient for the majority of researchers (Comrey, 1988; Floyd & Widaman, 1995; Gorsuch, 1974; Guadagnoli & Velicer, 1988; Kaiser & Rice, 1974; Leech, Barrett, & Morgan, 2015; Streiner, 1994). Additionally, it was observed that the KMO value was 0.95. Accordingly, it can be concluded that the sample was adequate for factor analysis and that an adequate number of items corresponded to each factor (Kaiser & Rice, 1974; Leech et al., 2015). The Bartlett test results, which analyzed whether the correlation matrix was different to the identity matrix, showed that it was ($\chi^2(190) = 21775.9$, p<0.01). On the other hand, Mahalanobis distances were calculated to analyze multivariate outliers. Among the 4,886 data in this sample, 145 Mahalanobis distances that provided significant results at the α=0.001 level were deleted, and a data set of 4,741 people was obtained. For the multicollinearity assumption, the variance inflation factor (VIF), tolerance value (TV), and conditional index (CI) were analyzed since there should be no multicollinearity. It was observed that the tolerance value was larger than 0.01, the VIF value was smaller than 10, and the CI value was smaller than 30. Accordingly, it can be concluded that there was no multicollinearity problem (Kline, 2016; Tabachnik & Fidell, 2012).

Tetrachoric correlation matrix and principal axis factoring as factor extraction methods were applied to data divided into two for EFA. First, parallel analysis was conducted to determine the number of factors and the analysis proposed a unidimensional construct. On the other hand, when the scree plot and eigenvalues were analyzed, it was observed that only the eigenvalue of the first factor was larger than one. The unidimensional construct explained 47.96% of total variance. Therefore, it was decided that the test was unidimensional. When factor loadings were analyzed, it was observed that loadings changed between 0.44 and 0.79. Accordingly, it can be concluded that a unidimensional structure was defined as the result of EFA.

Confirmatory factor analysis (CFA) was applied to the second half of the data set. Again, assumptions of the analysis were investigated. As it was observed that multivariate normal distribution was not held, tetrachoric correlation matrix and weighted least squares means and variance adjusted (WLSMV) estimation methods were applied for CFA (Li, 2016). On the other hand, Mahalanobis distances were calculated to analyze multivariate outliers. From among the 4,887 data in this sample, 69 Mahalonbis distances that provided significant results at the α=0.001 level were deleted, and a data set of 4,818 people was obtained. For multicollinearity assumption, the variance inflation factor (VIF), tolerance value (TV), and conditional index (CI) were analyzed since there should be no multicollinearity. It was observed that the TV was larger than 0.01, the VIF value was smaller than 10, and the CI value was smaller than 30. Accordingly, it can be concluded that there was no multicollinearity problem (Kline, 2016; Tabachnik & Fidell, 2012).

The results of CFA applied to the second half showed significant chi-square values ($\chi^2(170) = 753.45$, p<0.01). Accordingly, it can be said that the model data fit was not held. However, since this statistic has a tendency to be significant and high in large samples (Mueller, 1996), other fit statistics were examined. Accordingly, CFI and TLI values were observed as being 0.99. Additionally, all factor loadings had statistically significant t-values between 0.44 and 0.78. Error variances changed between 0.40 and 0.80. Based on these results, it can be concluded that data fitted with the unidimensional construct.

After determining that the data set was unidimensional, the total score, IRT ability estimation, and factor scores of each individual were calculated. The total scores of individuals were calculated from the total number of correct answers of individuals to each item. For ability estimation based on IRT, it was analyzed whether the data set held IRT assumptions (unidimensional, local independence and S-shape) (DeMars, 2010; Hambleton & Swaminathan, 1985; Lord, 1980). It was observed that unidimensional assumptions were held when the factor structure was investigated. Yen's (1984) Q_3 statistic was used to determine whether the local independence assumption was held. For this purpose, the model data fit was analyzed and which unidimensional IRT model (1-, 2-, or 3-parameter logistic model) fitted the data set was investigated. Accordingly, log likelihood values were examined. When models were compared, it was found that the three-parameter logistic model (3PLM) fitted data better ($\chi^2_{2PLM\text{-}3PLM}(20) = 1352.55$, $p<0.01$). Item parameters were estimated based on 3PLM, the residual matrix was created with residuals of each item, and the correlation between them was analyzed. It was observed that correlations were not higher than the 0.20 threshold value. Accordingly, it can be concluded that the local independence assumption was held (DeMars, 2010). Whether item characteristic functions were S-shaped was analyzed by plotting the item characteristic curve and it was observed that they had an S-shape. Ability estimation of individuals was obtained with the expected a posteriori (EAP) method for IRT.

