ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2019; 10(1);1-11

A Preliminary Study to Evaluate the Reproducibility of Factor Analysis Results: The Case of Educational Research Journals in Turkey*

Burak AYDIN ** Mehmet KAPLAN *** Hakan ATILGAN ****
Sungur GÜREL *****

Abstract
In quantitative research, an attempt to reproduce previously reported results requires at least a transparent definition of the population, sampling method, and the analyses procedures used in the prior studies. Focusing on the articles published between 2010 and 2017 by the four prestigious educational research journals in Turkey, this study aimed to investigate the reproducibility of the factor analysis results from a theoretical perspective. A total of 275 articles were subject to descriptive content analysis. Results showed that 77.8% of the studies did not include an explicit definition of the population under interest, and in 50.9% of the studies, the sampling method was either not clear or reported to be convenience sampling. Moreover, information about the missing data or a missing data dealing technique was absent in the 76% of the articles. Approximately, half of the studies were found to have inadequate model fit. Furthermore, in almost all studies, it could not be determined whether the item types (i.e., levels of measurement scales) were taken into consideration during the analyses. In conclusion, the majority of the investigated factor analysis results were evaluated to be non-reproducible in practice.

Key Words: Reproducibility, factor analysis, descriptive content analysis

INTRODUCTION

The Open Science Collaboration (OSC) team reviewed several academic articles published in three respected psychology journals, investigated the reproducibility of the reported results in a total of 100 experimental or correlational studies, and stated that most of the results in those articles could not be obtained again (Open Science Collaboration, 2015). This reproducibility crisis was subject to both negative criticisms (e.g., Gilbert, King, Pettigrew, & Wilson, 2016) and supportive reports (e.g., Anderson et al., 2016). The negative criticism by Gilbert et al. (2016) stated that the reproducibility study by the OSC team had three issues that are sampling error, low statistical power, and bias. Hence the authors concluded that the OSC team seriously underestimated the reproducibility. This conclusion however criticized by Anderson et al. (2016) stating that Gilbert et al. (2016)'s study was very optimistic and based on statistical misconceptions and selective interpretations. Following the crisis, several steps such as journal policies that encourage to share data sets and the software scripts, and academic collaborations that promote open science (e.g., Moshontz et al., 2018) have been taken into consideration to overcome reproducibility issues in scientific research. Especially in social science, negligence in appropriate use of sample selection procedures and data analysis are the two main sources of error that may reduce the reproducibility rates of the results.

* Preliminary results of this work were presented at the 5th Congress on Measurement and Evaluation in Education and Psychology, Antalya 2016.
** Adjunct Professor, Recep Tayyip Erdoğan University, Department of Education, Rize-Turkey, burak.r.aydin@gmail.com, ORCID ID: 0000-0003-4462-1784
***Assistant Professor, Artvin Çoruh University, Department of Education, Artvin-Turkey, mehmet.kaplan2@gmail.com, ORCID ID: 0000-0002-4175-3899
**** Associate Professor, Ege University, Department of Education, Izmir-Turkey, hakan.atilgan@ege.edu.tr , ORCID ID: 0000-0002-5562-3446
*****Assistant Professor, Siirt Üniversitesi, Department of Education, Siirt-Turkey, s.gurel@siirt.edu.tr, ORCID ID: 0000-0003-3425-858X

To cite this article:
Aydın, B., Kaplan, M., Atılgan, H., & Gürel, S. (2019). A preliminary study to evaluate the reproducibility of factor analysis results: The case of educational research journals in Turkey. *Journal of Measurement and Evaluation in Education and Psychology, 10*(1), 1-11. DOI: 10.21031/epod.482393

Received: 13.11.2018
Accepted: 13.02.2019

The sampling method is an important part of quantitative research because inaccurate representation of the population can threaten the external validity of the study. Sampling methods can be classified in various ways (e.g., Balcı, 2000, Lavrakas, 2008; Kish, 1965; Levy & Lemeshow, 2013; Neuman, 2013); however, a most common categorization is known as probability sampling or non-probability sampling methods. Regardless of the sampling method, the use of inadequately small sample size and the existence of non-response or response bias (Lewis-Beck, Bryman & Liao, 2004) can result in non-reproducible findings in quantitative research. In addition, selection bias resulting from the non-probability-based methods is also another source of non-reproducible results. It is also important to take the sampling method into consideration when analyzing the data. Sterba (2009) discussed Neyman's and Fisher's frameworks to address sampling techniques when making statistical inferences. Fisher's framework requires three prerequisites with non-probability-based sampling methods, a correct statistical model, a valid distributional assumption, and conditionality. The conditionality assumption is not satisfied if the sampling technique is not taken into consideration (i.e., clustered or stratified sampling) and if the non-random sample fails to mimic a random sample due to disproportionally selected cases. On the other hand, Neyman's framework was created exclusively for random sampling methods (Sterba, 2009). Thus, the appropriate selection of sampling method and adequate data analysis play a vital role to increase reproducibility of research findings.

Inspired by the OSC's work (Open Science Collaboration, 2015) and their definition of the direct replication as an attempt to recreate the conditions to obtain previous findings, this study aims to show whether a team of researchers will have difficulties if they attempt to recreate the conditions and reproduce the results in the educational research articles published by the journals headquartered in Turkey. Hence, a preliminary study was designed to conduct a descriptive content analysis to investigate the sampling methods and data analysis procedures in these journals. To create a manageable study, the content was narrowed to factor analyses.

