ANADOLU UNIVERSITY JOURNAL OF EDUCATION FACULTY
ISSN: 2602-2249
Volume:4 Issue:3 2020 JULY

R Yazılımı ile Açımlayıcı Faktör Analizi

Exploratory Factor Analysis with R Software

Abdullah Faruk KILIÇ[1]

Article Type: Technical Note

Application Date: 02.04.2020 **Accepted Date:** 13.07.2020

To Cite This Article: Kılıç, A. F. (2020). Exploratory factor analysis with R software. *Anadolu University Journal of Education Faculty (AUJEF), 4*(3), 276-293.

ÖZ: Açımlayıcı Faktör Analizi (AFA) eğitim ve sosyal bilimler alanında sıklıkla kullanılmaktadır. AFA özellikle ölçek geliştirme ve uyarlama çalışmalarında kullanılmaktadır. Bu çalışmalarda AFA sıklıkla kullanıldığı için araştırmalacılar AFA'nın nasıl gerçekleştirildiğine ilişkin kılavuza ihtiyaç duyabilmektedir. Bu nedenle bu çalışmada AFA'nın R yazılımında nasıl gerçekleştirileceği açıklanmıştır. AFA farklı yazılımlarla da gerçekleştirilebilir. Fakat R yazılımı esnek ve ücretsizdir. Bu nedenle mevcut çalışma AFA'nın R yazılımında gerçekleştirilmesine odaklanmıştır. İlk olarak veri setinin AFA varsayımlarını sağlayıp sağlamadığı kontrol edilmiştir. Bunun için bir fonksiyon yazılmıştır. Daha sonra faktör sayısına karar vermek için Paralel Analiz (PA), en küçük kısmi ortalamalar (MAP) ve yamaç grafiği kullanılmıştır. Faktör sayısına karar verdikten sonra, açımlayıcı faktör analizi gerçekleştirilerek raporlanmıştır Sonuçların Word belgesi olarak raporlanması için gerekli R kodları sunulmuştur. Bu çalışmada beş kategorili (1-5) veri seti ile iki boyutlu yapı incelenmiştir. Faktör döndürme yöntemi olarak da eğik döndürme yöntmelerinden oblimin kullanılmıştır. Araştırmacılar R kodlarını kendi veri setlerinin özelliklerini göz önünde bulundurarak düzenlemelidir.

Anahtar sözcükler: Açımlayıcı faktör analizi, R yazılımı, paralel analiz, en küçük kısmi ortalamalar

ABSTRACT: Exploratory Factor Analysis (EFA) is frequently used in educational and social sciences. EFA has been used in scale development and adaptation studies, in particular. Therefore, in this study, how to conduct EFA in R software has been explained. First of all, it is examined whether the data set holds the assumptions of EFA. When examining the assumptions of EFA, a function was written. Then, the number of factors was evaluated via parallel analysis (PA), minimum average partial (MAP), and scree plot. After deciding on the number of factors, EFA was conducted and reported. To report the results, R codes were provided to write the results in a Word document. Five categories and two-factorial data set were used in the current study. Oblimin was used as rotation method. Researchers should edit the R codes in terms of their data set properties.

Keywords: Exploratory factor analysis, R software, parallel analysis, minimum average partial

[1] Research Assistant Dr., Adıyaman University, abdullahfarukkilic@gmail.com, ORCID: 0000-0003-3129-1763 (Corresponding author)

1. INTRODUCTION

Factor analysis is one of the most widely used methods in psychology, social sciences and educational sciences (Fabrigar et al., 1999; Floyd & Widaman, 1995; P. Kline, 1994). In particular, factor analytical methods are frequently used in scale development and adaptation studies (Acar-Güvendir & Özer-Özkan, 2015; Boztunç Öztürk et al., 2015; Gül & Sözbilir, 2015; Kılıç & Koyuncu, 2017). Goretzko et al. (2019) examined *Psychological Assessment* and *the European Journal of Psychological Assessment (EJPA)* journals from 2007 to 2017 and stated that over 336 studies had used some type of exploratory factor analysis (EFA) procedure. The fact that it is a frequently used method has led to the publication of articles emphasizing the importance of correct use of factor analysis (Costello & Osborne, 2005; Fabrigar et al., 1999; Kahn, 2006; Watkins, 2018). Thus, it helps researchers to use and report the results of EFA correctly.

The covariances between variables are examined in factor analysis and it aims to obtain fewer latent variables than the number of indicators (Brown, 2015; Fabrigar & Wegener, 2012; Floyd & Widaman, 1995; Kahn, 2006; Watkins, 2018). There are two modes of factor analysis: the first is exploratory and the second is confirmatory mode (Price, 2017). The purpose of the exploratory mode, known as exploratory factor analysis (EFA), is to define the model or factor structure of a set of variables. The confirmatory mode, known as confirmatory factor analysis (CFA), is usually based on a strong theoretical or experimental structure. While EFA constructs a theory, CFA tests a theory (Stevens, 2009). The current study focuses on EFA procedures.

1.1. Exploratory Factor Analysis

In this part of the study, the assumptions of EFA, sample size, factor extraction methods, factor retention methods, interpretation and replication of the results are given. Then R codes are given for exploratory factor analysis.

1.1.1. *Assumptions of EFA*

Before conducting EFA, assumptions must be held or precautions should be taken in terms of violations of assumptions. For this purpose, the data set must be examined in terms of missing values. If there are missing values in the data set, the necessary precautions should be taken. Enders (2010) can be examined with a view to overcoming the missing value problem. In addition, the relationships between variables should be linear. Moreover, variables are assumed to have a multivariate normal distribution (Alpar, 2013). In addition, data set does not consist of multivariate outliers (Tabachnik & Fidell, 2012).

1.1.2. *Sample size*

There are different approaches to minimum sample size. Floyd and Widaman (1995) stated that there should be four or five individuals per indicator, while Gorsuch (1974) said that there should be five individuals. However, he emphasized that the sample size should not be less than 200. Streiner (1994) suggested five individuals per indicator like Gorsuch (1974) but he indicated that the sample size should not be less than 100. If the sample size were less than 100, there would have to be 10 individuals per indicator. Comrey (1988) stated that a sample size for 200 was sufficient in most cases but stressed that this was the case if the items in the scale did not exceed 40.

Guadagnoli and Velicer (1988) reported that factor loadings of items greater than 0.80 are stable, even if the sample size is less than 50, regardless of the number of variables. Pearson and Mundform (2010) reported that when binary data are normally distributed, a sample size of 50 is sufficient when there are 12 variables per factor. De Winter, Dodou, and Wieringa (2009) stated that a sample size below 50 could be sufficient. They emphasized that for unidimensional constructs, if factor loadings are 0.8 and there are 24 indicators, a sample size of six could be sufficient.