After obtaining total score and ability estimations, the Bartlett method was used to estimate factor scores. The Bartlett method was selected as it is unbiased when estimating real factor scores (Hershberger, 2005). After obtaining the total score, ability estimations, and factor scores of individuals, the relationship between three variables was analyzed using correlation analysis. Additionally, a scatter plot was used to visually represent the relationship between variables.

2.4. Data Analysis

In this study, to estimate EFA and factor scores, the psych package (Revelle, 2018) in R software (R Core Team, 2017) was used. M*plus* software (Muthén & Muthén, 2012) was used for CFA. IRT parameter estimations were performed using the irtoys package (Partchev, 2016) in R, and the BILOG engine. The sirt package (Robitzsch, 2017) was used to test the local independence assumption. ggplot2 (Wickham, 2016) in R software was used for plotting the graphs.

3. FINDINGS

In this study, the answer to the question "According to the relationship between total score, ability estimation, and factor score, can factor scores be used instead of ability estimation and total score?" was investigated. Accordingly, the relationship between total score, ability estimation, and factor scores was analyzed and presented in Table 1.

Table 1. *The relationship between total score, ability estimation, and factor score*

Variables	$\bar{X}$	S	Skewness	Kurtosis	Correlation		
					Total score	Ability Estimation	Factor Score
Total score	14.30	4.75	-0.64	-0.59	1		
Ability Estimation	0.00	0.93	-0.20	-0.71	0.975**	1	
Factor Score	0.00	1.23	-1.00	0.25	0.982**	0.960**	1

**p<0.01

When Table 1 was examined, a correlation between the total score, ability estimation, factor score, and descriptive statistics of variables was observed. Since the scales of total score, ability estimation, and factor score were different, it can be concluded that mean and standard deviation

values were different. When skewness and kurtosis values were examined, it was observed that skewness values were between -1.00 and -0.20 while kurtosis values were between -0.71 and 0.25. Accordingly, it can be concluded that variables have a univariate normal distribution (Byrne, 2016; Chou & Bentler, 1995; Curran, West, & Finch, 1996; Finney & DiStefano, 2013). Therefore, correlations between variables were calculated using the Pearson Product Moment (PPM) correlation coefficient. When correlations between variables were analyzed, it could be stated that there was a positive and high-level relationship. A scatter plot of the total score and ability estimation of individuals is presented in Figure 1.

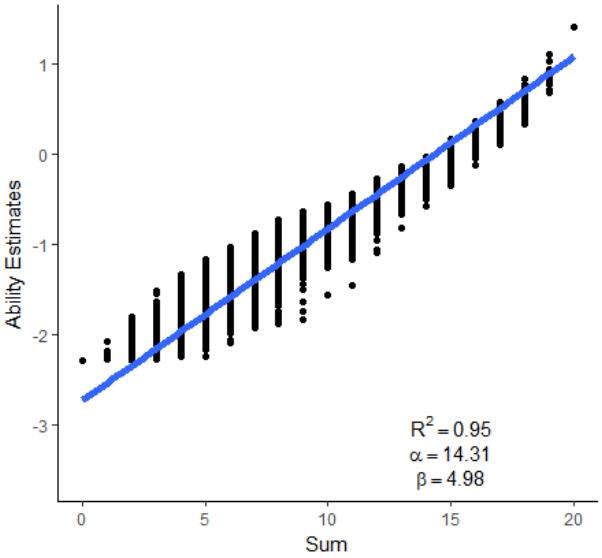

Figure 1. Total score and ability estimate distribution.

When Figure 1 is examined, the distribution of total score and ability estimation can be observed. The explained variance on the obtained linear regression equation was 95%. Accordingly, it can be stated that the variation in total score explained 95% of the variation in ability estimation. On the other hand, there was differentiation in the ability estimation for each total score category. For example, many students with a total score of 5 differed in ability estimation. A scatter plot of total scores and factor scores is presented in Figure 2.