Factor Analysis in Educational Research

Educational researchers might reach conclusions using scores derived from a measurement tool, and in such cases, the validity of the conclusions is not independent of the validity of the scores. Scale validity is a unitary concept; however, evidence to support validity can be sought through several dimensions. One of these dimensions is known as construct validity (Atılgan, Kan, & Aydın, 2017; Nunnally & Bernstein, 1994). A psychological construct cannot be defined unless it is measurable (Crocker & Algina, 1986; Lord & Novick, 1968) and one of the procedures to provide evidence for the construct validity is the factor analysis. The use of factor analysis in educational research has been popular when developing a new scale or adapting a scale for cross-validation using confirmatory factor analysis (CFA) or explanatory factor analysis (EFA). CFA is also common when using a developed scale in quantitative research. For example, Göktaş et al. (2012), focusing on the studies conducted in Turkey, investigated 2111 articles published in 19 journals between 2005 and 2009 and identified a measurement tool in 1794 studies. A similar finding was reported by Karadağ (2011) who examined 211 doctoral dissertations completed between 2003 and 2007. Yılmaz and Altınkurt (2012), Sözbilir, Güler, and Çiltaş (2012), Selçuk, Palancı, Kandemir, and Dündar (2014), Kozikoğlu and Senemoğlu (2016), Yalçın, Yavuz, and Dibek (2016), and Gökmen et al. (2017) also noticed the common use of measurement tools both in national and international journals. Scale development and adaptation studies are also common in national journals. For example, Öztürk, Eroğlu, and Kelecioğlu (2015) identified 108 adaptation studies published in 10 journals between 2005 and 2014. The common use of scale development and adaptation was also noticed by Gül and Sözbilir (2015). Readers interested in further details about factor analysis and their role in educational research are referred to Acar (2014), Büyüköztürk (2002), Çüm and Koç (2013), Erkuş (2016), Güvendir and Özkan (2015), Kline (2015), Öztürk, Eroğlu, and Kelecioğlu (2015), Prudon (2015), Yurdugül and Bayrak (2012), Worthington and Whittaker (2006), and Wright (2017).

Results obtained with factor analysis are not independent from the sample. For example, Simon (1979) completed one of the studies that revealed the importance of sample selection in factor analysis. The

author wanted to draw attention that an attitude scale validated with a sample of university students could work differently for non-university students. His first sample consisted of 188 students from a single university, while the second sample consisted of 188 different individuals with the help of a foundation operating on a national basis. The author used the same factor analysis techniques on two different samples and reached different factor structures. At this point, it should be noted that, in the factor analysis, the sample should not represent a country, a territory, or a society, but it needs to represent the behaviors to be measured. Another study, which put forth the importance of sample selection in EFA, was completed by Gaskin, Orellana, Bowe, and Lambert (2017). The authors studied the construct validity of a scale used by the World Health Organization to determine whether individuals were generally healthy. In a study, in which 31251 individuals over the age of 50 from six different countries were considered as the population, the authors tested two different sampling methods. In the first approach, 1000 different samples were selected using simple random sampling to reflect the skewed distribution of the 31251 individuals' total health scores. In the second approach, 1000 different samples were selected with stratified random sampling to reach normally distributed scores. Exploratory factor analyses were performed on selected samples. With random sampling, generally a single factor solution was reached, whereas with the stratified sample a two-factor structure was reached. The authors found the structure obtained by stratified random sampling to be more defensible. These results showed that the sample can support different factor structures even when using probability-based methods. In addition, these results emphasize the importance of using prior knowledge about the population in sampling (Smith, 1983). From the sample perspective, one of the factors that make reproducibility difficult is using convenience sampling. The convenience sampling method can compromise the accuracy of the results in exchange for saving time and money (Balcı, 2015). The probability that a sample reached by the convenience sampling method is representative of any population greater than itself is usually very low. The validity of the results obtained by convenience sampling method has a high degree of concern, and this has been the subject of several academic studies (Bornstein, Jager, & Putnick, 2013; Delice, 2010; Landers & Behrend, 2015; Peterson & Merunka, 2014; Tyrer & Heyman, 2016).

After determining a sampling method that can represent the population, another important issue for reproducibility is the sample size. The sample needs to be sufficiently large to achieve unbiased estimates in factor analysis. Using an appropriate sample size may vary depending on the complexity of the factor structure, the magnitude of the factor loadings and the missing data. To determine the appropriate sample size in their studies, researchers can use the Monte Carlo simulation studies (Wolf, Harrington, Clark, & Miller, 2013). In other words, the definition of the population, choice of the sampling method and the sample size play an important role in factor analysis, and they affect the accuracy of the psychometric properties of the measurement tool. The factors obtained by factor analysis are affected by the sample (Kline, 2015; Thompson, 2004).

From a technical point of view, factor analysis is a dimension reduction process. The responses to n different questions in a scale form an n × n covariance matrix, and the factor analysis searches for a solution to produce this matrix using a smaller number of variables (Crocker & Algina, 1986). In other words, the variance with the n different variables is tried to be represented by a smaller number of variables, i.e., factors. This dimension reduction process can be quite complex depending on, for example, the number of questions, the relationship between items, how the missing data is handled, and the characteristics of the estimation method. Several sources address all the technical parts of factor analysis (e.g., Büyüköztürk, 2002; Crocker & Algina, 1986; Kline, 2015; Prudon, 2015; Thompson, 2004). A structure revealed by an EFA or CFA may not be reproduced with a similar sample if the missing data technique is not known (Akbaş & Tavşancıl, 2015; Çüm & Gelbal, 2015; Kürşad & Nartgün, 2015) and if the estimation method is not clearly defined (Beauducel & Herzberg, 2006; Hox, 1995). In addition, it should be clear whether the items were treated as categorical or continuous variables (Rhemtulla, Brosseau-Liard, & Savalei, 2012, Yang-Wallentin, Jöreskog, & Luo, 2010). Model-data fit information can also provide clues for reproducible findings (Prudon, 2015).

