Turkish Journal of Education

TURJE

2012

www.turje.org | July, 2019 | Volume 8, Issue 3 | eISSN: 2147-2858

In this issue our readers will find;

"Nomad Games" Kyrchyn Valley, Kyrgyzstan (Fahri Tarhan, 2018)

Azerbaycan Avenue 64 / 22 46040 Onikişubat - Kahramanmaraş TURKEY

Journal Name: Turkish Journal of Education Türk Eğitim Dergisi
Acronym: TURJE
Turkish and English Quarterly Publishing Scientific Peer Reviewed Online Journal
Türkçe ve İngilizce olarak üç ayda bir yayımlanan Bilimsel Hakemli Elektronik Dergi
eISSN: 2147-2858
Year: 2019
Volume: 8 Issue: 3
URL: www.turje.org, Email: turjeorg@gmail.com, Telephone: +90344-300-1310
Address: Azerbaycan Bulvarı 64/22 46040 Onikişubat Kahramanmaraş TÜRKİYE

Turkish Journal of Education (TURJE) seeks to bridge and integrate the intellectual, methodological, and substantive diversity of educational scholarship, and to encourage a vigorous dialogue between educational scholars and practitioners.
TURJE publishes researches, theoretical statements, philosophical arguments, and critical syntheses.
First published in 2012

Table of Contents İçindekiler

Editorial Board Editörler Kurulu

Reviewers of the Issue Sayı Hakemleri

Volume (8) Issue (3)

Dr. Beyza Aksu Dünya (Bartın University, Turkey)
Dr. Burak Aydın, (Recep Tayyip Erdoğan University, Turkey)
Dr. Duygu Metin Peten, (Yozgat Bozok University, Turkey)
Dr. Filiz Bezci-Bircan, (Çankırı Karatekin University, Turkey)
Dr. Gülşen Taşdelen Teker, (Hacettepe University, Turkey)
Dr. Lokman Akbay, (Mehmet Akif Üniversity, Turkey)
Dr. Melek Gülşah Şahin, (Gazi University, Turkey)
Dr. Nagihan Boztunç Öztürk, (Hacettepe University, Turkey)
Dr. Ragıp Terzi, (Harran University, Turkey)
Dr. Sedat Şen, (Harran University, Turkey)

To join our Editorial or Reviewer Team
please send your CV to: turjeorg@gmail.com

Indexing Dizinler

EBSCO Education Source (Since January, 2013)
Education Source is an International Field Index on Educational Sciences, combines high-quality databases from EBSCO and H.W. Wilson and includes hundreds of unique full-text journals not previously available in any databases.

Emerging Sources Citation Index (ESCI), Web of Science™ Core Collection (Since January, 2015)
The Emerging Sources Citation Index (ESCI) was launched in late 2015 as a new database within Clarivate Analytics' (formerly Thomson Reuters') Web of Science. Journals indexed in the ESCI will not receive Impact Factors; however, the citations from the ESCI will now be included in the citation counts for the Journal Citation Reports, therefore contributing to the Impact Factors of other journals. Articles in ESCI indexed journals will be included in an author's *H-Index* calculation, and also any analysis conducted on Web of Science data or related products such as InCites.

ERIH PLUS European Reference Index for the Humanities (ERIH) to NSD (Since January, 2017)
The European Science Foundation (ESF) signed in January 2014 a memorandum of understanding with the NSD - Norwegian Centre for Research Data to transfer the maintenance and operations of the European Reference Index for the Humanities (ERIH) to NSD. The ERIH database operated by NSD is called ERIH PLUS.

Turkish Journal of Education is a DOAJ Indexing Open Access Journal

TURJE
www.turje.org

Volume: 8 Issue: 3

Turkish Journal of Education

DOI: 10.19128/turje.518636
Research Article

Received 28.01.2019
Accepted 21.05.2019

Comparison of factor retention methods on binary data: A simulation study

Abdullah Faruk Kılıç
Hacettepe University, Ankara, Turkey, abdullahfarukkilic@gmail.com
İbrahim Uysal
Bolu Abant İzzet Baysal University, Bolu, Turkey, ibrahimuysal@ibu.edu.tr

ABSTRACT In this study, the purpose is to compare factor retention methods under simulation conditions. For this purpose, simulations conditions with a number of factors (1, 2 [simple]), sample sizes (250, 1.000, and 3.000), number of items (20, 30), average factor loading (0.50, 0.70), and correlation matrix (Pearson Product Moment [PPM] and Tetrachoric) were investigated. For each condition, 1.000 replications were conducted. Under the scope of this research, performances of the Parallel Analysis, Minimum Average Partial, DETECT, Optimal Coordinate, and Acceleration Factor methods were compared by means of the percentage of correct estimates, and mean difference values. The results of this study indicated that MAP analysis, as applied to both tetrachoric and PPM correlation matrices, demonstrated the best performance. PA showed a good performance with the PPM correlation matrix, however, in smaller samples, the performance of the tetrachoric correlation matrix decreased. The Acceleration Factor method proposed one factor for all simulation conditions. For unidimensional constructs, the DETECT method was affected by both the sample size and average factor loading.

Keywords: *Exploratory factor analysis, Factor retention, Parallel analysis, Minimum average partial, DETECT*

Faktör sayısını belirleme yöntemlerinin karşılaştırılması: Bir simülasyon çalışması

ÖZ Bu araştırmada faktör sayının belirlenmesi amacıyla geliştirilen yöntemlerin simülasyon koşulları altında karşılaştırılması amaçlanmıştır. Bu amaç için faktör sayısı (1, 2 [basit]), örneklem büyüklüğü (250, 1000 ve 3000), madde sayısı (20, 30), ortalama faktör yükü (0.50, 0.70) ve kullanılan korelasyon matrisi (Pearson Momentler Çarpımı [PPM] ve Tetrakorik) simülasyon koşulu olarak araştırılmıştır. Her bir koşul için 1000 replikasyon yapılmış ve üretilen 24000 veri seti için PPM ve tetrakorik korelasyon matrisi üzerinden analizler gerçekleştirilmiştir. Araştırma kapsamında Paralel Analiz, Kısmi Korelasyonların En Küçüğü, DETECT, Optimal Koordinat ve İvmelenme Faktörü yöntemlerinin performansları doğru kestirim yüzdesi ve ortalama fark değerleri üzerinden karşılaştırılmıştır. Araştırma sonucunda hem tetrakorik hem de PPM korelasyon matrisiyle yürütülen MAP analizi en iyi performansı göstermiştir. PA da PPM korelasyon matrisiyle iyi performans göstermiş ancak küçük örneklemde tetrakorik korelasyon matrisiyle performansı düşmüştür. DETECT yöntemi tek boyutlu yapılarda örneklem büyüklüğü ve ortalama faktör yükünden etkilenmiştir.

Anahtar Kelimeler: *Açımlayıcı faktör analizi, Faktör sayısını belirleme, Paralel analiz, Kısmi korelasyonların en küçüğü, DETECT*

Citation: Kılıç, A.F. & Uysal, İ., (2019). Comparison of factor retention methods on binary data: A simulation study. *Turkish Journal of Education, 8*(3), 160-179. DOI: 10.19128/turje.518636

INTRODUCTION

When a construct of data set needs to be analyzed, factor analysis is one of the most common psychometric techniques (Osborne & Banjanovic, 2016). If researchers lack information regarding the construct, explanatory factor analysis (EFA) or unconstrained factor analysis is applied. If researchers have information regarding the construct, and if the compliance of the data sets with the construct in question is to be analyzed, confirmatory factor analysis (CFA) or constrained factor analysis is applied (Fabrigar & Wegener, 2012).

When EFA is applied, one of the obstacles is the determination of the number of dimensions. There are various methods to decide on the number of dimensions. For example, there is a method called the Kaiser K1, which accepts an eigenvalue number higher than 1 as the number of the dimension (Kaiser, 1960). In addition to this method, there is Parallel Analysis (PA), as proposed by Horn (1965), Minimum Average Partial test (MAP), as proposed by Velicer (1976), and the Scree Test, as proposed by Cattell (1966). The Scree Test has some limitations in terms of non-graphical solutions, and there are alternatives. Such methods include multiple regression, a t-value index, and the standard error of scree (SEscree) approach, as proposed by Zoski and Jurs (1993, 1996), the Scree Test Optimal Coordinate (n_{OC}), and the Scree Test Acceleration Factor (n_{AF}), as proposed by Raîche, Walls, Magis, Riopel, and Blais (2013). In addition to these methods, there is the NOHARM (Fraser & McDonald, 1988) method, which is based on latent trait theory and nonlinear harmonic approximations of the normal ogive error distribution. Additionally, DIMTEST (Nandakumar & Stout, 1993; Stout, 1987) and DETECT (Zhang & Stout, 1999) are non-parametric methods based on conditional covariances. Under the scope of this research, information about PA, MAP, DETECT, Optimal Coordinate (n_{OC}), and Acceleration Factor (n_{AF}) methods have been provided in detail.

Parallel Analysis (PA) determines the number of factors by generating a random variable for sample size N, and a p variable (Horn, 1965). In PA, eigenvalues obtained from the data set are compared with eigenvalues obtained from independent normal variables. Data sets, as well as the number of variables produced in PA are the same size as the researched data set. The number of factors in PA is decided by comparing the average eigenvalue obtained from the independent variables and eigenvalues obtained from the data set. The number of factors was proposed if the eigenvalue obtained from the data set was bigger than the mean of those obtained from the random uncorrelated data. (Ledesma & Valero-Mora, 2007). In PA, instead of this average eigenvalue, the median of an eigenvalue, or the 5th, 95th, or 99th percentile of an eigenvalue can be used (Cota, Longman, Holden, Fekken, & Xinaris, 1993; Glorfeld, 1995; Raîche et al., 2013). As Cota et al. (1993) proposed the use of the 95th percentile, in this research the 95th percentile was adopted.

The Minimum Average Partial (MAP) test was developed by Velicer (1976). The MAP test is based on principle component analysis (PCA). In an MAP test, after the PCA, a partial correlation matrix is formed. At first stage, the first fundamental components are separated from the correlation matrix obtained from the variables in the data set. Squares of off diagonal elements in the correlation matrix are calculated, and a partial correlation matrix is formed. At the second stage, two fundamental components are separated from the correlation matrix obtained from the variables in the data set. Squares of off diagonal elements in the correlation matrix are calculated, and, again, a partial correlation matrix is formed. This process is applied until one minus number of variables is attained. Average squared partial correlations obtained from these steps are ranged. The number of dimensions is defined as the number of steps necessary to analyze the smallest average squared partial correlation (Ledesma & Valero-Mora, 2007; O'connor, 2000; Velicer, 1976). Due to calculation, it is stated that, for each factor, the factor loading of at least two variables should be high (Zwick & Velicer, 1986).

The DETECT method is based on finding a positive conditional covariance between items that measure the same dimension and a negative conditional covariance between items that measure different dimensions (Zhang & Stout, 1999). This method can be applied with both a confirmatory and exploratory mode. With the confirmatory mode, the DETECT index is calculated over user-defined sections. With the exploratory mode, the DETECT method searches for partitions that will make the DETECT index maximum. The index calculated as a result of this method provides information regarding the multidimensionality of the test (Jang & Roussos, 2007; Kim, 1996). If the DETECT index is higher than 1.00, it can be interpreted as exhibiting a strong multidimensionality; if it is between 0.40 - 1.00, it can be interpreted as exhibiting a moderate multidimensionality; if it is between 0.20 -0.40, it can be interpreted as exhibiting a weak multidimensionality, and if it is smaller than 0.20, it can be interpreted as exhibiting essential unidimensionality (Jang & Roussos, 2007; Zhang, 2007). In this research, the criteria selected was thus: smaller than 0.20 for one dimensional constructs, and larger than 1 for two dimensional constructs. For one dimensional constructs, if the DETECT index value is smaller than 0.20, it is accepted as one dimensional; for a DETECT index higher than 0.20, the construct was evaluated as multi-dimensional. For two dimensional constructs, values larger than 1 were considered. Thus, this research followed a more conservative approach.

In the Scree Plot method, Cattell (1966) combines a previous eigenvalue coordinate and the next eigenvalue coordinate with a trace line to determine the location of the screen. Thus, a point with immediate change is determined as an important number of a factor. In the Scree Test Optimal Coordinate (n_{OC}) method, without any statistical test, the focal point is the elbow. For that, a two-point regression model is used. In this method, dimension size is decided based on if estimated eigenvalues from the regression model and observed eigenvalues are equal or higher. For a comparison of values (between observed and estimated eigenvalues), the K1 rules is adopted and eigenvalues higher than 1 are compared. If desired, the average eigenvalues obtained from the PA result or eigenvalue at 0.05, 0.95 quantile can be used (Raîche et al., 2013). This method can be expressed as:

$$n_{OC} = \sum_{i} I\,[(\lambda_i \geq 1)\;\&\;(\lambda_i \geq \hat{\lambda}_i] \qquad (1)$$

Here, λ_i represents observed eigenvalue, and $\hat{\lambda}_i$ represents estimated eigenvalue. $\hat{\lambda}_i$ Additionally, is named as the optimal coordinate (Raîche et al., 2013).

In the Scree Test Acceleration Factor (n_{AF}) method, the acceleration factor represents the point on the coordinate where the slope of the curve changes abruptly. The value presented in Eq.1 and represented as $\hat{\lambda}_i$ is calculated by taking a second derivative of the optimal coordinate. For more detailed information, Raîche et al., (2013) can be read. This method is expressed as:

$$n_{AF} = \sum_{i} I\,(\lambda_i \geq 1\;\&\; i < k)\; with\; k \equiv argmax_j\,(af_j) \qquad (2)$$

Here, $argmax_j\,(af_j)$ represents the maximum point of the second derivative of an optimal coordinate function (Raîche et al., 2013).