Erkuş (2014) emphasized that when EFA is used in the scale development process, none of the sample size recommendations in the literature may be valid and the sample size may change according to the nature of the measured feature.

1.1.3. Selection of Factor Extraction Method

There are a number of factor extraction methods in EFA, such as principal components, maximum likelihood (ML), alfa factoring, unweighted least squares (ULS), weighted least squares (WLS), principal axis factoring (PAF) and minimum residual (minRes).

Although the methods have advantages and weaknesses when compared to each other, it has been stated that the PAF method generally gives better results for data that are not distributed normally (Fabrigar et al., 1999). Costello and Osborne (2005) stated that if data are nearly normally distributed, ML is recommended. But if the normality assumption does not hold, the PAF method is recommended for most cases. In the present study, PAF, which is called the "principal factor solution" in the psych package, was used as the factor extraction method.

1.1.4. Factor Retention Methods

There are a number of methods for deciding the factor number such as the Kasier K1 rule (Kaiser, 1960), scree plot (Cattell, 1966), minimum average partial (MAP) analysis (Velicer, 1976), parallel analysis (PA) (Horn, 1965) and Stout's (1987) DIMTEST. Research shows that PA and MAP analysis has more accurate results (Buja & Eyuboglu, 1992; Cho et al., 2009; Cota et al., 1993; Garrido et al., 2011; Yang & Xia, 2015; Zwick & Velicer, 1986). In this research, the PA, MAP, scree plot and DETECT methods were used.

1.1.5. Factor rotation

Factor rotation methods, which are used in order to increase the interpretability of the factors (Osborne, 2015; Watkins, 2018), can be divided into two categories, namely orthogonal and oblique (Osborne, 2014). Orthogonal rotation assumes that factors are not correlated while oblique methods assume that factors are correlated and allow correlation between factors. If the construct is unidimensional, no rotation method is used. While explained total variance is not changed as a result of rotation (Costello & Osborne, 2005), the variance explained by the factors does change. The psych package has a lot of rotation methods, including oblimin, varimax, quartimax, equamax, bifactor and promax (Revelle, 2018). In this study, oblimin was used as an oblique rotation method.

1.1.6. Interpretation and replication

One can try to interpret factors by examining the variables with and without relation to factors in the interpretation stage (Gorsuch, 1974). Factor loadings, cross-loadings and explained variance as a result of EFA should be examined and evaluated.

Replication of EFA has been widely argued for by Osborne (2014), Osborne and Banjanovic (2016) and Osborne and Fitzpatrick (2012). Briefly, to conduct replication analysis for EFA results, divide the sample into two random samples and conduct EFA with the same factor extraction method, factor number and rotation technique. Then compare the results in terms of factor loadings and examine which items load the same factor.

In addition to the situations that should be considered in the use of factor analysis, the software used in factor analysis is also important for researchers. Factor analysis can be performed via different software. However, software can differ in terms of allowed factor extraction, factor rotation or correlation methods (e.g. SPSS does not allow tetrachoric or polychoric correlation matrix for binary or polytomous data). On the other hand, Factor software (Lorenzo-Seva & Ferrando, 2019) does not allow the principal axis factoring extraction method. While SPSS is commercial, Factor is not. R software (R Core Team, 2018) was used in the current research because it provides flexibility to researchers and is free.

2. R CODE FOR EXPLORATORY FACTOR ANALYSIS

In the current study, EFA was conducted on simulated data set. The data set has 20 indicators (observed variable). All of the indicators have 5 categories. To conduct EFA in R, we use some packages. In R software, "*Package*" means a collection of functions. Instead of writing functions one by one, we use instant functions in packages. Because of this, we install or call packages first of all. Figure 1 contains the codes used to import the data set into R software and to examine whether the assumptions of the analysis hold. In order to make the codes more understandable, comment lines have been added.

```
#if you have not previously installed these packages, please install these.
#install.packages(c("pastecs", "moments", "mctest", "dplyr", "psych", "polycor", "corrplot",
"nFactors", "ggplot2", "sirt", "data.table), "flextable")
#loading required packages
library(dplyr) # to use chain operator activate this package

#read data
data <- read.table("D:/efa_with_r/r_application.txt", header = FALSE)
#use your location and if your data has header please change the code as header = TRUE
#The data is .txt format. There were not row names and column names. It contains only individual
#responses like 1,2,3,4 or 5

assumptions <- function(x) { #x is a data frame includes item responses
#creating a summary data frame
#descriptive statistics were obtain via pastects package
  descr <- as.data.frame(matrix(NA, nrow = 8, ncol = ncol(x)))
  rownames(descr) <- c("Number_of_Observations",
                       "Number_of_missing_values",
                       "min_value", "max_value",
                       "mode_value", "median_value",
                       "_skewness_", "_kurtosis_")
  descriptives <- pastecs::stat.desc(x) #calculate descriptive statistics.

#This function taken from https://www.r-bloggers.com/computing-the-mode-in-r/
 Mode = function(x) {
    ta = table(x)
    tam = max(ta)
    if (all(ta == tam))
      mod = NA
```

```
    else
      if(is.numeric(x))
        mod = as.numeric(names(ta)[ta == tam])
    else
      mod = names(ta)[ta == tam]
    return(mod)
  }

  #calculate modes
  mods <- as.data.frame(apply(as.matrix(x), 2, Mode))
  descr[1:4, ] <- descriptives[c(1, 3, 4, 5), ]
  descr[5, ] <- mods[1:ncol(x), ]
  descr[6, ] <- descriptives[8, ]
  descr[7, ] <- moments::skewness(x, na.rm = T)
  descr[8, ] <- moments::kurtosis(x, na.rm = T)-3




#Calculate VIF and Tolerance values
  #To obtain IF and TV values describe the model
  x_new <- x
  x_new$rn <- 1:nrow(x)
  model_for_collinearity <-  lm(
    as.formula(paste(colnames(x_new)[ncol(x_new)], "~",
                     paste(colnames(x_new)[1:(ncol(x_new)-1)], collapse = "+"),
                     sep = ""
    )), data = x_new)
  mc_VIF_TOL <- as.data.frame(mctest::mctest(model_for_collinearity,
                 type = "i")$idiags[,1:2]) #calculate VIF and Tollerance values

#Calculate Condition Index
  mc_CI <- mctest::eigprop(mod = model_for_collinearity)$ci

#A data frame for summary of multicollinearity
  mc_control <- data.frame(min_VIF = min(mc_VIF_TOL$VIF),
                           max_VIF = max(mc_VIF_TOL$VIF),
                           min_TOL = min(mc_VIF_TOL$TOL),
                           max_TOL = max(mc_VIF_TOL$TOL),
                           min_CI = min(mc_CI),
                           max_CI = max(mc_CI) #giving a summary of multicollinearity

  )
  )

  #Mahalanobis Distance Calculation
  #To calculate mahalanobis distance, missing values are not accepted.
  distance <- as.matrix(mahalanobis(x, colMeans(x), cov = cov(x)))

#Those with Mahalanobis Distance p values bigger than 0.001 were considered as outliers.
  Mah_significant <- x %>%
      transmute(row_number = 1:nrow(x),
                Mahalanobis_distance = distance,
                Mah_p_value = pchisq(distance, df = ncol(x), lower.tail = F)) %>%
      filter(Mah_p_value <= 0.001)

  #Calculate Mardia's kurtosis value for multivariate normality
    mardia_kurt <- psych::mardia(x, plot = F)

  #Return a list consist of descriptive statistics, multicollinearity, multivariate normality
and Mahalanobis distance for multivariate outliers
    return(list(descriptives = round(descr, 2),
                multicollineartiy =  round(mc_control, 2),
                Mah_significant =  Mah_significant,
                n_outlier = nrow(Mah_significant),
                Mardia_Kurtosis = mardia_kurt$kurtosis,
                Mardia_Kurtosis_p_value = mardia_kurt$p.kurt ))
}
```