Overall, any attempt to reproduce results of a factor analysis requires detailed information about the sampling method and the analysis procedure. As stated earlier, the purpose of this study is to show

whether a team of researchers will have difficulties if they attempt to recreate the conditions and reproduce the factor analysis results reported in the educational research articles published by the journals headquartered in Turkey. The research questions are set to be:

1. Is the definition of the population explicit?
2. Which sampling methods are used?
3. What are the sample sizes, number of items and factors?
4. How is the missing data handled?
5. Which software is used?
6. Are the levels of measurement scales (categorical or continuous) taken into consideration and which estimators were used?
7. What is the reported data-model fit information?

METHOD

The scope of the study was limited to four internationally indexed educational research journals headquartered in Turkey, namely, Eurasian Journal of Educational Research (EJER), Educational Sciences Theory and Practice (ESTP), Hacettepe University Journal of Education (HUJE), and Education and Science (ES). Because it was not feasible to examine all the studies published in these journals with a small research team, the boundaries of this study were limited by the publication date and research topic. Specifically, the articles published between January 2010 and December 2017 including the keywords related to the factor analysis, which is one of the most commonly used data analysis method in educational research, were selected to be reviewed in this study. More specifically, to identify articles that reported factor analysis in the specified date range, keywords of *development, adaptation, factor analysis, structural equation modeling, validity, reliability, confirmatory, exploratory, CFA, EFA, Cronbach* or their Turkish translations were searched and a total of 341 academic articles were downloaded to be reviwed for the purpose of this study. Articles in each journal were examined by one of the four authors in our research team, and it was narrowed down to 275 out of 341 articles where CFA, EFA, or Principal Component Analysis (PCA) were used for the data analysis. These 275 articles were then investigated in a descriptive content analysis framework. The descriptive content analysis is one of the quantitative data analysis methods and usually includes reporting of basic statistics such as frequency, average, median, and variance (Gall, Gall, & Borg, 1996; Stapleton & Leite, 2005).

Data Collection

Title, publication year, publishing journal, and general purpose in 275 articles selected for this study were recorded. Specifically, the general purpose of the study was coded as scale development, scale adaptation, or other. The sampling characteristics, sampling method and the clear definition of the population were considered as the first dimension of reproducibility. The content of the sample used in those studies was coded as *students, teachers or prospective teachers, academicians, administrators*, or *other*. The data analysis procedures, which were considered as the second dimension of the reproducibility, were also examined in this study. Specifically, the following criteria were recorded: (i) whether the missing data was reported, (ii) whether the missing data was handled using an appropriate technique, (iii) whether an EFA and CFA were performed using the same sample, (iv) sample size, (v) number of items in scales, (vi) number of factors found, (vii) items types (e.g., Likert or yes/no), (viii) software, and (ix) model-data fit information. The data analysis techniques were coded as explanatory or confirmatory. It is worth to note that PCA was considered as an exploratory technique (Bryant & Yarnold, 1995). For the model fit information, the ratio of the chi-square to the degrees of freedom, the root of the square error of approximation (RMSEA), standardized root mean square residual (SRMR), comparative fit index (CFI), Tucker-Lewis index (TLI or NNFI), normative fit index (NFI), goodness of fit index (GFI), adjusted GFI (AGFI), incremental fit index (IFI), and relative fit index (RFI) were recorded. In addition, if more than one scale was used in an article, number of items, number of factors, type of the items, and fit information were recorded on a different row for the same

article. Also, if more than one model was tested for the same scale, only the information of the final model was recorded. As a result, the final data set consisted of 448 rows in total.

FINDINGS

The number of published articles selected for this study was 35 (12.7%) in 2010, 32 (11.6%) in 2011, 35 (12.7%) in 2012, 46 (16.7%) in 2013, 53 (19.3%) in 2014, 28 (10.2%) in 2015, 18 (6.5%) in 2016, and 28 (10.2%) in 2017. In addition, the frequency of the articles by the journals was 94 (34.2%), 56 (20.4%), 40 (14.5%), and 85 (30.9%) for the ES, EJER, HUJE, and ESTP, respectively. The frequency of studies in scale development was 108 (39.3%), in scale adaptation was 99 (36.0%), and in other topics was 68 (24.7%). Table 1 shows the frequencies of the 275 articles by year, journal, and research purpose.

Table 1. The Frequencies of the 275 Articles by Year, Journal, and Study Purpose.

Year	ES			EJER			HUJE			ESTP			Total
	SD	SA	O	SD	SA	O	SD	SA	O	SD	SA	O	
2010	3	5	5	1	3	1	4	4	0	6	3	0	35
2011	3	8	3	1	0	1	3	1	0	6	5	1	32
2012	4	3	2	2	3	0	3	6	0	9	1	2	35
2013	5	11	2	2	3	1	3	5	0	6	6	2	46
2014	12	12	4	5	0	5	3	1	0	5	3	3	53
2015	1	0	2	1	2	7	4	1	2	2	0	6	28
2016	0	2	2	2	2	4	0	0	0	2	4	0	18
2017	1	2	2	5	2	3	0	0	0	4	1	8	28
Total	29	43	22	19	15	22	20	18	2	40	23	22	275

Note: SD = Scale development, SA = Scale adaptation, O = Other.