When the literature was reviewed, various studies regarding Parallel Analysis were identified (Buja & Eyuboglu, 1992; Cho, Li, & Bandalos, 2009; Cota et al., 1993; Dinno, 2014; Glorfeld, 1995; Green, Levy, Thompson, Lu, & Lo, 2012; Green, Redell, Thompson, & Levy, 2016; Green, Thompson, Levy, & Lo, 2015; Kaya Kalkan & Kelecioğlu, 2016; Raîche et al., 2013; Weng & Cheng, 2005; Xia, Green, Xu, & Thompson, 2019; Yang & Xia, 2015; Zwick & Velicer, 1986). It was observed that these studies mostly focused on parallel analysis processes rather than comparison of performance of factor retention methods. In addition, most of these studies were conducted on continuous or ordinal data. There was limited researches with binary data (Kaya Kalkan & Kelecioğlu, 2016; Weng & Cheng, 2005). Current study focused to binary data which mostly used to educational and social researches. It can be said that the current study will contribute to the literature in this respect. In addition, there were studies about DIMTEST; however, in these studies, DIMTEST was often compared with NOHARM (Finch & Habing, 2005, 2007). But current study compares five factor retention methods (MAP, PA, optimal coordinate, DETECT and acceleration factor). Additionally, there were no studies comparing the methods adopted in the current study. For example Garrido, Abad, and Ponsoda (2011) examine only MAP procedure under different simulation conditions. But current study compares different factor retention methods under simulation conditions which frequently encountered in educational research. The current study differs from the others in the literature in terms of comparison of previously not compared factor retention methods. In other words, current study is important in terms of comparing methods not previously examined. Therefore, it is believed that the current research will contribute to the literature and the EFA that is commonly used in empirical studies (Conway & Huffcutt, 2003; Henson & Roberts, 2006) will be instructive to determine the number of factors.

METHODOLOGY

As predictions can be made by considering probabilities (Gilbert, 1999), this research was designed as a Monte Carlo simulation study. Simulation studies provide advantages, as such studies enable a comparison between real parameters and estimated parameters (Feinberg & Rubright, 2016). This is because real parameters are unknown in empirical studies. Therefore, analyses are carried out on data sets of known real parameters and the real and estimated parameters are compared.

Simulation Conditions

In this research, the simulations factors were identified as sample size, number of factors, test length, average factor loading, and type of correlation matrix. A full crossed pattern was used in this study. Since the focal point of this research was the tests that achieved a score of 1-0, binary data set were generated.

For the sample size factor, three conditions such as 250, 1000, and 3000 were determined. For the sample size conditions, a sample size of 250 was selected, as this size is both applicable and commonly used in studies (Beauducel & Herzberg, 2006; Hu, Bentler, & Kano, 1992). When 1000 people are selected as the sample size, the aim is to satisfy the sample size proposed in different studies for factor analysis (Comrey & Lee, 1992; Floyd & Widaman, 1995; Gorsuch, 1974; Guadagnoli & Velicer, 1988; Streiner, 1994). Generally, since estimated parameters rarely change after a sample size of 3000 (DeMars, 2010), this sample size was added to the research. Additionally, the effect of a sample size of more than 1000 people was analyzed to determine the number of factors acting on sample factor retention methods.

For number of factors, two conditions were included in this research. Here, the fact that achievement tests generally consisted of one dimension was considered (Bennett et al., 1990; Bennett, Rock, & Wang, 1991; Lissitz, Hou, & Slater, 2012; van den Bergh, 1990), and one-dimensional constructs were included in the research. However, it is impossible to analyze the performance of factor retention methods through the use of two-dimensional constructs. In addition, since it is possible that a method that operates with full accuracy in a unidimensional construct may fail to show the expected performance in a two dimensional one, two dimensional constructs were also included in this research. Yet, the relationship between these two dimensions was set around 0. Thus, the number of factor conditions was investigated as one and two (simple).

For the test length factor, 20 and 30 item conditions were determined, respectively. Turkish, mathematics and science tests consist of 20 items are used in the central examinations conducted by the

Ministry of Education in Turkey (MEB, 2019). Since it is rare for an achievement test with more than 30 items in the focal point, the research was limited to a 20 and 30 items condition.

The average factor loading was determined as 0.50 and 0.70. An average factor loading value below 0.50 is relatively rare for binary tests. This is because the lowest factor loading is proposed as 0.33 (Tabachnik & Fidell, 2012). When the factor loadings of other items were considered, a factor loading of 0.50 was added to the simulation conditions as the lower limit. Through an average factor loading of 0.70, the effect of a high relation between items and the factor on the factor retention method was evaluated. Attaining the same loading for all items was disregarded as a goal in this research. This is because the same factor loading for all items is rare. Therefore, the factor loadings of all items could differentiate, with only the average factor loading being kept within the desired conditions (e.g. for a 20-item test, the total factor loading was found, and an average factor loading was subsequently calculated by dividing by 20).

In this research, analyses were conducted for methods using correlation matrices (MAP and PA), including both the Pearson Product Moment (PPM) correlation matrix and the tetrachoric correlation matrix. Thus, the means by which such methods produce results according to correlation matrices was examined. Via this approach, the over factoring or under factoring effects of correlation matrices on methods were analyzed. In Table 1, the simulation factors and conditions are presented.

Table 1.
Simulation factor and conditions

Constant factors	Changing Factors				
Data Type	Number of factors	Sample Size	Test Length	Average Factor Loading	Correlation Matrix
1-0	1	250	20	0.50	PPM
	2 (simple)	1000 3000	30	0.70	Tetrachoric

As seen from Table 1, a fully crossed pattern was used in this research. Accordingly, a total of 2x3x2x2=24 conditions were evaluated: namely, the number of factors (2 conditions), sample size (3 conditions), test length (2 conditions), and average factor loading (2 conditions). For this purpose, a 24.000 data set was generated with 1.000 replications. Harwell, Stone, Hsu, and Kirisci (1996) proposed at least 25 replications. Robey and Barcikowski (1992) propose a formula for calculating the number of replications. The number of replications was determined to be 575 from the table created in accordance with this formula for research conditions. In order to increase the power of the research, 1.000 replicates were produced for this research. Both PPM and tetrachoric correlation matrix analysis on methods using correlation matrices were conducted for the same data set. However, no additional data set was produced for these conditions.

Data Generation and Analysis

Binary (1-0) data was produced with Psych (Revelle, 2016) package in R program (R Core Team, 2018). For each simulation condition, 1.000 replications were applied. For this purpose, the first 1.000 seed values were randomly generated, and, for these seed values, data was generated for the simulation conditions.

The analyses of the generated data were analyzed using different R packages. For parallel analysis and MAP analysis, Psych (Revelle, 2016) package was used; for the DETECT method, sirt (Robitzsch, 2017) package was used; for the Acceleration Factor (n_{AF}) and Optimal Coordinate (n_{OC}) methods, nFactors (Raiche, 2010) package was used.

For Parallel Analysis (PA), principal factor solution was used as factoring method. The number of randomly generated correlation matrices in PA was determined as 50. Since the cut-off score in the

DETECT method was suggested as 0.20 (Jang & Roussos, 2007; Zhang, 2007), in this research, DETECT values higher than 0.20 for one dimensional data sets, and DETECT value higher than 1 for two dimensions were accepted as multi-dimensional. Thus, the test evaluation followed a more conservative approach. To evaluate performance of methods, both a real and proposed number of dimension were compared and the percentage of correct estimates (PCE) were obtained. For this purpose:

$$P_r = \begin{cases} 1\ if\ \ Suggested\ Number\ of\ Dimensions\ = \ Actual\ Number\ of\ Dimensions \\ 0\ if\ \ Suggested\ Number\ of\ Dimensions\ \neq \ Actual\ Number\ of\ Dimensions \end{cases} \tag{3}$$

function was used. In this function, r represented replication. Accordingly, the calculated percentage of correct estimates (PCE) can be given with;

$$Percentage\ of\ Correct\ Estimates\ = \frac{\sum_{r=1}^{1000} P_r}{1000} .100 \tag{4}$$

equation. This way, the percentage replication of each method's production of correct results was determined.

Additionally, the mean difference (MD) between the real and estimated number of factors was calculated. For this purpose:

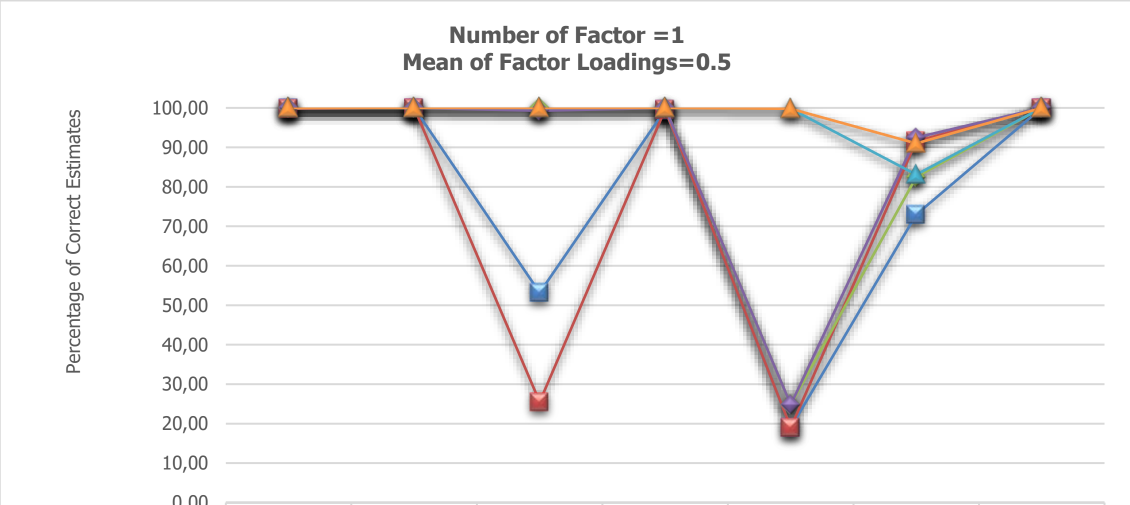
$$Mean\ Difference\ = \frac{\sum_{r=1}^{1000} (\widehat{m} - m)}{1000} \tag{5}$$

equation was used. Here, $\widehat{m}$ proposed number of dimensions reflected the m real number of dimension. Since the number of replications was 1.000, the average was calculated, thus determining whether the methods overestimated or underestimated the number of factors compared to PCE values.

FINDINGS

In this section, the findings obtained from simulation conditions are presented.

Comparison of Simulation Conditions for One Factor Constructs

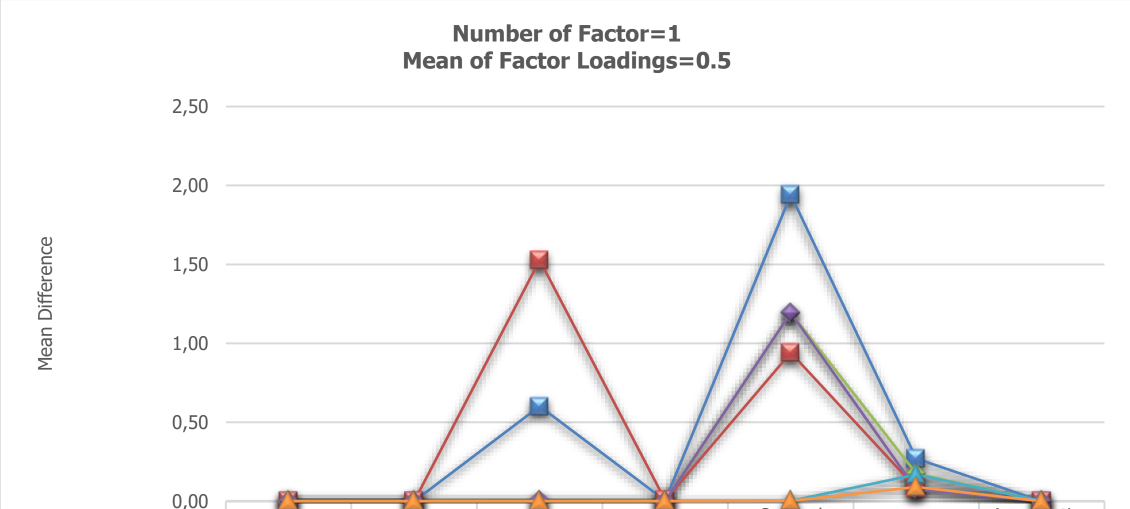


	MAP (T)	MAP (PPM)	PA (T)	PA (PPM)	Optimal Coordinate	DETECT	Accelaration Factor
250.20.1.05	100,00	100,00	53,20	99,50	19,00	73,00	100,00
250.30.1.05	99,80	100,00	25,50	99,60	19,00	91,60	100,00
1000.20.1.05	100,00	100,00	100,00	100,00	24,80	82,50	100,00
1000.30.1.05	100,00	100,00	99,30	100,00	24,80	92,50	100,00
3000.20.1.05	100,00	100,00	100,00	100,00	99,80	83,10	100,00
3000.30.1.05	100,00	100,00	100,00	100,00	99,80	91,00	100,00

Note: Coding System: First digit until the dot represented sample size, second digit represented number of item in the test, third digit represented number of factor, and fourth digit represented factor loading. (T): Tetrachoric correlation matrix, (PPM): Pearson Product Moment Correlation Matrix

Figure 1. Comparison of PCE values for simulation conditions when factor loading is 0.5 for one factor constructs

When Figure 1 was analyzed, the PCE values obtained through the factor retention method were observed. For an average factor loading of 0.5, when the results obtained for the sample size and the number of items method were considered, under MAP analysis with tetrachoric correlation matrix conditions (except for the sample size of 250, the number of items was 30 (99.8%)), all the conditions had 100% success rate. When MAP analysis was conducted with the PPM correlation matrix, the success rate was observed as 100%. However, when PA was conducted with the tetrachoric correlation matrix, it was impacted by the sample size. In case of the sample size of 250, PA had a significantly low percentage, while PCE had a rate of almost 100% for the sample sizes of 1000 and 3000. When PA was conducted with a PPM correlation matrix, a PCE value of almost 100% was attained for all sample sizes and numbers of items. When the Optimal Coordinate (n_{OC}) method was investigated, for the sample sizes of 250 and 1.000, it attained significantly low percentages; while, for the sample size of 3.000, the PCE value was almost 100%. In the DETECT method, as the number of items increased, it can be stated that PCE value also increased. For conditions with 30 items, the DETECT method had a PCE value of approximately 90%; whereas for conditions with 20 items, the PCE value was between 73-83.10%. Sample size had an effect when the number of items was higher than 20. As the sample size increased, the PCE value increased. It was observed that the Acceleration Factor (n_{AF}) method had a PCE value of 100% for all one factor constructs. Under simulation conditions with an average factor loading of 0.5 for one factor constructs, the MD value comparison is presented in Figure 2.