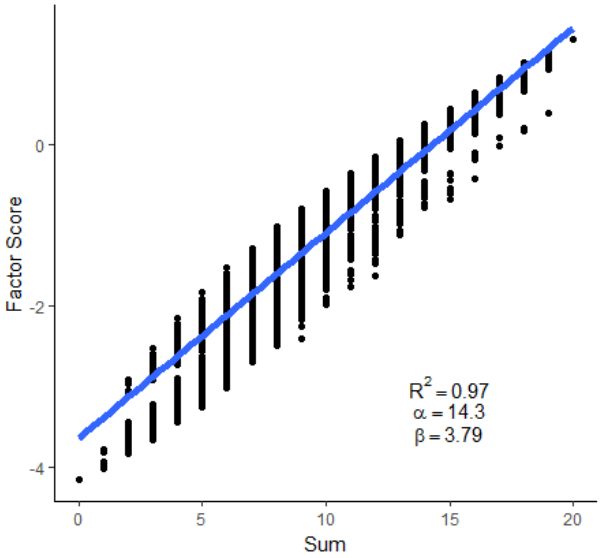

Figure 2. Total score and factor score distribution.

When Figure 2 is examined, the distribution of total score and factor score can be observed. The explained variance on the obtained linear regression equation was 97%. Accordingly, it can be stated that the variation in the total score explained 97% of the variation in the factor score. On the other hand, there was differentiation in the factor score for each total score category. For example, many students with a sum of 10 differed in their factor score. A scatter plot of ability estimation and factor score is presented in Figure 3.

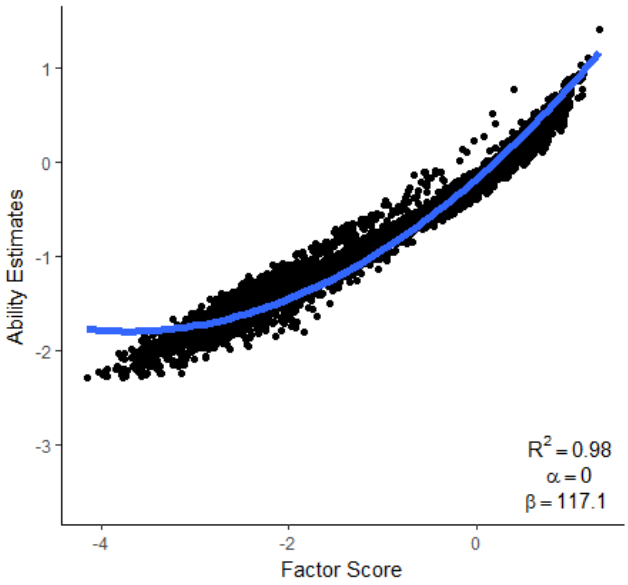

Figure 3. Factor score and ability estimates distribution.

When Figure 3 is examined, it can be concluded that the relationship between total score, ability estimation, and factor score was nonlinear. While the linear relationship presented in Table 1 explained 92.16% (0.962=0.921) of the variance, when the relationship between the two variables was considered as quadratic, 98% of the variance was explained. Accordingly, it can be stated that the relationship between the factor score and ability estimation was high.

4. DISCUSSION AND CONCLUSION

In the results of this study, a positive, statistically significant, and strong relationship between ability estimation and total score was observed. This finding is similar to findings of other studies in the literature (Bulut, 2018; Çelen & Aybek, 2013; Fava & Velicer, 1992; Grice, 2001; Macdonald & Paunonen, 2002; Progar & Sočan, 2008; Velicer, 1976). Based on this result, it can be said that the factor score, ability estimates, and total score obtained from the high-stakes achievement test are strongly related to each other. Due to the high relationship between the two variables, it is claimed that these variables can be used interchangeably. Tabachnik and Fidell (2012) stated that the relationship between the variables was 0.90 or higher, one of the variables was redundant, or this variable is a combination of other variables. From this perspective, a 0.975 value of correlation between variables showed that instead of total score, ability estimation can be used. When ability estimation was used, in contrast to total score, individual ability estimations in one category of sum differed. In this case, IRT influences ability estimation by using item parameters when estimating ability.

When the relationship between total score and factor scores was analyzed, it could be concluded that a similar high, positive, and statistically significant relationship was present for ability estimation. Similarly, in ability estimation, individuals with the same total score had different factor scores.

It was observed that the relationship between factor score and ability estimation was nonlinear. The relationship between these two variables signified a quadratic function and, in this case, it was observed that explained variance was extremely high. In other words, it could be claimed that factor score and ability estimation can be used interchangeably.