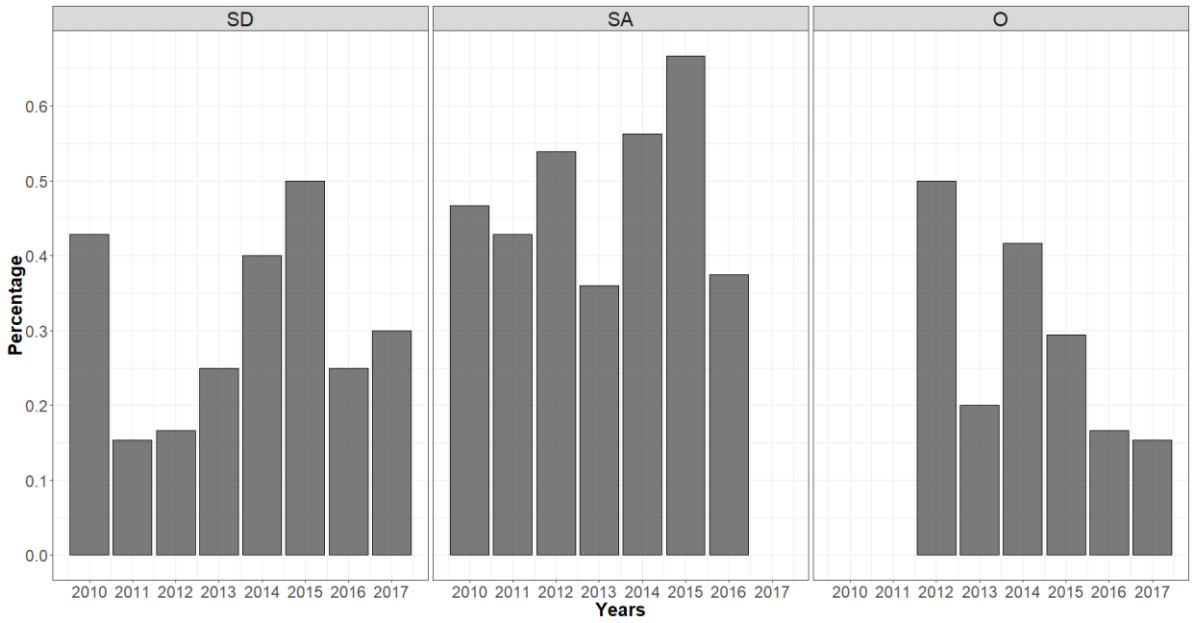
Definition of Population and Sampling Method

A clear definition of the population and the appropriate selection of the sampling method in quantitative research are important for ensuring the validity of the results. Based on the results, only 61 (22.2%) of the 275 articles reviewed in this study provided a clear definition of the population in their research. Table 2 shows the percentage of the studies that explicitly reported the population definition by year and study purpose. Scale development and adaptation studies included a clear definition approximately 1 in every 5 studies, whereas other studies had a rate of 1 in every 3. On the contrary to unclear definition of the population, the sampling method, whether probability-based or non-probability-based, was determined in 227 (82.6%) of the articles. More specifically, 169 of those 227 studies used a non-probability-based sampling, and 58 used a probability-based sampling. Of the 169 studies, the sampling technique was clearly stated in 112 articles where 92 of them were convenience, 11 of them were purposeful, 5 of them were stratified, 2 of them were maximum variation, 1 of them was snowball, and 1 of them was typical case sampling. In general, 48 (17.4%) of the 275 studies did not have a clear definition of the sampling method and 92 (33.5%) of the 275 stated that convenience sampling was used. Figure 1 shows the percentage of convenience sampling across years and study purpose. Overall 31%, 49% and 29% of the studies reported the use of convenience sampling for scale development, scale adaptation and other purposes respectively. In addition, the content of the sample was clearly defined in all articles. Specifically, 205 (74.6%) of the studies included only students, 40 (14.6%) of them included teachers or prospective teachers, 8 (2.9%) of them included only academicians, 4 (1.4%) of them included only administrators, and the remaining 18 (6.4%) of them included at least two of these groups or other individuals (e.g., parents and adults).

Table 2. Population and Missing Data Information of the 275 Articles by Year and Study Purpose

Year	Population information percentage			Missing data information percentage		
	SD	SA	O	SD	SA	O
2010	21	33	17	29	13	33
2011	15	21	40	31	21	40
2012	28	8	50	22	23	25
2013	25	12	40	13	20	0
2014	12	6	25	20	13	25
2015	0	67	41	25	67	29
2016	0	38	50	75	25	17
2017	40	20	8	10	40	46
All years	19	19	31	23	21	29
Overall	22			24		

Note: SD = Scale development, SA = Scale adaptation, O = Other.



Note: SD = Scale development, SA = Scale adaptation, O = Other.

Figure 1. Percentage of Convenience Sampling Across Years and Study Purpose

Sample Size, Number of Items, Number of Factors, and Item Types

The sample size, number of items, number of factors, and item types were recorded separately for 448 analyses in 275 articles. The median values of the observed sample sizes, the number of items used in the scale, and the number of obtained factors were 398, 25, and 3, respectively. In addition, the median value of the sample size per item was 14.8, and the number of items per factor was distributed with a median of 7. Table 3 shows the median values for sample size, item per factor and sample size per item by year, and study purpose. Item per factor median values were similar across years and purpose. However, sample size median values across all years were slightly lower for the scale development and adaptation, 381 and 400 respectively, compared to the median value for the other purposes which was 459. The sample size per item median values across all years were similar for the scale adaptation and other studies, 16.6 and 17.7, respectively, slightly larger compared to scale development values which was 12.2. Items with more than two categories (e.g., Likert) were employed in 228 (82.9%) of the 275 articles, whereas 7 (2.5%) studies used binary, 4 (1.5%) studies used continuously scaled

items, and the item type could not be determined for the remaining 36 (13.1%) studies. Furthermore, a total of 318 individual analyses out of 448 reported the item type, and out of these 318, 304 used items with more than two categories. The most preferred items (i.e., in 209 analyses) were the ones with five categories. Items with three, four, six, seven, nine, and ten categories were also used in 11, 36, 16, 24, 5, and 3 analyses, respectively.

Missing Data and Analysis Procedure

Of the 275 articles reviewed, only 66 (24%) reported how the missing data were handled. Of these 66 studies, 62 utilized listwise deletion and 3 utilized an imputation method. In addition, it was reported that there was no missing data at all in 1 study. Table 2 shows the percentage of the studies that included missing data information by year and study purpose. Similar to population definition rates, scale development and adaptation studies had a lower rate, 23% and 21% respectively, compared to other studies, 29%.