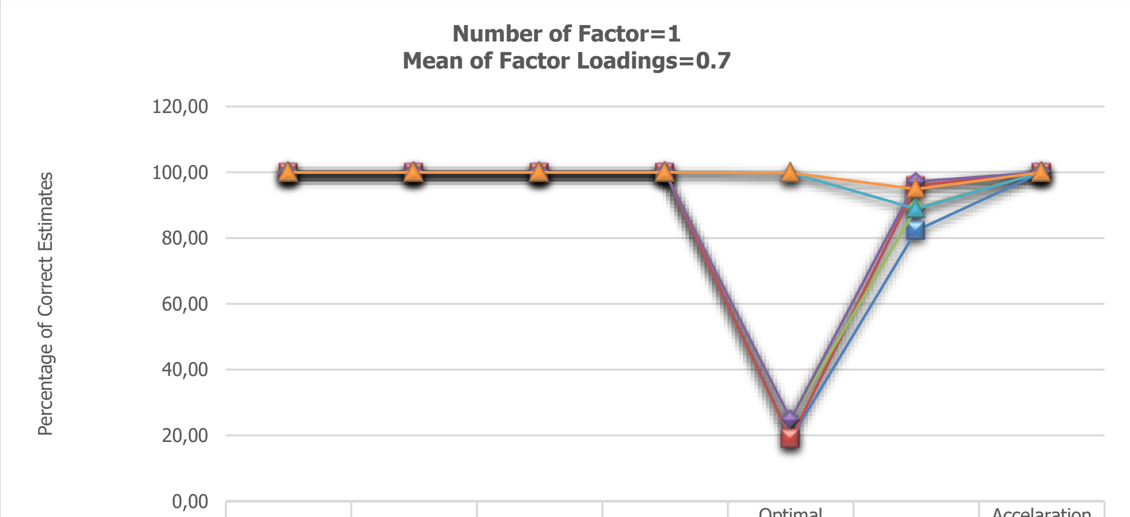
	MAP (T)	MAP (PPM)	PA (T)	PA (PPM)	Optimal Coordinate	DETECT	Acceleration Factor
250.20.1.05	0,00	0,00	0,60	0,00	1,94	0,27	0,00
250.30.1.05	0,00	0,00	1,53	0,01	0,94	0,08	0,00
1000.20.1.05	0,00	0,00	0,00	0,00	1,19	0,18	0,00
1000.30.1.05	0,00	0,00	0,01	0,00	1,19	0,08	0,00
3000.20.1.05	0,00	0,00	0,00	0,00	0,00	0,17	0,00
3000.30.1.05	0,00	0,00	0,00	0,00	0,00	0,09	0,00

Note: Coding System: First digit until the dot represented sample size, second digit represented number of item in the test, third digit represented number of factor, and fourth digit represented factor loading. (T): Tetrachoric correlation matrix, (PPM): Pearson Product Moment Correlation Matrix

Figure 2. Comparison of MD values for simulation conditions when factor loading is 0.5 for one factor constructs

In Figure 2, when methods with PCE values other than 100% were analyzed, the bias value for MAP analysis was 0. When PCE values were investigated (the tetrachoric correlation matrix of PCE value of the MAP analysis), it was 99.8% for 250.30.1.05 condition. Accordingly, in 2 of 1.000 replications, it can be stated that the number of factors obtained was other than 1. In MD value, this situation caused differentiation in the third digit after the comma. When a tetrachoric correlation matrix was conducted on the PA 250 sample size, positive MD values were observed. Accordingly, when a tetrachoric correlation matrix was used, it can be said that the PA had the tendency to produce more factors. However, results of PA conducted with PPM indicated that the MD values were close to 0. When the MD values for Optimal Coordinate (n_{OC}) method was investigated, it can be stated that the bias decreased as the sample size increased; yet, the biased results were positive, which meant that the number of factors was overestimated. Additionally, the Optimal Coordinate (n_{OC}) method produced an unbiased estimation for sample size of 3.000. When the MD values for the DETECT method were investigated, it can be stated that the number of items was more effective on MD values. As in all methods, in the DETECT method, the number of factors was overestimated. In the Acceleration Factor (n_{AF}) method, since a PCE value of 100% was obtained for all single factor constructs, the MD value was zero.

In Figure 3, when the average factor loading was 0.7, PCE values were compared for one factor construct simulation conditions.

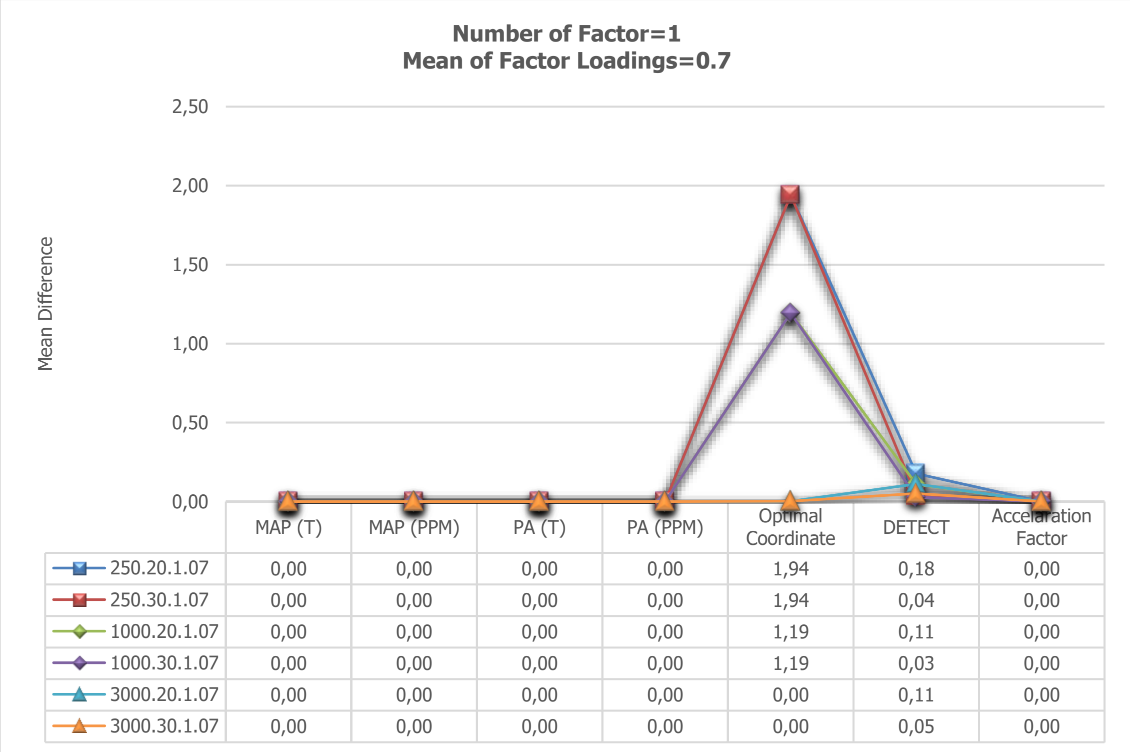


	MAP (T)	MAP (PPM)	PA (T)	PA (PPM)	Optimal Coordinate	DETECT	Accelaration Factor
250.20.1.07	99,90	100,00	100,00	100,00	19,00	82,30	100,00
250.30.1.07	100,00	100,00	100,00	100,00	19,00	96,00	100,00
1000.20.1.07	100,00	100,00	100,00	100,00	24,80	89,10	100,00
1000.30.1.07	100,00	100,00	100,00	100,00	24,80	97,20	100,00
3000.20.1.07	100,00	100,00	100,00	100,00	99,80	88,90	100,00
3000.30.1.07	100,00	100,00	100,00	100,00	99,80	94,90	100,00

Note: Coding System: First digit until the dot represented sample size, second digit represented number of item in the test, third digit represented number of factor, and fourth digit represented factor loading. (T): Tetrachoric correlation matrix, (PPM): Pearson Product Moment Correlation Matrix

Figure 3. Comparison of PCE values for simulation conditions when factor loading is 0.7 for one factor constructs

When Figure 3 was investigated, compared to an average factor loading of 0.5, in methods with the condition of an average factor loading of 0.7, it was observed that PCE values increased. While all conditions except one (250.20.1.07) had PCE values of 100%, when MAP analysis was conducted with a tetrachoric correlation matrix and a PPM correlation matrix, a PCE value of 100% was obtained for all conditions. For this condition cluster, the PCE value was observed as 100% for PA results conducted with both a tetrachoric and PPM correlation matrix. When conditions of an average factor loading of 0.5 and 0.7 were analyzed together (Figure 1 and Figure 3), it can be stated that Optimal Coordinate (n_{OC}) exhibits no differentiation for average factor loading. Under an average factor loading of 0.5 and the condition of a sample size of 3000, the PCE value was around 100%; whereas, with sample sizes of 250 and 1000, respectively, the PCE value was significantly low. In the DETECT method, when the average factor loadings increased, the PCE also increased. However, upon comparing Figure 1 and Figure 3, it can be stated that, as the number of items increased, the DETECT has become more effective at producing correct results. The PCE value was 94% when the DETECT method was most successful. In the Acceleration Factor (n_{AF}) method, the PCE value was observed as 100%. Under simulation conditions with average factor loading of 0.7 for one factor constructs, the MD value comparison is presented in Figure 4.



	MAP (T)	MAP (PPM)	PA (T)	PA (PPM)	Optimal Coordinate	DETECT	Acceleration Factor
250.20.1.07	0,00	0,00	0,00	0,00	1,94	0,18	0,00
250.30.1.07	0,00	0,00	0,00	0,00	1,94	0,04	0,00
1000.20.1.07	0,00	0,00	0,00	0,00	1,19	0,11	0,00
1000.30.1.07	0,00	0,00	0,00	0,00	1,19	0,03	0,00
3000.20.1.07	0,00	0,00	0,00	0,00	0,00	0,11	0,00
3000.30.1.07	0,00	0,00	0,00	0,00	0,00	0,05	0,00

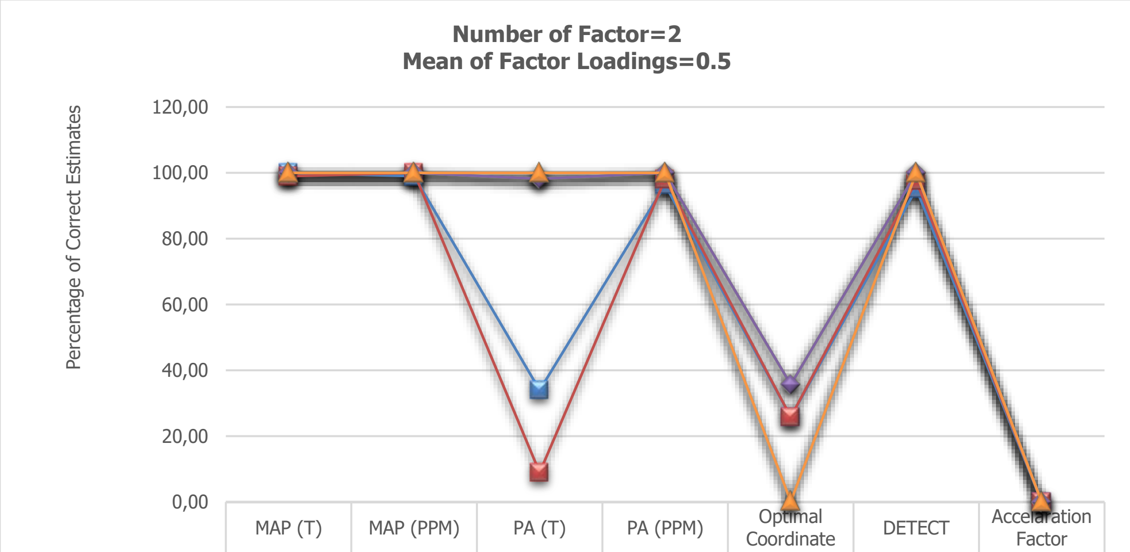
Note: Coding System: First digit until the dot represented sample size, second digit represented number of item in the test, third digit represented number of factor, and fourth digit represented factor loading. (T): Tetrachoric correlation matrix, (PPM): Pearson Product Moment Correlation Matrix

Figure 4. Comparison of MD values for simulation conditions when factor loading is 0.7 for one factor constructs

In Figure 4, when methods exhibiting PCE values other than 100% were analyzed, and since the MAP analysis only estimated the number of factors different to 1 in 1 of 1.000 replications, the PCE value was obtained as 99.90%. Therefore, the MD value differentiated in the third digit after the comma. Thus, it can be stated that the Optimal Coordinate (n_{OC}) method, as with the DETECT method, has a tendency to overestimate the number of factors. However, in the Optimal Coordinate (n_{OC}) method, while MD values decreased as the sample size increased, in the DETECT method, an increased number of item caused a rapid decrease in MD values. Additionally, the Optimal Coordinate (n_{OC}) method produced unbiased results for the sample size of 3.000.