Total score is practical in terms of calculation and interpretation. On the other hand, if ability estimation based on IRT is used, real abilities of individuals can be estimated. However, large samples are needed to hold IRT assumptions. DeMars (2010) stated that as the number of parameters increases, and as the ability distribution of the group moves away from normality, the sample size must grow. When the number of items is 20 and discrimination parameters are high, a sample size of at least 500 is needed. If the pseudo chance parameter is estimated, the sample size should be at least 2000 (DeMars, 2010). When these limitations of IRT is considered, it can be expressed that factor scores are more efficient. In many cases, factor scores can be estimated from a sample size of 250. Floyd and Widaman (1995) stated that there should be four or five individuals per item and the sample size should be as large as possible. Streiner (1994) suggested that each item should contain five individuals and the sample should not be smaller than 100. If the sample size should be smaller than 100, 10 individuals should be sampled for each item. Gorsuch (1974) suggested that each item should contain five individuals and the sample size should not be smaller than 200. Guadagnoli and Velicer (1988) stated that these calculations were baseless and EFA can be applied a sample size smaller than 50 when factor loadings are 0.80 or higher. Comrey (1988) stated that if the number of items did not exceed 40, a sample size of 200 individuals would be sufficient. When all these recommendations of these researchers were considered, it could be concluded that EFA needs a smaller sample size than IRT. Additionally, mainly EFA is conducted for the unidimensionality assumption of the IRT assumption evaluation stage. Therefore, it could be concluded that the practicability of factor score estimation is better than IRT ability estimation.

Based on the findings of this study, since it was observed that total score, ability estimation, and factor score can be used interchangeably, factor score is recommended. Because factor score requires smaller sample size, discriminates individuals better than total score, and produces very close result with IRT ability estimation. On the other hand, Erkuş (2014) suggested item weighting via factor loadings. He recommends that individual responses to items and that item factor loading multiplication (if individual score is 1 and factor loading is 0.48 the item score is 1x0.48 = 0.48 for that individual) for calculate total score. But factor scores are also calculated with factor loadings as well as other elements of factor analysis such as eigenvalues, communalities, and error variance. In this sense, using factor scores can be suggested. After estimating factor scores, linear transformation or T points can be used for reporting and results can be interpreted more easily.

The current study is limited by unidimensional constructs. Therefore, multidimensional constructs may be studied in future studies. On the other hand, since the data set of the current study was extremely large, a smaller sample size can be investigated in another study. As a result of this study, since it is thought that factor scores will make a positive contribution to the validity of decisions regarding individuals, the use of factor scores is suggested.

ORCID

Abdullah Faruk Kilic https://orcid.org/0000-0003-3129-1763

5. REFERENCES

Akyıldız, M., & Şahin, M. D. (2017). Açıköğretimde kullanılan sınavlardan Klasik Test Kuramına ve Madde Tepki Kuramına göre elde edilen yetenek ölçülerinin karşılaştırılması. *Açıköğretim Uygulamaları ve Araştırmaları Dergisi*, *3*(4), 141–159.

Anderson, T. W., & Rubin, H. (1956). Statistical inference in factor analysis. In J. Neyman (Ed.), *Proceedings of the Third Berkeley Symposium of Mathematical Statistics and Probability* (Vol. 5, pp. 111–150). Berkeley: University of California Press.

Bulut, G. (2018). Açık ve uzaktan öğrenmede şans başarısı : Klasik Test Kuramı (KTK) ve Madde Tepki Kurama (MTK) temelinde karşılaştırmalı bir analiz. *Açıköğretim Uygulamaları ve Araştırmaları Dergisi*, *4*(1), 78–93.

Büyüköztürk, Ş., Kılıç-Çakmak, E., Akgün, Ö. E., Karadeniz, Ş., & Demirel, F. (2013). *Bilimsel araştırma yöntemleri* (14. Baskı). Ankara: Pegem Akademi.

Byrne, B. M. (2016). *Structural equation modeling with AMOS: Basic concepts, applications, and programming* (3rd ed.). Routledge.

Çakıcı-Eser, D. (2013). PISA 2009 okuma testinden elde edilen iki kategorili verilerin BILOG programı ile incelenmesi. *Eğitim ve Öğretim Araştırmaları Dergisi*, *2*(4), 135–144.