For the data analysis method, it was determined that 84 (30.6%) of the studies employed only CFA, 57 (20.7%) employed only EFA, and 134 (48.7%) employed both EFA and CFA. 90 of 134 articles that employed both EFA and CFA conducted analyses using the same sample, or they divided the study sample into halves. The software information could be identified in 183 of the 275 articles. Specifically, SPSS and Lisrel together, Lisrel, SPSS, AMOS, SPSS and AMOS together, M*plus*, and EQS were used in 71, 49, 29, 19, 12, 2, and 1 studies, respectively.

Table 3. Median Values of Sample Size, Item per Factor and Sample Size per Factor of the 448 Analyses by Year and General Purpose

Year	Sample size median			Item per factor median			Sample size per item median		
	SD	SA	O	SD	SA	O	SD	SA	O
2010	464	358	367	8.9	8.3	8.5	12.2	21.3	10.0
2011	461	341	214	8.5	7.6	4.0	12.4	12.8	16.6
2012	336	529	258	6.1	7.0	12.5	10.8	13.6	6.0
2013	388	407	605	6.5	6.0	4.0	10.7	21.9	97.9
2014	317	436	256	6.0	5.0	7.0	12.9	25.6	14.9
2015	384	357	657	10.7	6.3	6.1	13.0	9.4	49.2
2016	330	462	556	4.3	5.3	7.0	12.3	15.5	20.4
2017	303	270	719	7.3	5.7	6.6	11.0	16.4	27.8
All years	381	400	459	7.3	6.5	7.0	12.2	16.6	17.7
Overall	398			7				14.8	

Note: SD = Scale development, SA = Scale adaptation, O = Other.

Data-model Fit Information

The ratio of chi-square by the degrees of freedom was reported in 183 studies, and it ranged between 1.01 and 9.45 with a median value of 2.66; RMSEA was reported in 245 studies ranged between 0 and 0.44 with a median of 0.06; SRMR was reported in 131 studies ranged between 0.004 and 0.11 with a median of 0.05; CFI was reported in 233 studies ranged between 0.70 and 1 with a median of 0.96; TLI was reported in 143 studies between 0.69 and 1 with a median of 0.96; NFI was reported in 146 studies ranged between 0.64 and 1 with a median of 0.95; GFI was reported in 197 studies ranged between 0.47 and 1 with a median of 0.92; AGFI was reported in 157 studies ranged between 0.07 and 1 with a median of 0.90; IFI was reported in 54 studies ranged between 0.81 and 1 with a median of 0.95; and finally RFI was reported in 41 studies ranged between 0.62 and 1 with a median of 0.96. The estimator was determined only in 39 analyses, and 30 of them utilized maximum likelihood, 7 used robust maximum likelihood, and 2 used least squares methods.

DISCUSSION and CONCLUSION

Inspired by the reproducibility crisis in psychology research (Open Science Collaboration, 2015), this preliminary study aimed to evaluate the reproducibility of the results using factor analysis in four prestigious educational research journals headquartered in Turkey. The authors examined 448 different analyses reported in 275 articles published between 2010 and 2017 based on sampling method and data analysis procedures which were considered as two of the main dimensions of reproducible research.

Factor analyses were generally employed with a purpose of either scale development or scale adaptation in 75.3% of the 275 articles and they were used with different purposes for the remaining 24.7% of the articles. A clear definition of the population was not found in 77.8% of the studies which can be evidence for the threat to the validity. The number of articles in which the sampling method could not be determined or determined as the convenience sampling was 140 (50.9%). In 76% of the studies, the information about how the missing data was handled could not be identified, and the ones where the missing data was reported used outdated techniques, such as listwise deletion and mean imputation. In 90 of 275 studies, both the EFA and CFA were utilized using the same sample. The results obtained by EFA and CFA using the same data have been a subject of debate (Erkuş, 2016; Van Prooijen & Van Der Kloot, 2001). Considering the importance of a clear definition of the population and the use of proper sampling method that can produce generalizable results, these findings were evaluated as the evidence of non-reproducible results in those articles. Handling of missing data is an important part in factor analysis (Allison, 2003; Çüm, & Gelbal, 2015), as for the social sciences in general (Schafer, 1997; Schlomer, Bauman, & Card, 2010). The fact that the missing data was not explicitly addressed in the examined studies increased the concern for non-reproducible results in those articles. The missing data issue in the educational research conducted in Turkey was also noticed by Demir and Parlak (2012). Çüm and Gelbal (2015) stated that in the case of misuse of missing data techniques, the results could be misleading, and this is directly related to the reproducibility of the results. It is not clear why the missing data or the missing data technique were not mentioned in three of the four examined studies, if there were no missing data at all and it was due to forced responses, this is also alarming in terms of reproducibility (Ray, 1990; Xiao, Liu & Li 2017).

In factor analysis, another important issue regarding reproducibility of results is to provide adequate sample size (Wolf, Harrington, Clark, & Miller, 2013). Selecting an adequate sample size depends on the complexity of the model and the magnitude of the factor loadings. Monte Carlo simulations are powerful techniques that can be used to determine the appropriate size, but in the literature, there are recommendations for the ratio of the number of participants to the number of items, for example, 1 to 20 and 1 to 10 (Hogarty, Hines, Kromrey, Ferron, & Mumford, 2005). In the articles examined in this study, the median value of this ratio was found to be approximately 15, and in general, it was evaluated that the importance of sample size was recognized. The average number of items per factor was fewer than seven in half of the studies. In theory, if there are multiple factors in the model, a factor can be defined with two items, but it is recommended to have at least 3, 4, or 5 items per factor (Kline, 2015). Increasing the number of items can allow for a strong definition of the structure, thus enhance the reproducibility. In general, it was evaluated that the importance of the number of items per factor was not recognized in the articles examined for this study.