Comparison of Simulation Conditions for Two Factors Constructs

Under simulation conditions with average factor loading of 0.5 for two factors constructs, the PCE value comparison is presented in Figure 5.

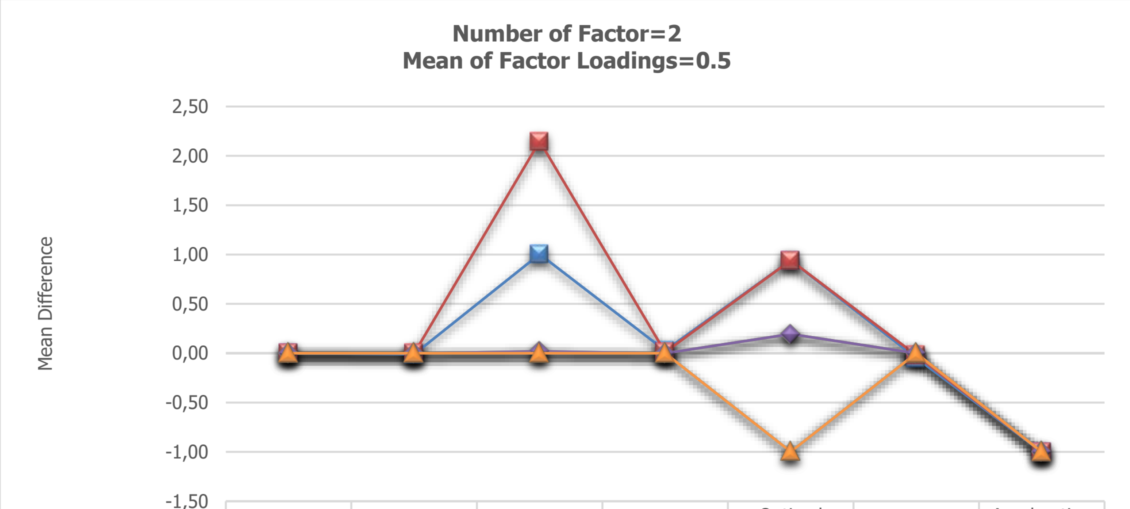


	MAP (T)	MAP (PPM)	PA (T)	PA (PPM)	Optimal Coordinate	DETECT	Acceleration Factor
250.20.2.05	100,00	99,00	34,00	96,40	25,80	95,20	0,00
250.30.2.05	99,00	100,00	9,00	98,10	25,80	97,60	0,00
1000.20.2.05	100,00	100,00	98,20	100,00	35,70	100,00	0,00
1000.30.2.05	100,00	100,00	98,20	100,00	35,70	100,00	0,00
3000.20.2.05	100,00	100,00	100,00	100,00	0,20	100,00	0,00
3000.30.2.05	100,00	100,00	100,00	100,00	0,20	100,00	0,00

Note: Coding System: First digit until the dot represented sample size, second digit represented number of item in the test, third digit represented number of factor, and fourth digit represented factor loading. (T): Tetrachoric correlation matrix, (PPM): Pearson Product Moment Correlation Matrix

Figure 5. Comparison of PCE values for simulation conditions when factor loading is 0.5 for two factors constructs

When Figure 5 was analyzed, simulation conditions with an average factor loading of 0.5 were compared in two factors constructs. When MAP analysis was conducted with a tetrachoric correlation matrix, the PCE value was around 100%. There was 99.0% PCE value in only one condition (250.30.2.05). When MAP analysis was conducted for the PPM correlation matrix, it was observed that the PCE value was 100%, excepting one condition (250.20.2.05). Accordingly, in the MAP analysis with the sample size of 250, it can be stated that the correct estimation was made with an error margin of 1%. When PA was conducted with a tetrachoric correlation matrix, for the sample sizes of 30 and 250 items, it had a PCE value of 9%. As the size of the sample increased, the PA conducted on a tetrachoric correlation matrix had a tendency to produced more accurate estimations. PA conducted for the PPM correlation matrix indicated a PCE value of around 100% for the 250 sample size, and 100% for the 1.000 and 3.000 sample sizes. For the simulation condition with a sample size of 250 and 20-item, PA results conducted with PPM showed a 96.4% PCE value. For the same sample size, as the number of items increased, the PCE value of the PA increased as well. In the Optimal Coordinate (n_{OC}) method, it can be stated that there was no differentiation for both average factor loading and number of items. This is valid for both one factor (Figure 1 and Figure 3) and two factors (Figure 5 and Figure 7) constructs. contrary to one dimensional constructs, while PCE value of 0% was exhibited in the 3.000 sample size, PCE values of 25% and 35% were obtained for the 250 and 1.000 sample size, respectively. It can be stated that the DETECT method has higher PCE values for two dimensional constructs. When the sample size was 1.000 or more, the DETECT method had a PCE value of 100% for all conditions with 2 dimensions and an average factor loading of 0.5. In the Acceleration Factor (n_{AF}) method, the PCE value was 0% for all conditions. When the dimension number suggested by this method was analyzed, it was observed that there was a one-dimension construct for all conditions. Under simulation conditions with an average factor loading of 0.5 for two factors constructs, the MD value comparison is presented in Figure 6.

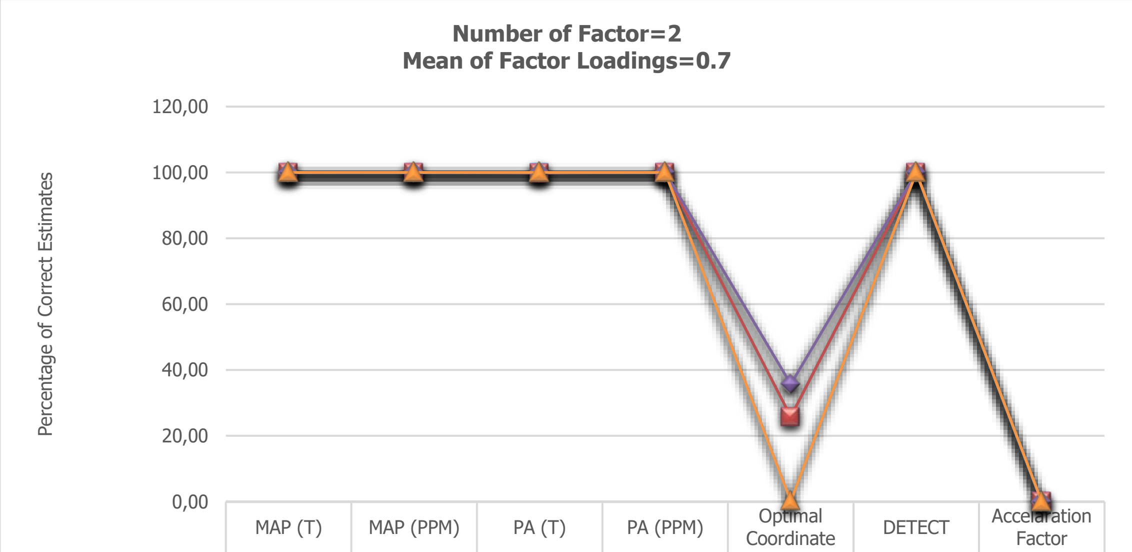


	MAP (T)	MAP (PPM)	PA (T)	PA (PPM)	Optimal Coordinate	DETECT	Accelaration Factor
250.20.2.05	0,00	-0,01	1,00	0,03	0,94	-0,05	-1,00
250.30.2.05	0,00	0,00	2,15	0,01	0,94	-0,02	-1,00
1000.20.2.05	0,00	0,00	0,02	0,00	0,19	0,00	-1,00
1000.30.2.05	0,00	0,00	0,02	0,00	0,19	0,00	-1,00
3000.20.2.05	0,00	0,00	0,00	0,00	-1,00	0,00	-1,00
3000.30.2.05	0,00	0,00	0,00	0,00	-1,00	0,00	-1,00

Note: Coding System: First digit until the dot represented sample size, second digit represented number of item in the test, third digit represented number of factor, and fourth digit represented factor loading. (T): Tetrachoric correlation matrix, (PPM): Pearson Product Moment Correlation Matrix

Figure 6. Comparison of MD values for simulation conditions when factor loading is 0.5 for two factors constructs

When the MD values presented in Figure 6 were analyzed, in MAP analysis conducted with PPM, the number of factors was underestimated for only one condition (250.20.2.05). For the sample size of 250, the results for PA conducted with a tetrachoric correlation matrix indicated that in a 30-item condition the PA had significantly high MD value. This is because PA estimated 2.15 times more factors than real number of factor. However, there was no such case when the analysis was conducted with PPM. In the Optimal Coordinate (n_{OC}) method with a sample size of 250 and 1.000, the number of factors was overestimated; however, for the 3.000 sample size, the number of factors was underestimated. While the DETECT method had negative MD values for the sample size of 250, it produced unbiased results for the other samples. The acceleration Factor (n_{AF}) method only proposed one factor for all conditions. Therefore, the MD value was -1. The PCE value analysis under two-dimensional construct simulation conditions with average factor loading of 0.7 is given in Figure 7.

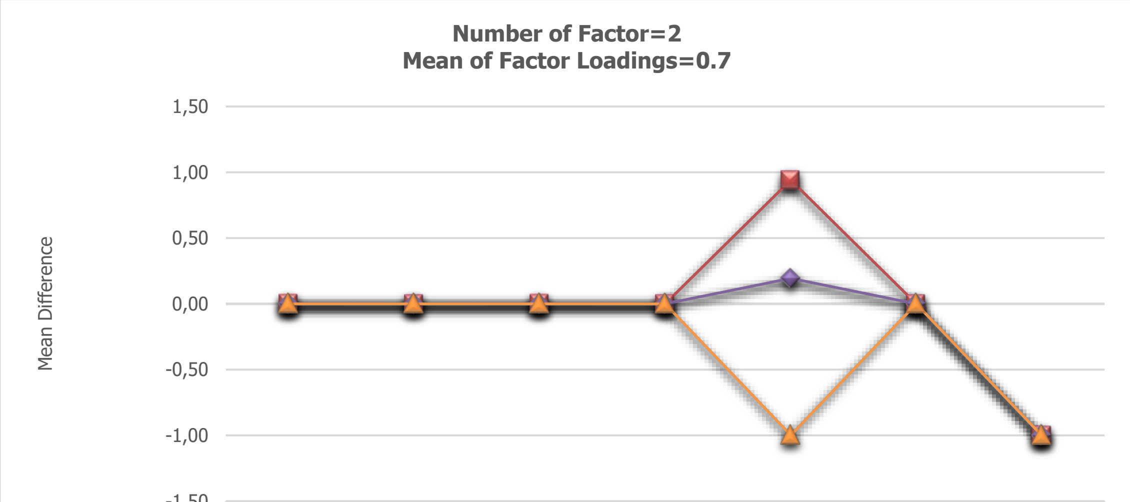


	MAP (T)	MAP (PPM)	PA (T)	PA (PPM)	Optimal Coordinate	DETECT	Acceleration Factor
250.20.2.07	100,00	100,00	100,00	100,00	25,80	100,00	0,00
250.30.2.07	99,90	100,00	99,80	100,00	25,80	100,00	0,00
1000.20.2.07	100,00	100,00	100,00	100,00	35,70	100,00	0,00
1000.30.2.07	100,00	100,00	100,00	100,00	35,70	100,00	0,00
3000.20.2.07	100,00	100,00	100,00	100,00	0,20	100,00	0,00
3000.30.2.07	100,00	100,00	100,00	100,00	0,20	100,00	0,00

Note: Coding System: First digit until the dot represented sample size, second digit represented number of item in the test, third digit represented number of factor, and fourth digit represented factor loading. (T): Tetrachoric correlation matrix, (PPM): Pearson Product Moment Correlation Matrix

Figure 7. Comparison of PCE values for simulation conditions when factor loading is 0.7 for two factors constructs

Figure 7 presented the PCE values obtained for conditions with two factors and an average factor loading of 0.7. When PA and MAP analysis were conducted with a tetrachoric correlation matrix, all conditions except for one (250.30.2.07) had a PCE value of 100%. When PA and MAP analysis were conducted with a PPM correlation matrix, all conditions had a PCE value of 100%. When both Figure 1 and 3 and Figure 5 and 7 were analyzed, it can be expressed that the Optimal Coordinate (n_{OC}) exhibited no differentiation for both average factor loading and number of items. The same results were observed for conditions with average factor loading of 0.5. contrary to one dimensional constructs, while two dimensional constructs had 0% PCE value for the 3.000 sample size, they had 25% and 35% PCE values for the 250 and 1000 sample sizes, respectively. In the Optimal Coordinate (n_{OC}) method, it can be stated that only sample size influenced two factor constructs. In the DETECT method, when average factor loadings increased, the PCE value also increased. The DETECT method had 100% PCE for this condition set. In the Acceleration Factor (n_{AF}) method, the PCE value was 0. Accordingly, it can be stated that dimensionality for any data set was accurately estimated. MD value analysis under two-dimensional construct simulation conditions with an average factor loading of 0.7 is given in Figure 8.



	MAP (T)	MAP (PPM)	PA (T)	PA (PPM)	Optimal Coordinates	DETECT	Accelaration Factor
250.20.2.07	0,00	0,00	0,00	0,00	0,94	0,00	-1,00
250.30.2.07	0,00	0,00	0,00	0,00	0,94	0,00	-1,00
1000.20.2.07	0,00	0,00	0,00	0,00	0,19	0,00	-1,00
1000.30.2.07	0,00	0,00	0,00	0,00	0,19	0,00	-1,00
3000.20.2.07	0,00	0,00	0,00	0,00	-1,00	0,00	-1,00
3000.30.2.07	0,00	0,00	0,00	0,00	-1,00	0,00	-1,00

Note: Coding System: First digit until the dot represented sample size, second digit represented number of item in the test, third digit represented number of factor, and fourth digit represented factor loading. (T): Tetrachoric correlation matrix, (PPM): Pearson Product Moment Correlation Matrix

Figure 8. Comparison of MD values for simulation conditions when factor loading is 0.7 for two factors constructs

When the MD values presented in Figure 8 were analyzed, for the Optimal Coordinate (n_{OC}) method samples of 250 and 1.000, there was a positive MD value; whereas, for the 3.000 sample size, there was a negative MD value. The acceleration Factor (n_{AF}) method estimated all data set as one factor. Therefore, all MD values obtained as -1. MD values of both PA and MAP conducted with tetrachoric and PPM correlation matrices were 0. Accordingly, it can be stated that PA and MAP create an unbiased estimation for this correlation matrix.