Figure 1: *Codes for Checking Whether the Data Set Holds the EFA Assumptions*

The function in Figure 1 was developed to examine the EFA assumptions. When the codes are examined, it can be seen that "x" refers to the data set in the function named "assumptions." In the function, firstly, descriptive statistics were examined. For this purpose, the number of observations, the number of missing values, and minimum, maximum, mode, median, skewness and kurtosis values for variables were added to the function. To obtain descriptive statistics, stat.desc() function in pastecs package was used (Grosjean & Ibanez, 2018). Then, to determine whether there were multicollinearity problems, variance inflation factor (VIF), tolerance value (TV) and conditional index (CI) values were calculated. The maximum and minimum values of VIF, TV and CI values were kept in a data frame. The Mahalanobis distance values were calculated to examine the multivariate outliers. To examine multicollinearity, mctest package was used (Ullah et al., 2019). The chi-square test was used to determine whether the multivariate outliers were significant. Mahalanobis distance values that were significant at the $\alpha = 0.001$ level were reported. To examine multivariate outliers, stats package was used. The assumptions function returns a list consisting of six different data frames. The first of these was descriptive statistics. Then maximum and minimum values of VIF, TV and CI values were returned, respectively. In the third order were Mahalanobis distance and their p-values that were statistically significant. In the fourth order, there was the number of outliers. In the fifth and six order, Mardia's (1970) multivariate kurtosis value and its p-value were found. To examine multivariate normality, psych package (Revelle, 2018) was used. In Figure 2, the output of the assumptions function is given.

```
#After creating function, examine whether data holds EFA assumptions.
                            V1      V2      V3      V4      V5      V6      V7
Number_of_Observations   2000.00 2000.00 2000.00 2000.00 2000.00 2000.00 2000.00
Number_of_missing_values    0.00    0.00    0.00    0.00    0.00    0.00    0.00
min_value                   0.00    0.00    0.00    0.00    0.00    0.00    0.00
max_value                   4.00    4.00    4.00    4.00    4.00    4.00    4.00
mode_value                  2.00    2.00    2.00    2.00    2.00    2.00    2.00
median_value                2.00    2.00    2.00    2.00    2.00    2.00    2.00
_skewness_                  0.01    0.01    0.01   -0.06   -0.01   -0.04   -0.02
_kurtosis_                 -0.47   -0.46   -0.42   -0.45   -0.50   -0.46   -0.49


                            V8      V9     V10     V11     V12     V13     V14
Number_of_Observations   2000.00 2000.00 2000.00 2000.00 2000.00 2000.00 2000.00
Number_of_missing_values    0.00    0.00    0.00    0.00    0.00    0.00    0.00
min_value                   0.00    0.00    0.00    0.00    0.00    0.00    0.00
max_value                   4.00    4.00    4.00    4.00    4.00    4.00    4.00
mode_value                  2.00    2.00    2.00    2.00    2.00    2.00    2.00
median_value                2.00    2.00    2.00    2.00    2.00    2.00    2.00
_skewness_                  0.04    0.01    0.06    0.04   -0.06    0.00    0.01
_kurtosis_                 -0.53   -0.53   -0.54   -0.51   -0.46   -0.47   -0.52

                           V15     V16     V17     V18     V19     V20
Number_of_Observations   2000.00 2000.00 2000.00 2000.00 2000.00 2000.00
Number_of_missing_values    0.00    0.00    0.00    0.00    0.00    0.00
min_value                   0.00    0.00    0.00    0.00    0.00    0.00
max_value                   4.00    4.00    4.00    4.00    4.00    4.00
mode_value                  2.00    2.00    2.00    2.00    2.00    2.00
median_value                2.00    2.00    2.00    2.00    2.00    2.00
_skewness_                  0.03    0.00   -0.02    0.00    0.02   -0.01
_kurtosis_                 -0.43   -0.49   -0.52   -0.44   -0.45   -0.45


$multicollineartiy
  min_VIF max_VIF min_TOL max_TOL min_CI max_CI
1    1.57    1.79    0.56    0.64      1  15.04

$Mah_significant
  row_number Mahalanobis_distance  Mah_p_value
1       1514             47.20826 0.0005487822

$n_outlier
```

```
[1] 1

$Mardia_Kurtosis
[1] -5.761804

$Mardia_Kurtosis_p_value
[1] 8.321945e-09
```

Figure 2: *Output of Assumptions Function*

When Figure 2 is examined, firstly descriptive statistics can be seen. There were 20 variables in the data set and 2000 observations. It can be seen that there were no missing values. If you have a missing value in your data set, it is recommended that you cope with the missing value using the appropriate methods (see Enders, 2010). As the focus of this study was on the EFA, there were no details about missing values. The maximum, minimum, mode and median values of the variables can also be found here. In addition, univariate skewness and kurtosis values of variables are included in the descriptive statistics. Thus, at first glance, the data can be understood as to whether they hold assumptions of EFA.