The model-data fit information used in factor analysis is a clue for the reproducibility of the results. Fit values would be low if there were unexplained variance sources or the model was not correctly specified, and this poses a risk for reproducibility. For the model-data fit information, what should be the cut-off values is the subject of several studies (Kline, 2015; Marsh, Balla, & McDonald 1988; O'Boyle & Williams 2011; Prudon, 2015), assuming RMSEA<0.06, SRMR <0.08, CFI, and TLI (NNFI, NFI, GFI, and AGFI> 0.95 indicate a good fit, nearly half of the studies examined were found to have difficulty in meeting these criteria. The ratio of the chi-square to the degrees of freedom was not taken into consideration in our evaluation, given that it should not be used (Kline, 2015). Furthermore, the fact that the estimator information was not identified in most of the analyses prevented us to determine whether the characteristics of the items were taken into consideration during the analysis process and this is another concern, as when the normality assumption is not met, treating categorical (e.g., Likert) variables as continuous is likely to harm reproducibility (Li, 2016).

Overall the majority of the investigated factor analysis results were evaluated to be non-reproducible in practice. This non-reproducibility issue seems to be more evident for the scale development and adaptation studies compared to studies with other quantitative purposes given that the later has better rates of a clear definition of the population and missing data, along with relatively larger sample sizes and decreasing number of convenience sampling utilization. This study has its limitations. One of them is that the scope is broad; however, as the title indicates, this is a preliminary study to show an alarming issue, namely, a possible reproducibility crisis of educational research studies published by Turkish Journals. Researchers are invited to conduct more in-depth reproducibility studies for example with a focus on particular scales, EFA and rotation options (e.g., Kline, 2015; Osborne, 2015; Saracli, 2011), CFA and modification issues (e.g., Asparouhov & Muthen, 2009; Mueller & Hancock, 2008). The second limitation is that model-fit information is affected at least by the sample size, estimator, and model specification; hence, the model-fit information was not considered as a main indicator of reproducibility, but rather considered as clues. The third limitation is that no guideline was provided for the practitioners. However, it was made clear that any attempt to recreate conditions to reproduce a practitioner`s results will fail if the population, sampling method, and the analyses procedures were not represented transparently. When these reproducibility basics are fulfilled, practitioners should take advantage of already published guidelines, for example, Büyüköztürk (2002), Erkuş (2016), Kline (2015), Öztürk, Eroğlu and Kelecioğlu (2015), Prudon (2015), Worthington and Whittaker (2006), and Wright (2017). It is also strongly recommended for practitioners to share their data-set and data analysis syntax whenever possible. The list of 275 articles investigated in this preliminary study and the data set including information from 448 analyses are provided as supplementary files.

REFERENCES

Acar, T. (2014). Ölçek geliştirmede geçerlik kanıtları: Çapraz geçerlik, sınıflama ve sıralama geçerliği uygulaması. *Kuram ve Uygulamada Eğitim Bilimleri*, *14*(2), 1-11. DOI: 10.12738/estp.2014.3.2107

Akbaş, U., & Tavşancıl, E. (2015). Farklı Örneklem büyüklüklerinde ve kayıp veri örüntülerinde ölçeklerin psikometrik özelliklerinin kayıp veri baş etme teknikleri ile incelenmesi. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi*, *6*(1) 38-57.

Allison, P. D. (2003). Missing data techniques for structural equation modeling. *Journal of Abnormal Psychology*, *112*, 545-557.

Anderson, C. J., Bahník, Š., Barnett-Cowan, M., Bosco, F. A., Chandler, J., Chartier, C. R., ... & Della, N. P. (2016). Response to Comment on" Estimating the reproducibility of psychological science". *Science (New York, NY)*, *351*(6277), 1037-1037.

Asparouhov, T., & Muthén, B. (2009). Exploratory structural equation modeling. *Structural equation modeling: a multidisciplinary journal*, *16*(3), 397-438.

Atılgan, H., Kan, A., & Aydın, B. (2017). *Eğitimde Ölçme ve Değerlendirme*. Ankara: Anı Yayıncılık.

Balcı, A. (2015) *Sosyal Bilimlerde Araştırma: Yöntem, Teknik ve İlkeler*. Ankara: Pegem Yayınları.

Beauducel, A., & Herzberg, P. Y. (2006). On the performance of maximum likelihood versus means and variance adjusted weighted least squares estimation in CFA. *Structural Equation Modeling*, *13*(2), 186-203. DOI: 10.1207/s15328007sem1302_2

Bornstein, M. H., Jager, J., & Putnick, D. L. (2013). Sampling in developmental science: Situations, shortcomings, solutions, and standards. *Developmental Review*, *33*(4), 357-370. DOI: 10.1016/j.dr.2013.08.003

Bryant, F. B., & Yarnold, P. R. (1995). Principal-components analysis and exploratory and confirmatory factor analysis. In L. G. Grimm & P. R. Yarnold (Eds.), *Reading and understanding multivariate statistics* (pp. 99-136). Washington, DC, US: American Psychological Association.

Büyüköztürk, Ş. (2002). Faktör analizi: Temel kavramlar ve ölçek geliştirmede kullanımı. *Kuram ve Uygulamada Eğitim Yönetimi*, *32*, 470-483.

Crocker, L. M., & Algina, J. (1986). *Introduction to classical and modern test theory*. New York: Holt, Rinehart, Winston.

Çüm, S., & Gelbal, S. (2015). Kayıp veriler yerine yaklaşık değer atamada kullanılan farklı yöntemlerin model veri uyumu Üzerindeki etkisi. *Mehmet Akif Ersoy Üniversitesi Eğitim Fakültesi Dergisi*, *1*(35), 87-111.