DISCUSSION and CONCLUSION

When factor retention methods were compared, the results of this research indicated that, when MAP analysis was conducted on both tetrachoric and PPM correlation matrices, PCE values of 99% or more were observed for all conditions. Accordingly, it can be stated that MAP analysis may be used to determine dimensionality. Garrido, Abad, and Ponsoda (2011) stated that, for MAP analysis results conducted with (2, 3, 4, 5, 6, 7) variables with different category numbers for polychoric and PPM correlation matrices, a polychoric correlation matrix was more appropriate. Additionally, as factor loading increased, PCE values increased in MAP analysis. In our research, MAP analysis was the best method for all conditions. In Garrido, Abad, and Ponsoda (2011), this could be caused by the fact that the factor loading of each variable was accepted as equal. In our research, the average factor loading was kept constant; however, the factor loading of each item was differentiated. Zwick and Velicer (1986) compared PA, MAP, Cattell's scree test, Bartlett's chi-square test, and the K1 rule methods and reported that the PA and MAP method presented the best results. In our current research, the results are consistent with the literature.

When PA was conducted with a tetrachoric correlation matrix, it was affected by average factor loading and sample size. In small samples with an average factor loading of 0.5, the performance of PA decreased. Yet, as the sample increased, the performance of PA increased as well. These results are similar to those of Yang and Xia (2015) and Cho et al. (2009). While sample size had a significant effect on PA conducted with a tetrachoric correlation matrix, the same effect is invisible with a PPM correlation matrix. Guilford (1952) stated that, to calculate a tetrachoric correlation matrix, data should have a large sample size (at least 400) for binary data. The low performance of PA with the sample size of 250 may be linked with this.

Optimal Coordinate (n_{OC}) showed a good performance for one dimensional data with a sample size of 3.000; however, the PCE value was below 50% for other conditions. This result is in line with Raîche et al. (2013). Raîche et al. (2013) worked with 36 and 72 variables, using 2 and 5 folds of variables for sample size (72, 180, 144, 360) and 0.5 and 0.8 average factor loadings. Results for the Optimal Coordinate (n_{OC}) method's PCE value varied between 20% (72 variables, 144 sample size, 0.5 factor loading) and 82% (72 variables, 360 sample size, 0.8 factor loading).

The acceleration Factor (n_{AF}) method estimated one dimension for all conditions under the scope of this research. In this case, it can be expressed that the Acceleration Factor (n_{AF}) method presented no differentiation for factors under all conditions. Raîche et al. (2013) determined that the Acceleration Factor (n_{AF}) method's PCE value varied between 17% and 50%. However, in this study, factor loadings were distributed over 3 components as 0.8 and 0.2. In our study, average factor loading was considered.

While the DETECT method had around 100% PCE value for the majority of two factor constructs, it was affected by both sample size and average factor loading for one factor constructs. Additionally, as the number of item increased, the DETECT method's performance increased as well. Similarly, as average factor loading increased, DETECT estimated more accurate results. van Abswoude, van der Ark, and Sijtsma (2004) stated that the DETECT method is affected by a sample's size and is more efficient for larger samples (n=2.000). Additionally, DETECT determined multi-dimensionality accurately in the case of low correlation between factors. This is in line with other studies in the literature (Zhang, Yu, & Nandakumar, 2003).

Based on the conditions of this research, the following recommendations can be made; 1) both a tetrachoric and PPM correlation matrix of MAP can be used, 2) the PA method produced more accurate results with a PPM correlation matrix , but in the case of a tetrachoric correlation matrix, samples size should be considered, 3) instead of Optimal Coordinate (n_{OC}) and Acceleration Factor (n_{AF}) methods, PA or MAP methods can be preferred, and 4) average factor loading and samples size should be considered before using the DETECT method. Additionally, for this method, both the number of items and dimension number has an effect. It can be stated that this method may be used after collectively evaluating all conditions. Additionally, the combined use and evaluation of PA and MAP can be recommended. Since only simple construct data was used in this research, in future studies factor size could be increased and simulations that manipulate the correlation between factors could be applied.

Acknowledgement

This study was presented in 6th. International Conference on Education, Zagreb, Croatia.

REFERENCES

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: A Multidisciplinary Journal*, *13*(2), 186–203. https://doi.org/10.1207/s15328007sem1302_2

Bennett, R. E., Rock, D. A., Braun, H. I., Frye, D., Spohrer, J. C., & Soloway, E. (1990). The relationship of expert-system scored constrained free-response items to multiple-choice and open-ended items. *Applied Psychological Measurement*, *14*(2), 151–162. https://doi.org/10.1177/014662169001400204

Bennett, R. E., Rock, D. A., & Wang, M. (1991). Equivalence of free-response and multiple-choice items. *Journal of Educational Measurement*, *28*(1), 77–92. https://doi.org/10.1111/j.1745-3984.1991.tb00345.x

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., & Lee, H. B. (1992). *A first course in factor analysis* (2nd ed.). London: Lawrence Erlbaum Associates.

Conway, J. M., & Huffcutt, A. I. (2003). A review and evaluation of exploratory factor analysis practices in organizational research. *Organizational Research Methods*, *6*(2), 147–168. https://doi.org/10.1177/1094428103251541

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

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

Dinno, A. (2014). *Gently clarifying the application of Horn's parallel analysis to principal component analysis versus factor analysis*. Unpublished manuscript, School of Community Health, Portland State University, Oregon, USA. Retrieved from http://doyenne.com/Software/files/ PA_for_PCA_vs_FA.pdf

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

Feinberg, R. A., & Rubright, J. D. (2016). Conducting simulation studies in psychometrics. *Educational Measurement: Issues and Practice*, *35*(2), 36–49. https://doi.org/10.1111/emip.12111

Finch, H., & Habing, B. (2005). Comparison of NOHARM and DETECT in item cluster recovery: Counting dimensions and allocating items. *Journal of Educational Measurement*, *42*(2), 149–169. https://doi.org/10.1111/j.1745-3984.2005.00008

Finch, H., & Habing, B. (2007). Performance of DIMTEST- and NOHARM- based statistics for testing unidimensionality. *Applied Psychological Measurement*, *31*(4), 292–307. https://doi.org/10.1177/0146621606294490

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

Fraser, C., & McDonald, R. P. (1988). NOHARM: Least squares item factor analysis. *Multivariate Behavioral Research*, *23*(2), 267–269. https://doi.org/10.1207/s15327906mbr2302_9

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

Gilbert, N. (1999). Simulation: A new way of doing social science. *American Behavioral Scientist*, *42*(10), 1485–1487. https://doi.org/10.1177/00027642990420100002

Glorfeld, L. W. (1995). An improvement on Horn's parallel analysis methodology for selecting the correct number of factors to retain. *Educational and Psychological Measurement*, *55*(3), 377–393. https://doi.org/10.1177/0013164495055003002

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

Green, S. B., Levy, R., Thompson, M. S., Lu, M., & Lo, W.-J. (2012). A proposed solution to the problem with using completely random data to assess the number of factors with parallel analysis. *Educational and Psychological Measurement*, *72*(3), 357–374. https://doi.org/10.1177/0013164411422252

Green, S. B., Redell, N., Thompson, M. S., & Levy, R. (2016). Accuracy of revised and traditional parallel analyses for assessing dimensionality with binary data. *Educational and Psychological Measurement*, *76*(1), 5–21. https://doi.org/10.1177/0013164415581898

Green, S. B., Thompson, M. S., Levy, R., & Lo, W.-J. (2015). *Type I and Type II error rates and overall accuracy of the revised parallel analysis method for determining the number of factors*. *Educational and Psychological Measurement* , *75*(3), 428-457. https://doi.org/10.1177/0013164414546566

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

Guilford, J. P. (1952). When not to factor analyze. *Psychological Bulletin*, *49*(1), 26–37. https://doi.org/10.1037/h0054935

Harwell, M., Stone, C. A., Hsu, T.-C., & Kirisci, L. (1996). Monte carlo studies in item response theory. *Applied Psychological Measurement*, *20*(2), 101–125. https://doi.org/10.1177/014662169602000201

Henson, R. K., & Roberts, J. K. (2006). Use of exploratory factor analysis in published research. *Educational and Psychological Measurement*, *66*(3), 393–416. https://doi.org/10.1177/0013164405282485
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
Hu, L., Bentler, P. M., & Kano, Y. (1992). Can test statistics in covariance structure analysis be trusted? *Psychological Bulletin*, *112*(2), 351–362. https://doi.org/10.1037/0033-2909.112.2.351
Jang, E. E., & Roussos, L. (2007). An investigation into the dimensionality of TOEFL using conditional covariance-based nonparametric approach. *Journal of Educational Measurement*, *44*(1), 1–21. https://doi.org/10.1111/j.1745-3984.2007.00024.x
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
Kaya Kalkan, Ö., & Kelecioğlu, H. (2016). The effect of sample size on parametric and nonparametric factor analytical methods. *Educational Sciences: Theory & Practice*, *16*(1), 153–171. https://doi.org/10.12738/estp.2016.1.0220
Kim, H. R. (1996). A new index of dimensionality-DETECT. *The Pure and Applied Mathematics*, *3*(2), 141–154.
Ledesma, R. D., & Valero-Mora, P. (2007). Determining the number of factors to retain in EFA: An easy-to-use computer program for carrying out parallel analysis. *Practical Assessment, Research & Evaluation*, *12*(2), 2–11.
Lissitz, R. W., Hou, X., & Slater, S. C. (2012). The contribution of constructed response items to large scale assessment: Measuring and understanding their impact. *Journal of Applied Testing Technology*, *13*(3), 1–50. Retrieved from http://www.jattjournal.com/index.php/atp/article/view/48366/39234
MEB. (2019). Sınavla öğrenci alacak ortaöğretim kurumlarına ilişkin merkezî sınav başvuru ve uygulama kılavuzu. Retrieved from https://www.meb.gov.tr/meb_iys_dosyalar/2019_04/03134315_Kilavuz2019.pdf
Nandakumar, R., & Stout, W. (1993). Refinements of Stout's procedure for assessing latent trait unidimensionality. *Journal of Educational Statistics*, *18*(1), 41-68. https://doi.org/10.2307/1165182
O'connor, B. P. (2000). SPSS and SAS programs for determining the number of components using parallel analysis and Velicer's MAP test. *Behavior Research Methods, Instruments, & Computers*, *32*(3), 396–402. https://doi.org/10.3758/BF03200807
Osborne, J. W., & Banjanovic, E. S. (2016). *Exploratory factor analysis with SAS®*. Cary, NC: SAS Intitute Inc.
R Core Team. (2018). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.r-project.org/.
Raiche, G. (2010). nFactors: An R package for parallel analysis and non graphical solutions to the Cattell scree test. Retrived from https://cran.r-project.org/web/packages/nFactors/nFactors.pdf.
Raîche, G., Walls, T. A., Magis, D., Riopel, M., & Blais, J.-G. (2013). Non-Graphical solutions for Cattell's scree test. *Methodology*, *9*(1), 23–29. https://doi.org/10.1027/1614-2241/a000051
Revelle, W. (2016). psych: Procedures for psychological, psychometric, and personality research. Evanston, Illinois. Retrieved from https://cran.r-project.org/package=psych
Robey, R. R., & Barcikowski, R. S. (1992). Type I error and the number of iterations in monte carlo studies of robustness. *British Journal of Mathematical and Statistical Psychology*, *45*, 283–288.
Robitzsch, A. (2017). sirt: Supplementary item response theory models. Retrieved from https://cran.r-project.org/package=sirt
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.). Boston: Pearson.
van Abswoude, A. A. H., van der Ark, L. A., & Sijtsma, K. (2004). A comparative study of test data dimensionality assessment procedures under nonparametric IRT models. *Applied Psychological Measurement*, *28*(1), 3–24. https://doi.org/10.1177/0146621603259277
van den Bergh, H. (1990). On the construct validity of multiple- choice items for reading comprehension. *Applied Psychological Measurement*, *14*(1), 1–12. https://doi.org/10.1177/014662169001400101
Velicer, W. F. (1976). Determining the number of components from the matrix of partial correlations. *Psychometrika*, *41*(3), 321–327.
Weng, L.-J., & Cheng, C.-P. (2005). Parallel analysis with unidimensional binary data. *Educational and Psychological Measurement*, *65*(5), 697–716. https://doi.org/10.1177/0013164404273941
Xia, Y., Green, S. B., Xu, Y., & Thompson, M. S. (2018). Proportion of indicator common variance due to a factor as an effect size statistic in revised parallel analysis. *Educational and Psychological Measurement*. 79(1), 75-107. https://doi.org/10.1177/0013164418754611
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

Zhang, J. (2007). Conditional covariance theory and DETECT for polytomous items. *Psychometrika*, *72*(1), 69–91. https://doi.org/10.1007/s11336-004-1257-7

Zhang, J., & Stout, W. (1999). The theoretical detect index of dimensionality and its application to approximate simple structure. *Psychometrika*, *64*(2), 213–249. https://doi.org/10.1007/BF02294536

Zhang, Y. O., Yu, F., & Nandakumar, R. (2003). *The impact of conditional scores on the performance of DETECT*. Paper presented at the annual meeting of the National Council on Measurement in Education. Chicago, IL. Retrieved from https://files.eric.ed.gov/fulltext/ED478170.pdf

Zoski, K., & Jurs, S. (1993). Using multiple regression to determine the number of factors to retain in factor analysis. *Multiple Linear Regression Viewpoint*, *20*(1), 5–9. Retrieved from http://www.glmj.org/archives/MLRV_1993_20_1.pdf

Zoski, K., & Jurs, S. (1996). An objective counterpart to the visual scree test for factor analysis: The standard error scree. *Educational and Psychological Measurement*, *56*(3), 443–451. https://doi.org/10.1177/0013164496056003006

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

TÜRKÇE GENİŞLETİLMİŞ ÖZET

Veri setinin yapısı araştırılmak istenildiğinde faktör analizi en sık kullanılan psikometrik tekniklerden biridir. Araştırmacıların yapıyla ilgili bilgileri yoksa bu durumda açımlayıcı faktör analizi ya da sınırlandırılmamış faktör analizi yapılır. Eğer yapı hakkında araştırmacıların fikri varsa ve verilerin bu yapıya uyum gösterip göstermediği araştırılacaksa bu durumda da doğrulayıcı faktör analizi diğer bir değişle sınırlandırılmış faktör analizi yapılmaktadır.