There was a multicollinearity output after the descriptive statistics. If there is no multicollinearity problem, tollerance values should be greater than 0.01, VIF values should be smaller than 10 and CI values should be smaller than 30 (R. B. Kline, 2011; Tabachnik & Fidell, 2012). Therefore, the maximum and minimum values of the TV, VIF and CI were reported. After the multicollinearity output, Mahalanobis distances that are significant at the $\alpha = 0.001$ level, their row numbers and p-values are reported. A number of multivariate outliers were located after Mahalanobis distances. The data set used in the study had 1 outlier. The last output of the function was Mardia's (1970) kurtosis and its p-value. The code in Figure 3 was used to subtract the outliers form the data set.

```
new_data <- data[-control_assumptions$Mah_significant$row_number, ]
```

Figure 3: *The Code for Subtraction of Outliers from the Data Set*

After subtraction of outliers from the data set, the “new_data” variable was used for EFA. Correlations between the variables in the data set were examined to conduct EFA after checking the assumption of EFA and making the necessary regularization. Figure 4 contains codes for examining the correlations between variables.

```
#____________Examination of Correlation between Variables_____________
cor_of_variables <- psych::polychoric(new_data)
only_number <- corrplot::corrplot(cor_of_variables, method = "number")
#Correlations between variables
number_and_circle <- corrplot::corrplot.mixed(cor_of_variables,
                        lower.col = "black",
                        number.cex = .7,
                        order = "FPC") #Ordered Correlations
```

Figure 4: *Investigation of Correlations Between Variables*

When the outputs of the assumptions function were examined, it was seen that the maximum and minimum values of the variables are 4 and 0, respectively. Therefore, the polychoric correlation matrix, which is suitable for polythomous data which has 5 or lower categories (Finney & DiStefano, 2013), was used. There was a visualization of correlations in Figure 4 after calculation of correlations between variables. The outputs of these codes are shown in Figure 5 and Figure 6, respectively.

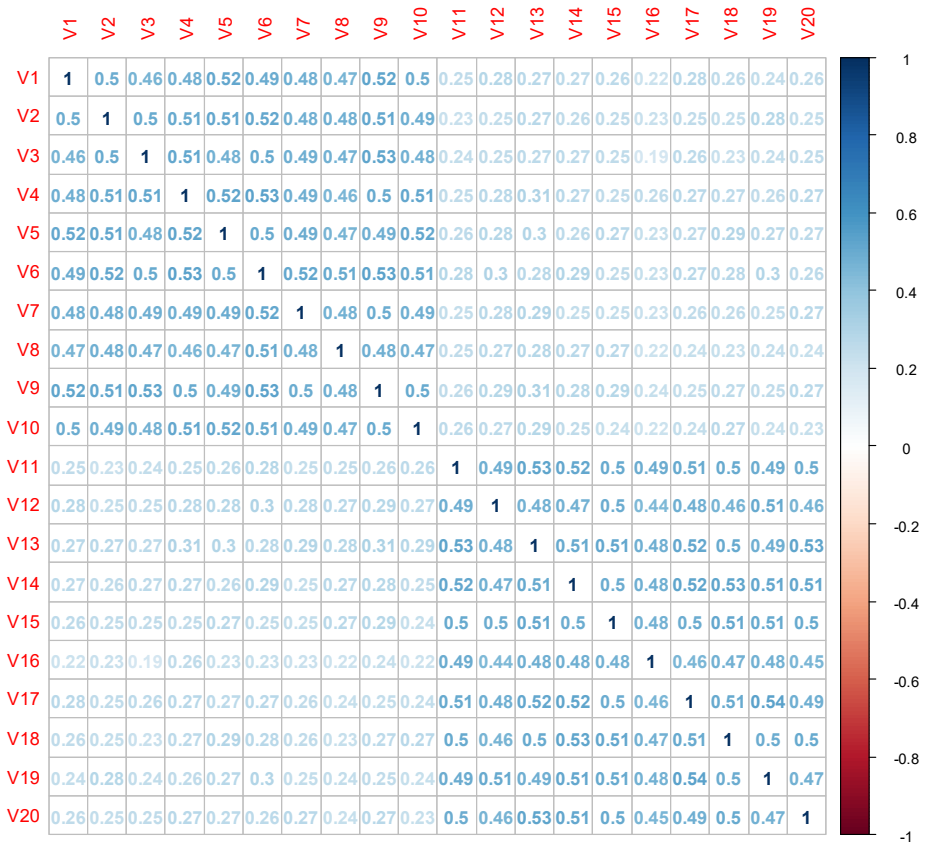

Figure 5: *Output of "only_number" Variable*

Figure 5 shows the numerical values of the correlations between variables. The correlations are positive when the color is closer to red, while the correlations are negative when the color is closer to blue. The shade of the colors gives an idea about the strength of the correlations.

In Figure 6, the correlations are sorted from large to small and the correlations are visualized with the help of circles in the upper triangle. Thus, when the data set has more variables, the correlations between the variables can be quickly examined.