Çüm, S., & Koç, N. (2013). Türkiye'de psikoloji ve eğitim bilimleri dergilerinde yayımlanan ölçek geliştirme ve uyarlama çalışmalarının incelenmesi. *Journal of Educational Sciences & Practices*, *12*(24), 115-135.

Delice, A. (2010). Nicel araştırmalarda örneklem sorunu. *Kuram ve Uygulamada Egitim Bilimleri*, *10*(4), 1969-2018.

Demir, E., & Parlak, B. (2012). Türkiye'de eğitim araştırmalarında kayıp veri sorunu. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi*, *3*(1), 230-241.

Erkuş, A. (2016). Ölçek geliştirme ve uyarlama çalışmalarındaki sorunlar ile yazım ve değerlendirilmesi. *Pegem Atıf İndeksi*, 1211-1224.

Gall, M. D., Borg, W. R., & Gall, J. P. (1996). *Educational research: An introduction*. White Plains, NY: Longman USA.

Gaskin, C. J., Orellana, L., Bowe, S. J., & Lambert, S. D. (2017). Why sample selection matters in exploratory factor analysis: Implications for the 12-item world health organization disability assessment schedule 2.0. *BMC Medical Research Methodology*, *17*(1), 40.

Gilbert, D. T., King, G., Pettigrew, S., & Wilson, T. D. (2016). Comment on estimating the reproducibility of psychological science. *Science*, *351*(6277), 1037-1037.

Gökmen, Ö. F., Uysal, M., Yaşar, H., Kirksekiz, A., Güvendi, G. M., & Horzum, M. B. (2017). Türkiye'de 2005-2014 yılları arasında yayınlanan uzaktan eğitim tezlerindeki yöntemsel eğilimler: Bir İçerik analizi. *Eğitim ve Bilim*, *42*(189), 1-25.

Göktaş, Y., Küçük, S., Aydemir, M., Telli, E., Arpacık, Ö., Yıldırım, G., & Reisoğlu, I. (2012). Türkiye'de eğitim teknolojileri araştırmalarındaki eğilimler: 2000-2009 dönemi makalelerinin içerik analizi. *Kuram ve Uygulamada Eğitim Bilimleri Dergisi*, *12*(1), 177-199.

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.

Güvendir, M. A., & Özkan, Y. Ö. (2015). Türkiye'deki eğitim alanında yayımlanan bilimsel dergilerde ölçek geliştirme ve uyarlama konulu makalelerin incelenmesi. *Elektronik Sosyal Bilimler Dergisi*, *14*(52), 23-33.

Hogarty, K. Y., Hines, C. V., Kromrey, J. D., Ferron, J. M., & Mumford, K. R. (2005). The quality of factor solutions in exploratory factor analysis: The influence of sample size, communality, and overdetermination. *Educational and Psychological Measurement*, *65*(2), 202-226. DOI: 10.1177/0013164404267287

Hox, J. J. (1995). Amos, EQS, and Lisrel for windows: A comparative review. *Structural Equation Modeling: A Multidisciplinary Journal*, *2*(1), 79-91.

Karadağ, E. (2011). Eğitim bilimleri doktora tezlerinde kullanılan Ölçme araçlar: Nitelik düzeyleri ve analitik hata tipleri. *Kuram ve Uygulamada Eğitim Bilimleri*, *11*(1), 311-334.

Kish, L. (1965). *Survey sampling*. Oxford. England:John Wiley & Sons.

Kline, R. B. (2015). *Principles and practice of structural equation modeling*. New York, NY, US: Guilford publications.

Kozikoğlu, I., & Senemoğlu, N. (2016). Eğitim programları ve Öğretim alanında yapılan doktora tezlerinin içerik analizi (2009-2014). *Eğitim ve Bilim*, *40*(182), 29-41.

Kürşad, M. Ş., & Nartgün, Z. (2015). Kayıp veri sorununun çözümünde kullanılan farklı yöntemlerin Ölçeklerin geçerlik ve güvenirliği bağlamında karşılaştırılması. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi*, *6*(2), 254-267. DOI: 10.21031/epod.95917

Landers, R. N., & Behrend, T. S. (2015). An inconvenient truth: Arbitrary distinctions between organizational, mechanical turk, and other convenience samples. *Industrial and Organizational Psychology*, *8*(2), 142-164.

Lavrakas, P. J. (2008). *Encyclopedia of survey research methods*. Thousand Oaks: Sage Publications.

Levy, P. S., & Lemeshow, S. (2013). *Sampling of populations: Methods and applications*. John Wiley & Sons.

Lewis-Beck, M. S., Bryman, A. , & Liao, T. F. (2004). *The Sage Encyclopedia of social science research methods.* Thousand Oaks: Sage.

Li, C. H. (2016). Confirmatory factor analysis with ordinal data: Comparing robust maximum likelihood and diagonally weighted least squares. *Behavior Research Methods*, *48*(3), 936-949. DOI: 10.3758/s13428-015-0619-7.

Lord, F. M., and Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, Mass.: Addison-Wesley

Marsh, H. W., Balla, J. R., & McDonald, R. P. (1988). Goodness-of-fit indexes in confirmatory factor analysis: The effect of sample size. *Psychological Bulletin*, *103*(3), 391-410. DOI: 10.1037/0033-2909.103.3.391

Moshontz, H., Campbell, L., Ebersole, C. R., IJzerman, H., Urry, H. L., Forscher, P. S., ... & Castille, C. M. (2018). The Psychological Science Accelerator: Advancing psychology through a distributed collaborative network. *Advances in Methods and Practices in Psychological Science*,*1*(4),501-515. DOI: 10.1177/2515245918797607

Mueller, R. O., & Hancock, G. R. (2008). Best practices in structural equation modeling. In J. W. Osborne (Ed.), *Best practices in quantitative methods* (pp. 488–508). Thousand Oaks: Sage Publications Inc.