AFA için sıklıkla karşılaşılan güçlüklerden biri de boyut sayısına karar vermektir. Boyut sayısına karar vermek için geliştirilen birçok yöntem bulunmaktadır. Bu yöntemlere; Örneğin 1'den büyük özdeğer sayısını boyut sayısı kabul eden ve Kaieser'in K1 kuralı (Kaiser, 1960), Horn (1965) tarafından önerilen Paralel Analiz (PA), Velicer (1976) tarafından önerilen Kısmi Korelasyonların En Küçüğü (MAP) testi, Cattell (1966) tarafından önerilen Yamaç Grafiği yöntemleri de bulunmaktadır. Yamaç grafiğine alternatif olarak Zoski ve Jurs (1993, 1996) tarafından önerilen çok regresyon t-değeri ve yamacın standart hatası (SEscree) yaklaşımları, Raîche, Walls, Magis, Riopel ve Blais (2013) tarafından önerilen amaç grafiği optimal koordinatlar ve yamaç grafiği ivmelenme faktörü gibi yöntemler bulunmaktadır. Bu yöntemlerin yanında koşullu kovaryanslara dayanan ve nonparametrik bir yöntemler olan DIMTEST (Nandakumar & Stout, 1993; Stout, 1987) ve DETECT (Zhang & Stout, 1999) yöntemleri vardır. Mevcut araştırmada Paralel Analiz, Kısmi Korelasyonların En Küçüğü, DETECT, Optimal Koordinat ve İvmelenme Faktörü yöntemleri karşılaştırılmıştır.

İhtimalleri dikkate alınarak önerilerde bulunulabilmesi nedeniyle (Gilbert, 1999) bu araştırma, Monte Carlo simülasyon çalışması olarak tasarlanmıştır. Simülasyon çalışmaları gerçek parametreler ile kestirilen parametrelerin karşılaştırılmasını sağlaması nedeniyle avantaj sağlamaktadır (Feinberg & Rubright, 2016).

Araştırmada simülasyon faktörleri, örneklem büyüklüğü (250, 1000 ve 3000), faktör sayısı (tek ve iki [basit yapıda] boyutlu) test uzunluğu (20 ve 30 madde), ortalama faktör yükü (0.50 ve 0.70) ve kullanılan korelasyon matrisi (Pearson Momentler Çarpımı [PPM] ve tetrakorik) olarak belirlenmiştir. Araştırmada tamamen çaprazlanmış desen kullanılmıştır. Buna göre 2x3x2x2=24 koşul üzerinde çalışmış ve 1.000 replikasyon yapılmıştır. Araştırmada ikili (1-0) yapıdaki veri kullanılmıştır.

Verinin üretimi için R yazılımındaki Psych paketi kullanılmıştır. Paralel analiz ve MAP analizi için Psych (Revelle, 2016) paketi, DETECT yöntemi için sirt (Robitzsch, 2017), ivmelenme faktörü ve optimal koordinat yöntemi için de nFactors (Raiche, 2010) paketleri kullanılmıştır.

PA için faktörleştirme yöntemi olarak temel faktör çözümlemesi kullanılmıştır. PA'da rassal olarak oluşturulan korelasyon matrisi sayısı ise 50 olarak belirlenmiştir. DETECT yönteminde kesme puanı olarak 0.20 önerildiğinden (Jang & Roussos, 2007; Zhang, 2007) araştırmada tek boyutlu veri seti için 0.20 üzerinde DETECT değeri elde edilen analizlerde çok boyutlu, iki boyutlu veri seti için ise DETECT değeri 1'den büyük olan sonuçlar çok boyutlu olarak kabul edilmiştir. Böylece testin değerlendirilmesinde daha tutucu bir yaklaşım izlenmiştir. Yöntemlerin performanslarının değerlendirilmesi için gerçek boyut sayısı ile önerilen boyut sayıları karşılaştırılarak doğru kestirim yüzdesi elde edilmiştir. Bunun için,

$$P_r = \begin{cases} 1, & \text{Önerilen Boyut Sayısı} = \text{Gerçek Boyut Sayısı} \\ 0, & \text{Önerilen Boyut Sayısı} \neq \text{Gerçek Boyut Sayısı} \end{cases} \qquad (3)$$

fonksiyonu kullanılmıştır. Bu fonksiyondaki r, replikasyonu ifade etmektedir. Buna göre hesaplanan doğru kestirim yüzdesi ise;

$$Doğru\ Kestirim\ Yüzdesi\ = \frac{\sum_{r=1}^{1000} P_r}{1000}.100 \qquad (4)$$

eşitliğiyle ifade edilebilir. Ayrıca gerçek ve kestirilen faktör sayıları arasındaki Ortalama Fark hesaplanmıştır. Bunun için;

$$Ortalama\ Fark\ = \frac{\sum_{r=1}^{1000} (\widehat{m} - m)}{1000} \qquad (5)$$

eşitliği kullanılmıştır. Burada $\widehat{m}$ önerilen boyut sayısını m gerçekteki boyut sayısını göstermektedir. Replikasyon sayısı 1000 olduğu için ortalaması alınmıştır. Böylece doğru kestirim yüzdesi değerinin yanında yöntemlerin faktör sayısını olduğundan daha az ya da daha fazla kestirip kestirmediği araştırılmıştır.

Aaraştırma sonucunda MAP analizi, hem tetrakorik hem de PPM korelasyon matrisiyle yürütüldüğünde araştırma kapsamındaki tüm koşullar için %99 ve üzerinde doğru kestirim yüzdesine sahip olduğu gözlenmiştir. Ayrıca faktör yükü arttıkça MAP analizinin doğru kestirim yüzdesi yükselmiştir. Buna göre MAP analizinin boyutluluk belirlemede kullanılabileceği söylenebilir.

PA, tetrakorik korelasyon matrisiyle yürütüldüğünde ortalama faktör yükünden ve örneklem büyüklüğünden etkilenmektedir. Küçük örneklemde ve ortalama faktör yükü 0.5 olduğunda PA'nın performansı düşmektedir. Ancak örneklem büyüdükçe PA'nın performansı da artmaktadır. Tetrakorik korelasyon matrisiyle yürütülen PA üzerinde örneklem büyüklüğü oldukça etkili iken PPM korelasyon matrisi üzerinde aynı etki mevcut değildir.

Optimal Koordinat yöntemi tek boyutlu veride örneklem büyüklüğünün 3000 olduğu durumda iyi performans göstermiş ancak diğer durumlarda doğru kestirim yüzdesi %50'nin altında kalmıştır. İvmelenme faktörü yöntemi araştırma kapsamındaki tüm koşullarda tek boyut önermiştir. Buna göre bu yönteminin araştırma kapsamındaki koşullar için, faktörleri ayrıştıramadığı söylenebilir.

DETECT yöntemi iki faktörlü yapıların büyük kısmında %100'e yakın PCE değerine sahipken tek faktörlü yapılarda örneklem büyüklüğü ve ortalama faktör yükünden etkilendiği gözlenmiştir. Ayrıca madde sayısının artması da DETECT'in performansında artışa neden olmaktadır. Ortalama faktör yükü arttıkça benzer şekilde DETECT daha doğru sonuçlar vermektedir.

Araştırmada yer alan koşullara dayalı olarak; 1) MAP analizini hem tetrakorik hem de PPM korelasyon matrisiyle kullanılabileceği, 2) PA yönteminin PPM korelasyon matrisiyle daha doğru sonuçlar verdiği ancak tetrakorik korelasyon matrisiyle yürütüldüğü durumda örneklem büyüklüğünün göz önüne alınması gerektiği, 3) Optimal koordinat ve ivmelenme faktörü yöntemlerinin yerine PA ya da MAP yöntemlerinin tercih edilmesi, 4) DETECT yönteminin ortalama faktör yükü ve örneklem büyüklüğü göz önünde bulundurularak kullanılması önerilmektedir.

TURJE
www.turje.org

Volume: 8 Issue: 3

Turkish Journal of Education

DOI: 10.19128/turje.519235
Research Article

Received 29.01.2019
Accepted 12.07.2019

The use of polychoric and Pearson correlation matrices in the determination of construct validity of Likert type scales

Hasan Fehmi Özdemir
Ankara University, Ankara, Turkey, hozdemir@ankara.edu.tr
Çetin Toraman
Çanakkale Onsekiz Mart University, Çanakkale, Turkey, toramanacademic@gmail.com
Ömer Kutlu
Ankara University, Ankara, Turkey, omer.kutlu@ankara.edu.tr

Check for Updates

ABSTRACT No matter how strong the theoretical infrastructure of a study is, if the measurement instruments do not have the necessary psychometric qualities, there will be a question of trust in interpreting the findings, and it will be inevitable to make wrong decisions with the results. One of the important steps in scale development/adaptation studies is to provide evidence of the experimental validity. In order to reveal evidence of construct validity of Likert type scales, to identify factor structures, to confirm previously predicted structures, factor analysis is used. The primary issue to be examined is the level of measurement of the variable and one of the leading decisions that must be taken is which relation matrix will be used. This descriptive research is based on the effects of using Pearson or polychoric correlation matrix in the factor analysis. It is determined that items showed different "item-total correlations", "loading values" and “correlation coefficients”, different factor numbers emerged, different items were removed out of the scale, confirmation status of the structure has changed.

Keywords: *Factor analysis, Polychoric correlation matrix, Pearson correlation matrix, Likert type scales, Construct validity*

Likert tipi ölçeklerin yapı geçerliğinin belirlenmesi sürecinde polikorik ve Pearson korelasyon matrisinin kullanımı

ÖZ Bir bilimsel çalışmanın teorik altyapısı ne kadar sağlam olursa olsun kullanılan ölçme araçları gerekli psikometrik nitelikleri taşımıyorsa, bulguların yorumlanmasında güven problemi olacak, elde edilen sonuçlarla yanlış kararlar alınması ise kaçınılmaz olacaktır. Ölçek geliştirme-uyarlama çalışmalarında, kuşkusuz en önemli adımlardan biri, aracın psikometrik niteliklerine dair deneysel geçerlilik kanıtlarının ortaya konmasıdır. Bu niteliklerden biri de yapı geçerliğidir. Likert tipi ölçek geliştirme-uyarlama çalışmalarında, yapı geçerliğine ilişkin kanıtların ortaya konması, faktör yapılarının ortaya çıkarılması ya da daha önceden kestirilen faktör yapılarının doğrulanması amacıyla faktör analizi kullanılır. Faktör analizi öncesinde sorgulanması gereken hususların başında verilerin hangi ölçek düzeyinde toplandığı gelmektedir. Analiz sürecinde alınması gereken önemli kararlardan biri ise hangi ilişki matrisinin kullanılacağıdır. Faktör analizinde, Pearson ya da polikorik korelasyon matrisi kullanmanın analiz sonuçları üzerindeki etkisini incelemeyi ve sonuçlarını karşılaştırmayı temel alan bu araştırma betimsel bir araştırmadır. Farklı korelasyon matrisi temelli faktör analizi sonuçlarının birbirinden farklılaştığı, maddelerin farklı “madde toplam korelasyonu”, “yük değeri” ve farklı yönde korelasyon değeri gösterebildiği, farklı faktör sayılarının ortaya çıktığı, farklı maddelerin ölçek dışında bırakılabildiği ve test edilen yapının doğrulanma durumunun değiştiği belirlenmiştir.

Anahtar Kelimeler: *Faktör analizi, Polikorik korelasyon matrisi, Pearson korelasyon matrisi, Likert tipi ölçekler, Yapı geçerliği*

Citation: Özdemir, H.F., Toraman, Ç., & Kutlu, Ö., (2019). The use of polychoric and Pearson correlation matrices in the determination of construct validity of Likert type scales. *Turkish Journal of Education, 8*(3), 180-195. DOI: 10.19128/turje.519235

INTRODUCTION

Psychological measurement instruments are used in many fields of social sciences, which study human behaviors, especially in psychology, educational sciences and sociology, and also in psychiatry, which has a close relationship with social sciences, within its biopsychosocial model. In general terms, psychological tests are systematic approaches that provides information about individuals' abilities, skills, performances, motivations and attitudes, and contribute to make various decisions according to what is obtained from them (Öner, 1987). Studies regarding scales have an important place in every area where psychological activities are conducted.

There are numerous measurement instruments, which have been developed or adapted for different purposes. In addition to the existing ones, there is need to new instruments to be used in the new studies, which especially assess different psychological features, and which can be used for different samples and age groups etc. This need is met through the development of new measurement instruments or adaptation of the existing ones. While both methods have advantages and disadvantages, the most important thing is that validity, reliability and even possible standardization evidences should be established for these instruments by use of appropriate methods and following the correct processes. No matter how strong the theoretical infrastructure of a scientific study is, if the measurement instruments that are used to collect data do not have the necessary psychometric qualities, there will be a question of trust in interpreting the findings obtained by that study, and it will be inevitable to make wrong decisions with the results obtained from the application of these instruments.