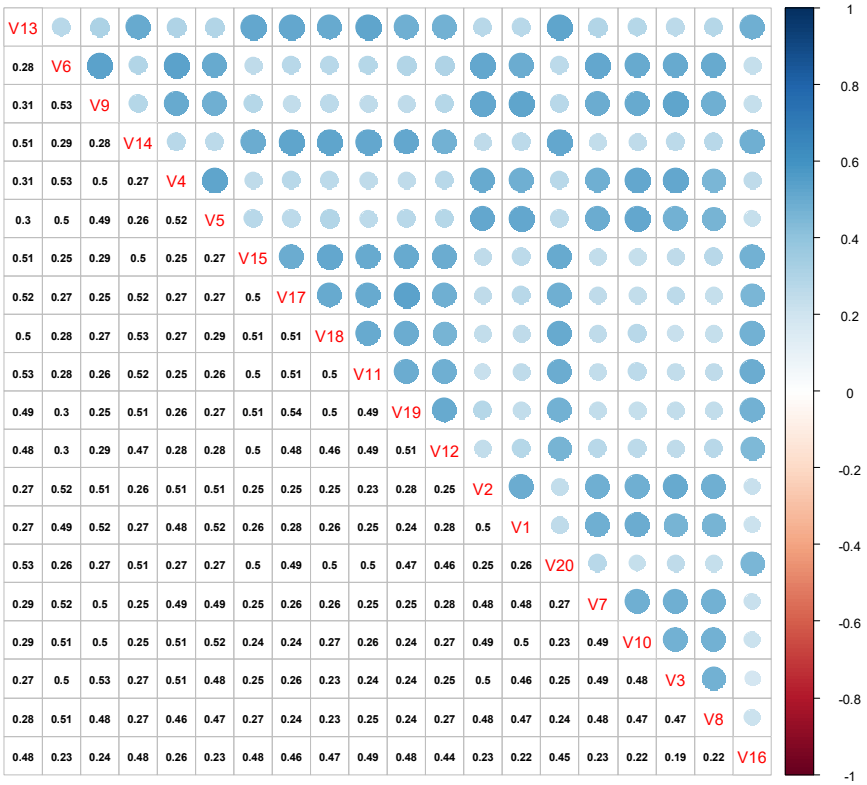

Figure 6: *Output of "number_and_circle" Variable*

When Figure 6 is examined, it can be seen that the correlations are sorted from large to small. Thus, the correlations vary between 0.62 and 0.25. After examining the correlations between the variables, the KMO value of the sample size for EFA was evaluated. Then, Bartlett's test, which gives information about whether the correlation matrix differs from the unit matrix, was conducted. The KMO value and the R codes for the Bartlett test are given in Figure 7.

```
#____________________KMO Value and Bartlett Test for EFA_____________________

kaiser <- psych::KMO(cor_of_variables)
bart <- psych::cortest.bartlett(cor_of_variables, n = nrow(new_data), diag = TRUE)

#####################################################
#Interpratation of KMO Value  (Kaiser & Rice, 1974)#
#KMO > 0.90 ==> Marvelous,                          #
#0.80 < KMO < 0.90 ==> Meritorious,                 #
#0.70 < KMO < 0.80 ==> Middling,                    #
#0.60 < KMO < 0.70 ==> Mediocre,                    #
#0.50 < KMO < 0.60 ==> Miserable,                   #
#KMO < 0.50 ==> Unacceptable.                       #
#####################################################
interpretation_KMO <- dplyr::case_when(
  kaiser$MSA >= 0.90 ~ "Marvelous",
  kaiser$MSA >= 0.80 & kaiser$MSA < 0.90 ~ "Mertitourious",
  kaiser$MSA >= 0.70 & kaiser$MSA < 0.80 ~ "Middling",
  kaiser$MSA >= 0.60 & kaiser$MSA < 0.70 ~ "Medicore",
  kaiser$MSA >= 0.50 & kaiser$MSA < 0.60 ~ "Miserable",
  kaiser$MSA <0.50 ~ "Unacceptable"
)

Bart_KMO <- data.frame(KMO = round(kaiser$MSA, 3),
                       Interpretation_KMO = interpretation_KMO,
                       Bartlett_Chi = bart$chisq,
                       Bartlett_df = bart$df,
                       Bartlett_sig = sprintf("%.3f",bart$p.value))
```

Figure 7: *Examination of KMO Value and Bartlett's Test Results*

As can be seen in Figure 7, correlations between variables were used to calculate the KMO value and conduct the Bartlett test. The KMO value was compared with the criterion value of KMO given by Kaiser and Rice (1974) and then the result was saved in a data frame named "Bart_KMO." The results of the Bartlett test and calculation of the KMO value are presented in Figure 8. To obtain KMO value and conduct Bartlett test pscyh package was used (Revelle, 2018).

```
> Bart_KMO
    KMO Interpretation_KMO Bartlett_Chi Bartlett_df Bartlett_sig
1 0.967          Marvelous     18591.93         190        0.000
```

Figure 8: *KMO Value and Bartlett's Test Results*

The KMO value, its interpretation, Bartlett's test chi-square value, its degree of freedom and p-value are reported in Figure 8. Scree plot, parallel analysis (PA) and minimum average partial (MAP) methods were used to determine the number of factors. The codes used for this analysis are presented in Figure 9.

```
#_____________________________Factor Retention in EFA_____________________________
# There is some method for factor retention. nFactors package has many methods.

#Scree Plot for determine number of factors
eigenvalues <- nFactors::eigenComputes(x = cor_of_variables)
eigen_for_graph <- data.frame(item_number = 1:ncol(new_data), eigenvalues)
library(ggplot2)
scree_plot <- ggplot(data = eigen_for_graph) +
  geom_point(aes(x = item_number, y = eigenvalues )) +
  geom_line(aes(x = item_number, y = eigenvalues )) +
  xlab("Factor Number")+
  ylab ("Eigenvalues")+
  theme_classic()+
  scale_x_continuous(breaks = seq(from = 1, to = ncol(new_data), by = 1))

#MAP analysis for examine number of dimensions
map_analysis <- psych::vss(new_data, n = (ncol(new_data) - 1))
map_factors <- which(map_analysis$map == min(map_analysis$map))

#Parallel analysis for examine number of dimensions
#Conduct Parallel analysis with Pearson Correlation Matrix
PA_pearson <- psych::fa.parallel(new_data, fa = "both", cor = "cor")

#Conduct Parallel Analysis with Polychoric Correlation Matrix
PA_poly <- psych::fa.parallel(new_data, fa = "both", cor = "poly")

results_factor_retentation <- list(MAP_Result = map_factors,
                                   Parallel_Analysis_Pearson = PA_pearson$nfact,
                                   Parallel_Analysis_Polychoric = PA_poly$nfact,
                                   Scree_Plot = "Look at the Plots Section for Scree Plot",
                                   scree_plot)
```

Figure 9: *Examining Factor Retention Methods' Results*

It can be seen that the scree plot was created first in Figure 9 via ggplot2 package (Wickham, 2016). To create the scree plot, first, the eigenvalues were calculated and then the eigenvalues are shown in a graphic. The psych package (Revelle, 2018) was used to conduct PA and MAP analysis. Both the Pearson product moment (PPM) correlation matrix and the polychoric correlation matrix were used to conduct PA. Finally, the results of all factor retention methods are reported in a list named "results_factor_retention." Figure 10 shows the value of the "results_factor_retention" variable and Figure 11 shows the scree plot of eigenvalues.

```
> results_factor_retentation
$MAP_Result
[1] 2

$Parallel_Analysis_Pearson
[1] 2

$Parallel_Analysis_Polychoric
[1] 2

$Scree_Plot
[1] "Look At The Plots Section for Scree Plot"

[[5]]
```