Cappelleri, J. C., Jason Lundy, J., & Hays, R. D. (2014). Overview of Classical Test Theory and Item Response Theory for the quantitative assessment of items in developing patient-reported outcomes measures. *Clinical Therapeutics*, *36*(5), 648–662. https://doi.org/10.1016/j.clinthera.2014.04.006

Çelen, Ü., & Aybek, E. C. (2013). Öğrenci başarısının öğretmen yapımı bir testle klasik test kuramı ve madde tepki kuramı yöntemleriyle elde edilen puanlara göre karşılaştırılması. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi*, *4*(2), 64–75. Retrieved from http://dergipark.ulakbim.gov.tr/epod/article/view/5000045503

Chou, C. P., & Bentler, P. M. (1995). Estimates and tests in structural equation modeling. In R. H. Hoyle (Ed.), *Structural equation modeling: Concepts, issues, and applications*. Thousand Oaks, CA: SAGE.

Comrey, A. L. (1988). Factor-analytic methods of scale development in personality and clinical psychology. *Journal of Consulting and Clinical Psychology*, *56*(5), 754–761. https://doi.org/10.1037/0022-006X.56.5.754

Comrey, A. L., & Lee, H. B. (1992). *A first course in factor analysis* (2nd ed.). London: Lawrence Erlbaum Associates, Inc. https://doi.org/10.1017/CBO9781107415324.004

Costello, A. B., & Osborne, J. W. (2005). Best practices in explatory factor analysis: Four recommendations for getting the most from your analysis. *Practical Assessment, Research & Evaluation*, *10*(7), 27–29. https://doi.org/10.1.1.110.9154

Creswell, J. W. (2013). *Research design: Qualitative, quantitative, and mixed Methods approaches* (4. Edition). Thousand Oaks, CA: Sage Publications.

Curran, P. J., West, S. G., & Finch, J. F. (1996). The robustness of test statistics to nonnormality and specification error in confirmatory factor analysis. *Psychological Methods*, *1*(1), 16–29. https://doi.org/10.1037/1082-989X.1.1.16

de Ayala, R. J. (2009). *The theory and practice of item response theory*. New York, NY: The Guilford Press. https://doi.org/10.1073/pnas.0703993104

DeMars, C. (2010). *Item response theory*. New York: Oxford University Press.

DiStefano, C., Zhu, M., & Mîndrilă, D. (2009). Understanding and using factor scores: Considerations for the applied researcher. *Practical Assessment, Research & Evaluation*, *14*(20), 1–11.

Erkuş, A. (2014). *Psikolojide ölçme ve ölçek geliştirme-I: Temel kavramlar ve işlemler* (2. Baskı). Ankara: Pegem Akademi.

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Pscyhological Methods*, *4*(3), 272–299. https://doi.org/10.1037/1082-989X.4.3.272

Fava, J. L., & Velicer, W. F. (1992). An empirical comparison of factor, image, component, and scale scores. *Multivariate Behavioral Research*, *27*(3), 301–322.

https://doi.org/10.1207/s15327906mbr2703_1

Finney, S. J., & DiStefano, C. (2013). Nonnormal and categorical data in structural equation modeling. In G. R. Hancock & R. O. Mueller (Eds.), *Structural equation modeling: A second course* (2nd ed., pp. 439–492). Charlotte, NC: IAP.

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

Fraenkel, J. R., Wallen, N. E., & Huyn, H. H. (2012). *How to design and evaluate research in education* (8. Edition). New York: McGraw-Hill.

Gorsuch, R. L. (1974). *Factor analysis* (1st ed.). Toronto: W. B. Saunders Company.

Green, B. F. (1976). On the factor score controversy. *Psychometrika*, *41*(2), 263–266. https://doi.org/10.1007/BF02291843

Grice, J. W. (2001). Computing and evaluating factor scores. *Psychological Methods*, *6*(4), 430–450. https://doi.org/10.1037//1082-989X.6.4.430

Guadagnoli, E., & Velicer, W. F. (1988). Relation of sample size to the stability of component patterns. *Psychological Bulletin*, *103*(2), 265–275.

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

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and appliccations*. New York: Springer Science & Business Media, LLC.