Undoubtedly, one of the most important steps in scale development or adaptation studies is to provide evidence of experimental validity of the psychometric properties of the developed or adapted instrument (Crocker & Algina, 2008). In this context, it is necessary to examine whether the instrument measures the intended feature exactly and accurately. The determination of other psychometric qualities of a scale can be done after this examination. Healthy and precise decisions about the scores to be drawn from the application of an instrument can only be taken in the light of the validity evidences.

The notion of validity, which began to evolve in the 1930s thanks to the efforts of practitioners who wanted to break hegemony of the academics in the American Psychological Association (APA), was first defined in 1954 in testing standards report published by a commission established by the American Educational Research Association (AERA), the American Psychological Association (APA), and the National Council on Measurement in Education (NCME) (AEAB, APB, EÖUK, 1997). Although the scope is defined under four categories, namely the content, predictive, concurrent and construct validity, some researchers argue that the concept of validity must be gathered under the framework of construct validity through ongoing discussions (Cronbach & Meehl, 1955; Cronbach, 1984; Messick, 1988, 1993; Şencan, 2005; Urbina, 2014). The investigation process of construct validity is not much different from following the general scientific procedures for development and/or confirmation of a theory (Lemke and Wiersma, 1976). In addition, all the information provided by any validity procedure is relevant to construct validity (Anastasi, 1988). This gives construct validity both a comprehensive status, and also importance and reliability. Therefore, the rest of the validity types can be subsumed under the concept of construct validity. Construct validity has also been associated with factor analysis, which is the most commonly used method to obtain evidence of such validity by many researchers who have even used the concept of factor validity instead of construct validity (Nunnally and Bernstein, 1994).

In scale development and adaptation studies in behavioral sciences; Factor Analysis (FA), which is one of the multivariate statistical methods that Nunnally and Bernstein (1994, p. 111) defined as the heart of the measurement of psychological constructs to reveal evidence of construct validity, to identify factor structures, or to confirm previously predicted factor structures, is used. FA is used in a large number of scientific studies dealing with complex research questions along with other multivariate statistics.

Naturally, in the social sciences such as education, psychology and sociology, where multiple variables are being studied simultaneously and the univariate statistical processes are insufficient to explain; multivariate methods such as multivariate analysis of variance, regression analysis, discriminant analysis, multidimensional scaling and factor analysis are needed. Assuming that these methods together constitute a conceptual basis of the topics covered in scientific research and an arch-like structure that leads to the understanding of causality principles, it can be said that the keystone of this structure is the "factor analysis" (Cattell, 1978).

The purpose of this research is to compare the effect of using Pearson or polychoric correlation matrix on the results of the factor analysis, which is applied in order to provide evidence of construct validity during the development process of likert type scales used in social sciences. Although, similar studies aiming to illustrate the advantages of using polychoric rather than Pearson correlations by means of simulation studies held in the field (Holgado-Tello et al., 2010), one of the most important points that differentiates this study from the previous ones is working with real data. Also some new issues in the scope of this study, such as number of items remained and items' additivity states, which were not analyzed in previous simulation studies, increases the importance of this work.

Factor Analysis

In data reduction with FA, unlike many statistical methods that are used to examine the relationship between dependent and independent variables, there is a summarizing process that aims to understand the underlying cause-and-effect relationships in data sets. Gorsuch (1974) stated that the main purpose of the FA is to develop both the theoretical constructs in a particular field and to reveal the operational representatives of these theoretical constructs. Cattell's (1978, p. 4) definition of FA as the queen of correlational methods is supportive of Pedahazur and Schmelkin's (1991) argument, which states that one of the best method to be used in examining the internal structure of a group of variables or indicators is the factor analytical method. It can be argued that the FA is a statistical technique used to demonstrate construct validity (Atılgan, Kan, & Doğan, 2006; Crocker & Algina, 2008; Bowden, 2004; Erkuş, 2003; Kieffer, 1999; Pedhazur & Pedhazur Schmelkin, 1991; Urbina, 2014). The information obtained in the FA provides a road map for subsequent validity and reliability studies and other statistical analyses to be performed on the basis of the scores obtained from the measurement instrument (Çokluk, Şekercioğlu, & Büyüköztürk, 2010, p.177).

Factor analysis, one of the multivariate statistical methods with a wide range of applications, was developed for use in the field of psychology with the groundbreaking studies of Pearson (1901) and Spearman (1904) at the very beginning of the 20th century. However, it has been used in various scientific fields since the second half of the 20th century. Factor analysis is a set of methods rather than a specific scientific method for determining the fundamental dimensions of a data matrix structure. In particular, researchers, who need to explain an individual's behavior, intelligence, and abilities in a mathematical model, have been compelled to develop this scientific method (Albayrak, 2006, p. 107).

Factor analysis can also be defined as a multivariate statistic that aims to find and discover a smaller number of conceptually meaningful new variable(s) by combining a large number of interrelated variables (Büyüköztürk, 2002). It is divided into two general categories: "exploratory factor analysis" and "confirmatory factor analysis". The exploratory factor analysis attempts to discover the connection between observed variables with unknown latent variables, while the confirmatory factor analysis attempts to confirm the aforementioned structure with the data obtained from the measurement instruments (Çokluk, Şekercioğlu, & Büyüköztürk, 2010). In addition to these two types of factor analysis, in modern factor analysis we see hybrid factor analysis, in which confirmatory rotation methods are used after extraction method (as cited in Henson & Roberts, 2006, p. 395).

Exploratory Factor Analysis (EFA)

When solving the encountered problems, humankind try to reach the solution by reducing the total work. Known as “the Least Effort Law" in the scientific world, we find this theory in all scientific fields as an attempt to explain the most by using the least. Behavioral scientists also try to explain the observed variables with fewer latent variables. In doing so, they use exploratory factor analysis to derive new and independent k variables (factor), equal or less, by taking advantage of the correlation or covariance matrix attained from data set which consists of *j* associated variables (Özdamar, 2013). Since latent structures or factors are thought to summarize the observed variables, they are closely related to the evaluation of the validity of the theory development and observed scores. Theorizing and measurement of structures are processes linked to each other through organic bonds (Henson & Roberts, 2006). As pointed out by Kieffer (1999), the use of factor analytical techniques in social sciences has been integrated with the evaluation of construct validity of both theory development and measurement. Whatever the purpose of the use of exploratory factor analysis is, the significance of the latent variables is directly related to researchers’ definition. As Mulaik (1987) states, it is not the EFA that defines things about psychological features such as intelligence, personality, but it is actually the researchers themselves who make definitions for taking decisions about how to use such concepts. The analytical results, as Thompson and Daniel (1996) argue, will provide information for the definitions to be made, but nevertheless the researcher has full responsibility for the decisions made in this elaboration process. At this point, the role of exploratory factor analysis does not go beyond being a tool. For these reasons, researchers’ decisions are needed to be thought thoroughly in the process of exploratory factor analysis (Henson & Roberts, 2006).

Despite having a fairly wide range of uses, the use of exploratory factor analysis in research is under serious criticism (Fabrigar, Wegener, MacCallum, & Strahan, 1999). Most of these criticisms focus on the subjective decision-making necessity arising from the nature of analysis during the execution of exploratory factor analysis. Tabachnick and Fidell (2013) described the absence of a criterion variable that could be used to test outcomes as one of the problems experienced in factor analysis. The interpretation of the analysis results is, to a large extent, based on the presumed judgments of the researchers, who are also assumed to know the analysis sufficiently.

Researchers, who will apply exploratory factor analysis in their studies, should question some basic concepts, follow certain steps in the analysis process, and make decisions throughout the process. In addition to these, they need to be able to master accurate and complete reporting practices in order to increase the contributions they will make to the literature and to the subsequent studies.

At the beginning of the analysis, the primary issues to be examined are the level of measurement of the data, sampling size, missing values and/or outliers, normality, linearity, multi-collinearity and singularity. How to determine the sample size, and whether it is sufficient or not; how to follow a path if there are missing values and/or outliers; how to test the normality of the data, how to obtain useful results from the data set in cases where the normal distribution cannot be obtained, and how to differentiate conditions where the multivariate normal distribution is a necessity are the issues researchers consider.

In the process of analysis, the leading decisions that must be taken by the researchers are; which relation matrix (which correlation matrix or variance-covariance matrix) will be used, which factors in predicting factor/factorization method to be used (main/principal component method, basic/common factor analysis, maximum likelihood method, generalized least squares method, principal axis factorization method, alpha factorization method, image factorization method), which rules will be based on when determining the number of factors (Kaiser criterion/rule-all factors with eigenvalues greater than one, Cattell Scree test or scree plot, explained variance criterion, Joliffe’s criterion), in which cases the factor rotation to be applied, while doing this which rotation method (orthogonal or oblique) will be selected (varimax, quartimax and equamax or direct obliques and promax), and the theoretical and practical basis for the decisions that might largely differentiate results, assumptions and the advantages and disadvantages should also be known.

Confirmatory Factor Analysis (CFA)

In contrast to EFA, which produces theory, CFA is another factor analysis method that tests the theory. The factor analysis method, which is used to reach the answers of the questions, such as whether the relation between the factors belonging to a structure revealed by EFA and the variables are sufficient or not, which variables are related to which factors, whether the factors are independent of each other, whether these factors are sufficient enough to explain the original structure, is called CFA (Özdamar, 2013).

Exploratory factor analysis is used to reveal the best factor model for the observed data set when researchers do not have any idea of the underlying factors before the application of a psychometric measurement instrument. On the other hand, in the event that there is a previously defined and bounded structure, in other words a hypothesis about the underlying factors, then the methodology that the researchers refer to when testing this model/factor structure systematically is confirmatory factor analysis (Bryant, Yarnold, & Michelson, 1999). CFA is used to test a previously validated and reliable instrument's usability in a new culture, in a field, and/or to a target group. It has been indicated that CFA is a more appropriate method to be used by researchers in assessing the construct validity (Stapleton, 1997), and it has been emphasized that between the two main factor analytical methods, CFA is both theoretically more important and should be used more widely (Gorsuch, 1983).

CFA offers important advantages such as comparing the different factorial structures put forth by EFA, providing an opportunity to correct the conceptual and statistical susceptibility of the model, which has been thought to be weak, by transforming it into an improved model with a more reliable and efficient structure.

As in the EFA, it is necessary for the researchers to be in control of a number of issues to be done before and after the analysis in the CFA, and also accurate and complete reporting practices in order to increase the contributions they will provide to the literature and for the follow-up studies.

Before CFA, the relationship matrix to be used depending on the level of measurement of the data available should be determined. The analysis then begins with a description of the model in the direction of theoretical bases. In doing so, the model is determined by fixing or releasing certain parameters (factor coefficients, factor correlation coefficients, variance-covariance of the measurement error) in line with the theoretical expectations of a researcher. This is followed by an analysis of the fit statistics obtained from the estimations of the collected data set and model parameters by using certain programs (AMOS, LISREL, etc.). By using different fit statistics (Chi-Square Goodness of Fit, Goodness-of-Fit Index-GFI, the Root Mean Square Error of Approximation-RMSEA, Residual Averages and Root Mean Square Residual-RMR, Comparative Fit Index-CFI, Normalized Compliance Index-NFI and Non-Normative Compliance Index-NNFI and Parsimony Goodness-of-Fit Index-PGFI), the analysis is continued by evaluating the resulting model fit. If the fit indices cannot meet the acceptance levels, the modification indices must first be evaluated and, if necessary, the model must be redefined and the process must be repeated from the first step. Factor structure, the number of factors, the degree to which each item has a high load to which factor, the model supported by a previous study or a developed theory, and even the amount of error should be implicitly defined.

As in all other statistical methods, the leading issue to be questioned before the factor analysis is at what measurement level the data are collected. Theoretical bases and methods used for factor analysis with continuous variables have been fairly developed. In practice, however, some of the observed and/or measured variables are at the level of ordinal scale, which is often overlooked and is incorrectly treated as if they had the numerical metric property representing the sequential categories like 1, 2, 3, 4 (Jöreskog & Moustaki, 2001). The use of appropriate statistical techniques by accepting the data obtained from these scales as if they are at interval levels, especially due to the item response formats of the scales (Likert type, etc.), which are used frequently in all fields covered by social sciences has been criticized in advance. Researchers who have expressed this criticism indicate that the data obtained from such scales are at the ordinal level of measurement and that appropriate statistical techniques should be used for these data (Stevens, 1946; Thomas, 1982; Jamieson, 2004). When the ordinal scale level variables considered, the necessity of using "tetrachoric" (for two categorical data) or "polychoric" (for three or more categorical data) correlation matrices have been emphasized, while estimating the relationship between variables or conducting correlation-based analysis (Holgado-Tello, Chacón-Moscoso, Barbero- García, & Vila-Abad, 2010; Uebarsax, 2015). Jöreskog and Sörbom (2002) found that the most consistent, reliable, and strongest predictors of factor analysis can be achieved by using the polychoric correlation matrix. However, the Pearson correlation matrix is often used during the application of both exploratory and confirmatory factor analysis used in the scale development and adaptation processes. When the level of the relationship between categorical data are studied, various reasons are suggested why Pearson correlations are not appropriate (Holgado-Tello et al., 2010). Firstly, categorical variables are variables providing information at the level of ordinal scale, and Pearson correlation requires measurements at equal interval level. The information obtained from categorical variables is limited only to the number of observations per cell/categorization in the contingency table. If Pearson correlation is used under these conditions, the restriction/limitation of categorization/classification will lead to the artificial restriction/limitation of the relationship between measurements (Guilley & Uhlig, 1993), all observations / persons placed on the intervals, where the boundaries of each category determined, will be scored by accepting in one of the categories, which will result in a reduction in variability, in other words, a loss of data, even if the data are actually different from each other. Considering that in homogeneous samples, Pearson's correlation gets lower values than usual, the restrictions/limitations arising from this way of assigning scores to the observations/individuals will cause to determine the degree of the relationship between the observed variables lower as well, and consequently it will cause a decline/decrease in the factor loadings obtained by factorization of the correlation matrix (DiStefano, 2002). For this reason, when factor analysis will be used to test the validity of an instrument, it is of paramount importance that the measurement level at which the data is collected is taken into account. On the other hand, frequent use of the Pearson correlation matrix in analyses is due to the fact that researchers do not have enough knowledge about the subject or the limitations of the computer package programs they use. The best example of this is the fact that IBM-SPSS (Statistical Package for the Social Sciences), one of the most popular statistical package programs applied in the social sciences, does not recognize a preference for the correlation matrix to be used during factor analysis, and uses Pearson correlation matrix as a default. Significant amount of measurement error, especially random and systematic errors, are intermingled from different sources, in social sciences, and distorting the estimates of the relationship between the variables involved in the research, which leads to incorrect results. In this context, at the beginning of the negative consequences of the use of different types of correlations in the statistical analysis in terms of methodological investigations, it is possible that the researchers make mistaken inferences about the construct validity, which is the cornerstone of the basic and applied sciences.