Figure 10: *Results Results of Factor Retention Methods*

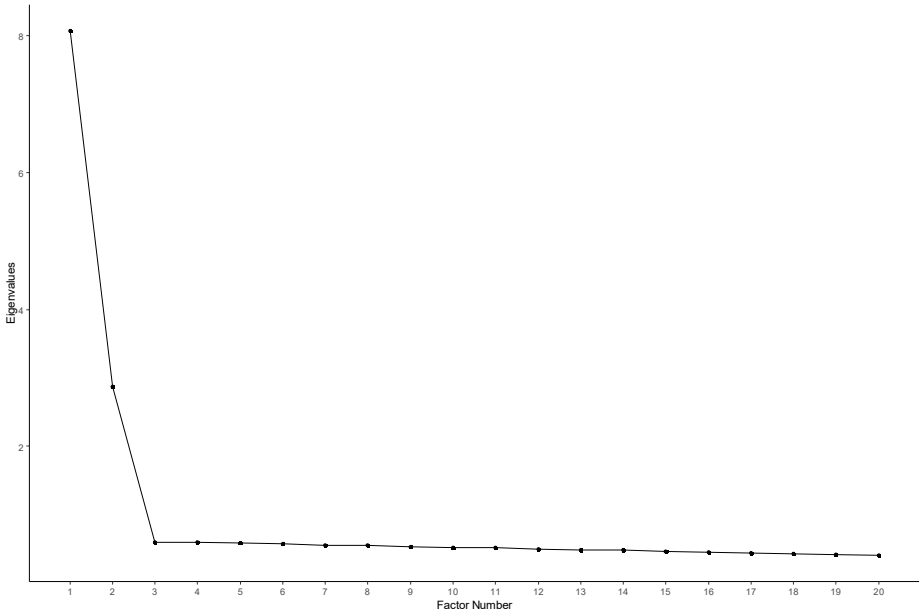

Figure 11: *Scree Pilot*

Figure 10 shows the "results_factor_retention" output. In addition, Figure 11 shows the scree plot. When the values of this variable are examined, it can be seen that all of the methods suggest two factors construct. Both PA with PPM correlation and PA with polychoric correlation suggests a two-factorial construct. When the scree plot is examined, it can be said that the construct is two-factorial. Thus, four different factor retention method results are reported. In the present study, due to all factor retentation methods suggests two factorial constructs, analyses were continued with two factorial construct. R codes created for EFA are presented in Figure 12.

```
#_____________________________Factor Analysis Results_____________________________
#1) Polychoric correlation matrix was used. Because of polythomous data.
#2) Principle Axis Factoring method was used for factor extraction method. Because multivariate
normality didn't hold.
#3) Number of factors was fixed to 2 because of PA, MAP and scree plot's results.

efa_results_two_factors <- psych::fa(r = new_data, cor = "poly", nfactors = 2, fm = "pa",
rotate = "oblimin")

#examine the results in terms of factor loading, interpretation and teorical framework
efa_results_two_factors
```

```
#Reporting EFA Results
#Create a summary data frame for Reporting
EFA_summary <- data.frame(item_number = 1:ncol(new_data),
                          std_dev = as.vector(apply(new_data, 2, sd)),
                          skewness = as.numeric(control_assumptions$descriptives[7, ]),
                          kurtosis = as.numeric(control_assumptions$descriptives[8, ]),
                          factor_loading = efa_results_two_factors$loadings[1:ncol(new_data), ],
                          communalities = efa_results_two_factors$communality )
EFA_summary
```

Figure 12: *EFA Codes for R*

R codes for EFA are given in Figure 12. First, the number of dimensions was determined considering factor retention methods. However, this was not a final decision. Therefore, possible structures should be tested. In this study, i picked up two factorial solution. The two-factorial solution was found by using principal axis factoring as the factor extraction method with the polychoric correlation matrix. I use oblimin rotation beacuse of most constructs related each other in social science. If orthogonal rotation is desired, “rotate = "varimax"” argument can be used. EFA results can be examined in terms of factor loadings, interfactor correlation, explained variance. First EFA results examined than a summary table was created via the analysis results. Figure 13 shows the “efa_results_two_factors” and “EFA_summary” variable output.

```
> efa_results_two_factors
Factor Analysis using method =  pa
Call: psych::fa(r = new_data, nfactors = 2, rotate = "oblimin",
    fm = "pa", cor = "poly")
Standardized loadings (pattern matrix) based upon correlation matrix
      PA1   PA2   h2   u2 com
V1   0.69  0.00 0.48 0.52   1
V2   0.72 -0.02 0.50 0.50   1
V3   0.71 -0.02 0.49 0.51   1
V4   0.71  0.01 0.51 0.49   1
V5   0.70  0.02 0.50 0.50   1
V6   0.72  0.01 0.53 0.47   1
V7   0.69  0.00 0.48 0.52   1
V8   0.67  0.01 0.45 0.55   1
V9   0.72  0.01 0.52 0.48   1
V10  0.71 -0.02 0.50 0.50   1
V11 -0.02  0.73 0.51 0.49   1
V12  0.05  0.65 0.46 0.54   1
V13  0.04  0.70 0.52 0.48   1
V14  0.00  0.72 0.52 0.48   1
V15 -0.01  0.72 0.51 0.49   1
V16 -0.04  0.68 0.44 0.56   1
V17 -0.01  0.72 0.51 0.49   1
V18  0.00  0.71 0.50 0.50   1
V19 -0.01  0.71 0.50 0.50   1
V20  0.00  0.70 0.49 0.51   1

                       PA1  PA2
SS loadings           4.98 4.96
Proportion Var        0.25 0.25
Cumulative Var        0.25 0.50
Proportion Explained  0.50 0.50
Cumulative Proportion 0.50 1.00

 With factor correlations of
     PA1  PA2
PA1 1.00 0.52
PA2 0.52 1.00

Mean item complexity =  1
Test of the hypothesis that 2 factors are sufficient.

The degrees of freedom for the null model are  190  and the objective function was  9.34
with Chi Square of  18591.93
```

```
The degrees of freedom for the model are 151  and the objective function was  0.11

The root mean square of the residuals (RMSR) is  0.01
The df corrected root mean square of the residuals is  0.01

The harmonic number of observations is  1999 with the empirical chi square  111  with prob <
0.99
The total number of observations was  1999  with Likelihood Chi Square =  222.09  with prob
<  0.00015

Tucker Lewis Index of factoring reliability =  0.995
RMSEA index =  0.015  and the 90 % confidence intervals are  0.011 0.02
BIC =  -925.57
Fit based upon off diagonal values = 1
Measures of factor score adequacy
                                                PA1  PA2
Correlation of (regression) scores with factors   0.95 0.95
Multiple R square of scores with factors          0.91 0.91
Minimum correlation of possible factor scores     0.82 0.82
```

```
> EFA_summary
     item_number std_dev skewness kurtosis factor_loading.PA1 factor_loading.PA2 communalities
V1             1   1.016     0.01    -0.47              0.692              0.005          0.483
V2             2   1.004     0.01    -0.46              0.720             -0.020          0.504
V3             3   1.009     0.01    -0.42              0.709             -0.023          0.486
V4             4   1.006    -0.06    -0.45              0.707              0.012          0.509
V5             5   1.034    -0.01    -0.50              0.700              0.018          0.503
V6             6   1.021    -0.04    -0.46              0.723              0.013          0.532
V7             7   1.031    -0.02    -0.49              0.692              0.004          0.482
V8             8   1.018     0.04    -0.53              0.666              0.010          0.451
V9             9   1.019     0.01    -0.53              0.718              0.008          0.521
V10           10   1.043     0.06    -0.54              0.714             -0.017          0.497
V11           11   1.013     0.04    -0.51             -0.020              0.727          0.513
V12           12   1.009    -0.06    -0.46              0.051              0.650          0.459
V13           13   0.992     0.00    -0.47              0.043              0.697          0.519
V14           14   1.011     0.01    -0.52              0.005              0.718          0.519
V15           15   1.014     0.03    -0.43             -0.006              0.716          0.509
V16           16   1.012     0.00    -0.49             -0.038              0.685          0.444
V17           17   1.018    -0.02    -0.52             -0.007              0.717          0.508
V18           18   1.000     0.00    -0.44             -0.004              0.712          0.504
V19           19   1.017     0.02    -0.45             -0.007              0.713          0.503
V20           20   1.010    -0.01    -0.45              0.002              0.696          0.486
```