Hershberger, S. L. (2005). Factor score estimation. In B. S. Everitt & D. C. Howell (Eds.), *Encyclopedia of Statistics in Behavioral Science* (Vol. 2, pp. 636–644). Chichester, UK, UK: John Wiley & Sons, Ltd. https://doi.org/10.1002/0470013192.bsa726

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, *30*(2), 179–185. https://doi.org/10.1007/BF02289447

İlhan, M. (2016). Açık uçlu sorularda yapılan ölçmelerde Klasik Test Kuramı ve çok yüzeyli Rasch modeline göre hesaplanan yetenek kestirimlerinin karşılaştırılması. *Hacettepe Un iversity Journal of Education*, *31*(2), 346- 368. https://doi.org/10.16986/HUJE.2016015182

Kaiser, H. F., & Rice, J. (1974). Little Jiffy, Mark IV. *Educational and Psychological Measurement*, *34*(1), 111–117. https://doi.org/10.1177/001316447403400115

Kline, R. B. (2016). *Principle and practice of structural equation modelling* (4th ed.). New York, NY: The Guilford Press.

Leech, N. L., Barrett, K. C., & Morgan, G. A. (2015). *IBM SPSS for intermediate statistics* (5. Baskı). East Sussex: Routledge.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. *Journal of Chemical Information and Modeling* (Vol. 53). New Jersey: Lawrence Erlbaum Associates, Inc. https://doi.org/10.1017/CBO9781107415324.004

Macdonald, P., & Paunonen, S. V. (2002). A monte carlo comparison of item and person statistics based on item response theory versus classical test theory. *Educational and Psychological Measurement*, *62*(6), 921- 943. https://doi.org/10.1177/0013164402238082

Mardia, K. V. (1970). Measures of multivariate skewness and kurtosis with applications. *Biometrika*, *57*(3), 519–530.

Mueller, R. O. (1996). *Basic principles of structural equation modeling: An introduction to LISREL and EQS*. *Design* (Vol. 102). New York: Springer Science & Business Media, LLC. https://doi.org/10.1016/j.peva.2007.06.006

Muthén, L. K., & Muthén, B. O. (2012). Mplus statistical modeling software: Release 7.0. *Los Angeles, CA: Muthén & Muthén*.

Partchev, I. (2016). irtoys: A collection of functions related to item response theory (IRT). Retrieved from https://cran.r-project.org/package=irtoys

Price, L. R. (2017). *Psychometric methods: Theory and practice*. New York, NY: The Guilford Press.

Progar, S., & Sočan, G. (2008). An empirical comparison of Item Response Theory and Classical Test Theory. *Horizons of Psychology*, *17*(3), 5–24.

R Core Team. (2017). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.r-project.org/.

Revelle, W. (2018). psych: Procedures for Psychological, Psychometric, and Personality Research. Evanston, Illinois. Retrieved from https://cran.r-project.org/package=psych

Robitzsch, A. (2017). sirt: Supplementary item response theory models. Retrieved from https://cran.r-project.org/package=sirt

Stage, C. (1998a). *A comparison between item analysis based on item response theory and classical test theory. A study of the SweSAT Subtest ERC*. *Educational Measurement*. Retrieved from http://www.sprak.umu.se/digitalAssets/59/59551_enr2998sec.pdf

Stage, C. (1998b). *A comparison between item analysis based on item response theory and classical test theory. A study of the SweSAT Subtest WORD*. *Educational Measurement*. Retrieved from http://www.sprak.umu.se/digitalAssets/59/59551_enr2998sec.pdf

Streiner, D. L. (1994). Figuring out factors: The use and misuse of factor analysis. *Canadian Journal of Psychiatry*, *39*(3), 135–140.

Tabachnik, B. G., & Fidell, L. S. (2012). *Using multivariate statistics* (6. ed.). Boston: Pearson.

Velicer, W. F. (1976). The relation between factor score estimates, image scores, and principal component scores. *Educational and Psychological Measurement*, *36*(1), 149–159. https://doi.org/10.1177/001316447603600114

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York. Retrieved from http://ggplot2.org

Williams, J. S. (1978). A definition for the common-factor analysis model and the elimination of problems of factor score indeterminacy. *Psychometrika*, *43*(3), 293–306. https://doi.org/10.1007/BF02293640

Xu, T., & Stone, C. A. (2012). Using IRT trait estimates versus summated scores in predicting outcomes. *Educational and Psychological Measurement*, *72*(3), 453–468. https://doi.org/10.1177/0013164411419846

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement*, *8*(2), 125–145. https://doi.org/10.1177/014662168400800201