Neuman, W. L. (2013). *Social research methods: Qualitative and quantitative approaches*.UK: Pearson education.

Nunnally, J., & Bernstein, I. (1994). *Psychometric Theory (3rd Edition).* New York: McGraw-Hill, Inc.

O'Boyle, E. H., Jr., & Williams, L. J. (2011). Decomposing model fit: Measurement vs. theory in organizational research using latent variables. *Journal of Applied Psychology, 96*(1), 1-12.

Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science, 349*(6251). doi: 10.1126/science.aac4716

Osborne, J. W. (2015). What is rotating in exploratory factor analysis. *Practical assessment, research & evaluation, 20*(2), 1-7.

Öztürk, N. B., Eroğlu, M. G., & Kelecioğlu, H. (2015). Eğitim alanında yapılan ölçek uyarlama makalelerinin incelenmesi. *Eğitim ve Bilim, 40*(178), 123-137.

Peterson, R. A., & Merunka, D. R. (2014). Convenience samples of college students and research reproducibility. *Journal of Business Research, 67*(5), 1035-1041. DOI: 10.1016/j.jbusres.2013.08.010

Prudon, P. (2015). Confirmatory factor analysis as a tool in research using questionnaires: A critique. *Comprehensive Psychology, 4*, 1-19. DOI: 10.2466/03.CP.4.10

Ray, J. J. (1990). Acquiescence and problems with forced-choice scales. *The Journal of Social Psychology, 130*(3), 397-399. DOI: 10.1080/00224545.1990.9924595

Rhemtulla, M., Brosseau-Liard, P. É., & Savalei, V. (2012). When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. *Psychological Methods, 17*(3), 354-373. DOI: 10.1037/a0029315

Saracli, S. (2011). Faktör analizinde yer alan döndürme metotlarının karşılaştırmalı incelenmesi üzerine bir uygulama. *Düzce Üniversitesi Sağlık Bilimleri Enstitüsü Dergisi, 1*(3), 22-26.

Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. Chapman and Hall: CRC.

Schlomer, G. L., Bauman, S., & Card, N. A. (2010). Best practices for missing data management in counseling psychology. *Journal of Counseling Psychology, 57*(1), 1-10.

Selçuk, Z., Palancı, M., Kandemir, M., & Dündar, H. (2014). Eğitim ve bilim dergisinde yayınlanan araştırmaların eğilimleri: İçerik analizi. *Eğitim ve Bilim, 39*(173), 428-449.

Simon, A. (1979). Effects of selective sampling on a factor analysis. *The Journal of General Psychology, 101*(2), 259-264. DOI: 10.1080/00221309.1979.9920079

Smith, T. (1983). On the validity of inferences from non-random sample. *Journal of the Royal Statistical Society. Series A (General), 146*(4), 394-403. DOI: 10.2307/2981454

Sözbilir, M., Güler, G., & Çiltaş, A. (2012). Türkiye'de matematik eğitimi araştırmaları: Bir içerik analizi Çalışması. *Kuram ve Uygulamada Eğitim Bilimleri, 12*(1), 565-580.

Stapleton, L. M., & Leite, W. L. (2005). Teacher's corner: A review of syllabi for a sample of structural equation modeling courses. *Structural Equation Modeling, 12*(4), 642-664.

Sterba, S. K. (2009). Alternative model-based and design-based frameworks for inference from samples to populations: From polarization to integration. *Multivariate Behavioral Research, 44*(6), 711-740.

Thompson, B. (2004). *Exploratory and confirmatory factor analysis: Understanding concepts and applications.* Washington, DC: American Psychological Association.

Tyrer, S., & Heyman, B. (2016). Sampling in epidemiological research: Issues, hazards and pitfalls. *BJPsych Bulletin, 40*(2), 57-60.

Van Prooijen, J. W., & Van Der Kloot, W. A. (2001). Confirmatory analysis of exploratively obtained factor structures. *Educational and Psychological Measurement, 61*(5), 777-792. DOI: 10.1177/00131640121971518

Wolf, E. J., Harrington, K. M., Clark, S. L., & Miller, M. W. (2013). Sample size requirements for structural equation models: An evaluation of power, bias, and solution propriety. *Educational and Psychological Measurement, 73*(6), 913-934.

Worthington, R. L., & Whittaker, T. A. (2006). Scale development research: A content analysis and recommendations for best practices. *The Counseling Psychologist, 34*(6), 806-838.

Wright, A. G. (2017). The current state and future of factor analysis in personality disorder research. *Personality Disorders: Theory, Research, and Treatment, 8*(1), 14-25.

Xiao, Y., Liu, H., & Li, H. (2017). Integration of the Forced-Choice Questionnaire and the Likert Scale: A Simulation Study. *Frontiers in Psychology, 8*, 806. DOI: 10.3389/fpsyg.2017.00806

Yalçın, S., Yavuz, H. C., & Dibek, M. I. (2016). En yüksek etki faktörüne sahip eğitim dergilerindeki makalelerin İçerik analizi. *Eğitim ve Bilim, 40*(182), 1-28.

Yang-Wallentin, F., Jöreskog, K. G., & Luo, H. (2010). Confirmatory factor analysis of ordinal variables with misspecified models. *Structural Equation Modeling, 17*(3), 392-423.

Yılmaz, K., & Altınkurt, Y. (2012). An examination of articles published on preschool education in Turkey. *Educational Sciences: Theory and Practice, 12*(4), 3227-3241.

Yurdugül, H., & Bayrak, F. (2012). Ölçek geliştirme Çalışmalarında kapsam geçerlik Ölçüleri: Kapsam geçerlik indeksi ve kappa istatistiğinin karşılaştırılması. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi, Özel Sayı*, *2*, 264-271.