Figure 13: *Summary for Reporting EFA Results*

Under the title “Standardized loadings (pattern matrix) based upon correlation matrix”, there are factor loadings (PA1 and PA2), communalites (h2) and uniquness values. In this example, first 10 item (1-10) loaded first factor and second 10 item (11-20) loaded second factor. The factor loadings of the first factor ranged from 0.69 to 0.72, while the factor loadings of the second factor ranged from 0.44-0.52. The analysis output includes the explained variance ratio after this section. First factor explained 25% of the total variance. Second factor explained 25% of the total variance, too. Cumulative varince row indicates that two-factorial construct explain 50% of the total variance. The analysis output includes the interfactor correlation after the explained variance ratio section. In this example interfactor correlation matrix was 2x2 because of two-factorial solution. Interfactor correlation was 0.52 for this solution. There was some analysis details which was not used frequenlty for EFA after the interfactor correlation section.

“EFA_summary” which was created as a summary of the analysis in Figure 13 have standard deviation (std_dev) skewness, kurtosis, factor loadings of first factor (factor_loading.PA1) factor loadings of second factor (factor_loading.PA2) and communalities variables. The factor loadings obtained by EFA and the descriptive statistics of the items are given in Figure 13 via “EFA_summary” variable. The codes that write the “EFA_summary” variable output into the Word document to provide convenience for researchers are presented in Figure 14. A screenshot of the Word documents is given in Figure 15.

```
library(data.table)
library(flextable)
library(dplyr)

#writing results in Word document
#efa summary
EFA_summary %>%
  round(.,3) %>%
  regulartable() %>%
  set_formatter_type(fmt_double="%.02f") %>%
  align(align = "center") %>%
print(preview = "docx")

#explained variance ratio
  data.frame( type = rownames(efa_results_two_factors$Vaccounted),
efa_results_two_factors$Vaccounted) %>%
  regulartable() %>%
  print(preview = "docx")

#interfactor correlations
data.frame( type = rownames(efa_results_two_factors$Phi), efa_results_two_factors$Phi) %>%
    regulartable() %>%
    print(preview = "docx")
```

Figure 14: *Writing "EFA_summary" Output to Word Document*

item_number	std_dev	skewness	kurtosis	factor_ loading.PA1	factor_ loading.PA2	communalities
1.00	1.02	0.01	-0.47	0.69	0.00	0.48
2.00	1.00	0.01	-0.46	0.72	-0.02	0.50
3.00	1.01	0.01	-0.42	0.71	-0.02	0.49
4.00	1.01	-0.06	-0.45	0.71	0.01	0.51
5.00	1.03	-0.01	-0.50	0.70	0.02	0.50
6.00	1.02	-0.04	-0.46	0.72	0.01	0.53
7.00	1.03	-0.02	-0.49	0.69	0.00	0.48
8.00	1.02	0.04	-0.53	0.67	0.01	0.45
9.00	1.02	0.01	-0.53	0.72	0.01	0.52
10.00	1.04	0.06	-0.54	0.71	-0.02	0.50
11.00	1.01	0.04	-0.51	-0.02	0.73	0.51
12.00	1.01	-0.06	-0.46	0.05	0.65	0.46
13.00	0.99	0.00	-0.47	0.04	0.70	0.52
14.00	1.01	0.01	-0.52	0.00	0.72	0.52
15.00	1.01	0.03	-0.43	-0.01	0.72	0.51
16.00	1.01	-0.00	-0.49	-0.04	0.68	0.44
17.00	1.02	-0.02	-0.52	-0.01	0.72	0.51
18.00	1.00	0.00	-0.44	-0.00	0.71	0.50
19.00	1.02	0.02	-0.45	-0.01	0.71	0.50
20.00	1.01	-0.01	-0.45	0.00	0.70	0.49

type	PA1	PA2
SS loadings	4.9760369	4.9569952
Proportion Var	0.2488018	0.2478498
Cumulative Var	0.2488018	0.4966516
Proportion Explained	0.5009585	0.4990415
Cumulative Proportion	0.5009585	1.0000000

type	PA1	PA2
PA1	1.0000000	0.5220286
PA2	0.5220286	1.0000000

Figure 15: *Summary Results of EFA Output*

Descriptive statistics, factor loadings, communalities, total explained variance and interfactor correlations obtained from EFA were written in the Word documents. A screenshot of the Word documents are given in Figure 15 to show what it looked like.

3. RESULT

In this research, how EFA, which is frequently used in social sciences and educational sciences, can be conducted in R software has been exemplified. The current study was limited to the construct of the data set and the data set used. In the current study, two-factorial construct and five categories data were used. The rotation method was oblimin in this example. In addition, EFA was performed with the polychoric correlation matrix because of the data set being five categories. It would be appropriate for researchers to choose the appropriate correlation matrix for their own data sets.

REFERENCES

Acar-Güvendir, M., & Özer-Özkan, Y. (2015). Türkiye'deki eğitim alanında yayımlanan bilimsel dergilerde ölçek geliştirme ve uyarlama konulu makalelerin incelenmesi. *Elektronik Sosyal Bilimler Dergisi*, *14*(52), 23–33. https://doi.org/10.17755/esosder.54872

Alpar, R. (2013). *Uygulamalı çok değişkenli istatistiksel yöntemler* (4. Baskı). Detay Yayıncılık.

Boztunç Öztürk, N., Eroğlu, M. G., & Kelecioğlu, H. (2015). Eğitim alanında yapılan ölçek uyarlama makalelerinin incelenmesi. *Eğitim ve Bilim*, *40*(178), 123–137. https://doi.org/10.15390/EB.2015.4091

Brown, T. A. (2015). *Confirmatory factor analysis for applied research* (2nd ed.). The Guilford.

Buja, A., & Eyuboglu, N. (1992). Remarks on parallel analysis. *Multivariate Behavioral Research*, *27*(4), 509–540. https://doi.org/10.1207/s15327906mbr2704_2

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research*, *1*(2), 245–276. https://doi.org/10.1207/s15327906mbr0102_10

Cho, S.-J., Li, F., & Bandalos, D. L. (2009). Accuracy of the parallel analysis procedure with polychoric correlations. *Educational and Psychological Measurement*, *69*(5), 748–759. https://doi.org/10.1177/0013164409332229

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

Costello, A. B., & Osborne, J. W. (2005). Best practices in exploratory 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

Cota, A. A., Longman, R. S., Holden, R. R., Fekken, G. C., & Xinaris, S. (1993). Interpolating 95th percentile eigenvalues from random data: An empirical example. *Educational and Psychological Measurement*, *53*(3), 585–596. https://doi.org/10.1177/0013164493053003001

de Winter, J. C. F., Dodou, D., & Wieringa, P. A. (2009). Exploratory factor analysis with small sample sizes. *Multivariate Behavioral Research*, *44*(2), 147–181. https://doi.org/10.1080/00273170902794206

Enders, C. K. (2010). *Applied missing data analysis*. Guilford.

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

Fabrigar, L. R., & Wegener, D. T. (2012). *Exploratory factor analysis*. Oxford University.

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

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). 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. https://doi.org/10.1037/1040-3590.7.3.286

Garrido, L. E., Abad, F. J., & Ponsoda, V. (2011). Performance of Velicer's minimum average partial factor retention method with categorical variables. *Educational and Psychological Measurement*, *71*(3), 551–570. https://doi.org/10.1177/0013164410389489

Goretzko, D., Pham, T. T. H., & Bühner, M. (2019). Exploratory factor analysis: Current use, methodological developments and recommendations for good practice. *Current Psychology*, 1–12. https://doi.org/10.1007/s12144-019-00300-2

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

Grosjean, P., & Ibanez, F. (2018). *pastecs: Package for Analysis of Space-Time Ecological Series*.

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

Gül, Ş., & Sözbilir, M. (2015). Fen ve matematik eğitimi alanında gerçekleştirilen ölçek geliştirme araştırmalarına yönelik tematik içerik analizi. *Eğitim ve Bilim*, *40*(178), 85–102. https://doi.org/10.15390/EB.2015.4070

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

Kahn, J. H. (2006). Factor analysis in counseling psychology research, training, practice: Principles, advances, and applications. *The Counseling Psychologist*, *34*(5), 684–718. https://doi.org/10.1177/0011000006286347

Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and Psychological Measurement*, *20*(1), 141–151. https://doi.org/10.1177/001316446002000116

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

Kılıç, A. F., & Koyuncu, İ. (2017). Ölçek uyarlama çalışmalarının yapı geçerliği açısından incelenmesi. In Ö. Demirel & S. Dinçer (Eds.), *Küreselleşen dünyada eğitim* (pp. 1202–1205). Pegem Akademi.

Kline, P. (1994). *An easy guide to factor analysis*. Routledge. https://doi.org/10.1016/0191-8869(94)90040-X

Kline, R. B. (2011). *Principles and practise of structural equating modeling* (3. Baskı). The Guilford Press.

Lorenzo-Seva, U., & Ferrando, P. J. (2019). *Factor (Version 10.10.01) [Computer software]* (10.8.04). Universitat Rovira i Virgili.

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

Osborne, J. W. (2014). Best practices in exploratory factor analysis. In *Best Practices in Quantitative Methods* (Issue August, pp. 86–99). CreateSpace Independent Publishing. https://doi.org/10.4135/9781412995627.d8

Osborne, J. W. (2015). What is rotating in exploratory factor analysis? *Practical Assessment Research & Evaluation*, *20*(2), 1–7.

Osborne, J. W., & Banjanovic, E. S. (2016). *Exploratory factor analysis with SAS®*. SAS Intitute Inc.

Osborne, J. W., & Fitzpatrick, D. C. (2012). Replication analysis in exploratory factor analysis: What it is and why it makes your analysis better. *Practical Assessment, Research & Evaluation*, *17*(15). http://pareonline.net/getvn.asp?v=17&n=15

Pearson, R. H., & Mundform, D. J. (2010). Recommended sample size for conducting exploratory factor analysis on dichotomous data. *Journal of Modern Applied Statistical Methods*, *9*(2), 359–368. https://doi.org/10.22237/jmasm/1288584240

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

R Core Team. (2018). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.r-project.org/.

Revelle, W. (2018). *psych: Procedures for psychological, psychometric, and personality research*. https://cran.r-project.org/package=psych

Stevens, J. P. (2009). *Applied multivariate statistics for the social science* (5th ed.). Taylor & Francis.

Stout, W. (1987). A nonparametric approach for assessing latent trait unidimensionality. *Psychometrika*, *52*(4), 589–617. https://doi.org/10.1007/BF02294821

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* (6th ed.). Pearson.

Ullah, M. I., Aslam, M., Altaf, S., & Ahmed, M. (2019). Some new diagnostics of multicollinearity in linear regression model. *Sains Malaysiana*, *48*(9), 2051–2060. https://doi.org/10.17576/jsm-2019-4809-26

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

Watkins, M. W. (2018). Exploratory factor analysis: A guide to best practice. *Journal of Black Psychology*, *44*(3), 219–246. https://doi.org/10.1177/0095798418771807

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

Yang, Y., & Xia, Y. (2015). On the number of factors to retain in exploratory factor analysis for ordered categorical data. *Behavior Research Methods*, *47*(3), 756–772. https://doi.org/10.3758/s13428-014-0499-2

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. https://doi.org/10.1037/0033-2909.99.3.432