METHODOLOGY

This study is a descriptive research. Descriptive research is the study of describing existing conditions without being interested in the relationships or differences between variables. In this context, descriptive studies serve for the purpose of describing science, as well as to provide an insight into the production of hypothesis for further researches (Erkuş, 2013).

Participants

A methodological comparison is planned in this research. The purpose of this research is neither to adapt a new instrument from a different culture, nor to develop a new instrument. For this reason, it has been deemed appropriate to carry out the methodological discussion through an instrument, which has been already developed. The data obtained during the development process of the Gender Equality Scale (GES), which was developed by Gözütok, Toraman and Acar Erdol (2017), were also used in this research with the permission of the researchers. Therefore, the participants are the same participants whose data were collected for the purpose of conducting exploratory and confirmatory factor analyzes in the mentioned study.

In Gözütok et al. (2017) study, data were collected from two separate groups of high school students to perform exploratory and confirmatory factor analysis. The group to collect data for the exploratory factor analysis consists of 396 students and for confirmatory factor analysis consists of 265 students. It was also noted that while groups were formed, groups were heterogeneous (different genders, different class levels [9^{th}, 10^{th}, 11^{th} and 12^{th} grade], in terms of cultural activities such as going to cinema, theater, reading newspapers and books) and balanced group distribution in terms of gender.

Data Collection Tools

In this study, the Gender Equality Scale (GES), and the data collected for factor analysis during the development of this scale were used.

Analysis of the Data

Gözütok et al. (2017) study was analyzed through IBM-SPSS based on the Pearson correlation matrix. Typical analyses used in many scale development studies were used in this analysis. They are namely; examination of item total correlations, Kaiser-Meyer-Olkin (KMO), Bartlett Sphericity test, examination of eigenvalues, varimax axis rotation. The Cronbach Alpha reliability coefficient was used to obtain evidence of reliability (Büyüköztürk, 2013; Özdamar, 2013). Factor analysis findings and results of Gözütok et al. (2017) were used for comparison with the permission of researchers.

In the factor analysis, "FACTOR" software developed by Lorenzo-Seva and Ferrando (2006) was used to obtain a comparative polychoric correlation matrix and to carry out the analyses. This software is free, useful and small-only focuses on factor analysis. By means of this software, it is possible to obtain a tetrachoric correlation matrix so that factor analysis of achievement tests, which can be coded as 0 and 1, can be realized. In addition, as in the case of Likert scales, which have multiple categories, a polychoric correlation matrix can be obtained and factor analysis is carried out over this matrix. The data file was uploaded onto the software and analyzed. In the analysis, the total correlations of the items were examined, the eigenvalues were examined, the Kaiser-Meyer-Olkin (KMO) and Bartlett Sphericity tests were performed, and the varimax rotation method was used when multiple factors were tested. The reliability of the factor software and the model fit indices were taken into account.

The data collected for the CFA were transferred to the AMOS 22 program, and the analyses were carried out with this program.

FINDINGS

According to the researchers who argue that when the measurement instruments developed in social sciences are at the ordinal scale, it is necessary to use the polychoric correlation matrix in the development process of Likert scales (Holgado-Tello, Chacón-Moscoso, Barbero- García, & Vila-Abad, 2010; Jöreskog & Sörbom, 2002; Uebarsax, 2015). Based on this argument, the first analysis was conducted without the number of factors being specified. The FACTOR software not only proposes factor numbers, provides information on eigenvalues, but also gives information regarding the contribution of the items (communality) to scale. As shown in Figure 1, after reviewing this information, the number of factors and the items to be deducted from the analysis were decided.

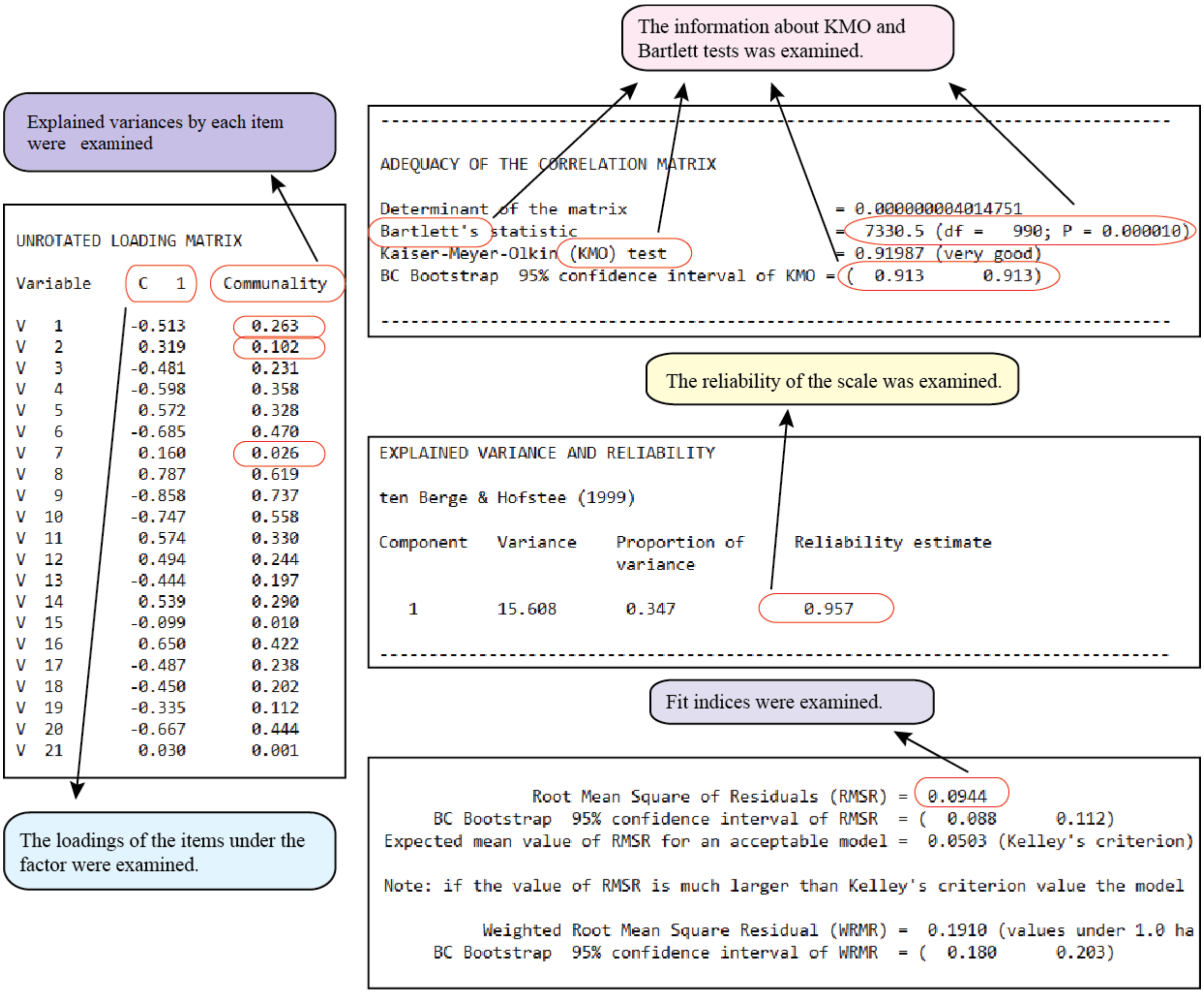
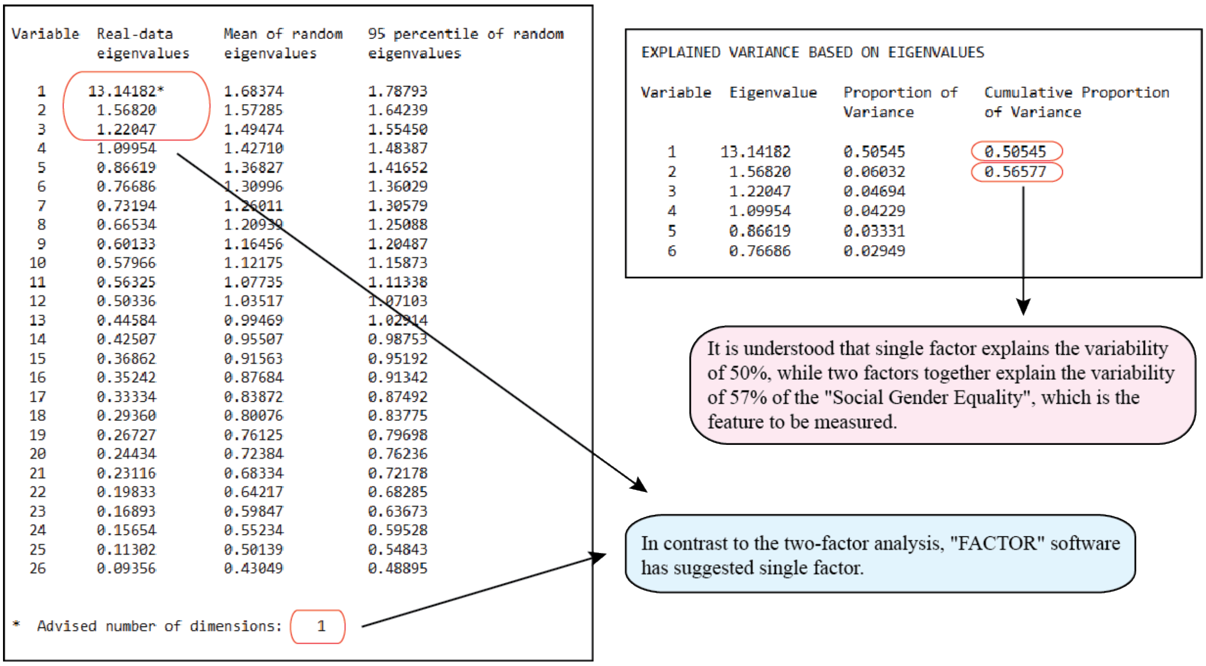


Figure 1. Analysis output file of "FACTOR" software

At the end of the contribution evaluation, where the input values and the items are scaled without specifying the factor number; it is understood that items 1, 2, 3, 7, 12, 13, 14, 15, 17, 18, 19, 21, 22, 29, 35, 37, 38, 39, and 40 contributed to the factor (lower than 0.300) with low correlation level. By removing these items and having known the factor structure of the scale in Gözütok et al. (2017) study, the analysis was repeated with two factors.



Variable	Real-data eigenvalues	Mean of random eigenvalues	95 percentile of random eigenvalues
1	13.14182*	1.68374	1.78793
2	1.56820	1.57285	1.64239
3	1.22047	1.49474	1.55450
4	1.09954	1.42710	1.48387
5	0.86619	1.36827	1.41652
6	0.76686	1.30996	1.36029
7	0.73194	1.26011	1.30579
8	0.66534	1.20939	1.25088
9	0.60133	1.16456	1.20487
10	0.57966	1.12175	1.15873
11	0.56325	1.07735	1.11338
12	0.50336	1.03517	1.07103
13	0.44584	0.99469	1.02914
14	0.42507	0.95507	0.98753
15	0.36862	0.91563	0.95192
16	0.35242	0.87684	0.91342
17	0.33334	0.83872	0.87492
18	0.29360	0.80076	0.83775
19	0.26727	0.76125	0.79698
20	0.24434	0.72384	0.76236
21	0.23116	0.68334	0.72178
22	0.19833	0.64217	0.68285
23	0.16893	0.59847	0.63673
24	0.15654	0.55234	0.59528
25	0.11302	0.50139	0.54843
26	0.09356	0.43049	0.48895

* Advised number of dimensions: 1

EXPLAINED VARIANCE BASED ON EIGENVALUES

Variable	Eigenvalue	Proportion of Variance	Cumulative Proportion of Variance
1	13.14182	0.50545	0.50545
2	1.56820	0.06032	0.56577
3	1.22047	0.04694	
4	1.09954	0.04229	
5	0.86619	0.03331	
6	0.76686	0.02949	

Figure 2. Analysis results obtained from "FACTOR" software by attempting two factors

As seen in Figure 2, the single factor, explained the 50% of the variability in SGE. The software suggested that analysis should be with single factor. The analysis was continued with a single factor structure. The results obtained were summarized in table 1.

Table 1.
Exploratory factor analysis and reliability analysis results based on polychoric correlation matrix

Item No	Item Loadings	Item Total Correlation	Item No	Item Loadings	Item Total Correlation
M4	-0.609	0.370	M27	0.796	0.633
M5	0.584	0.341	M28	0.847	0.718
M6	-0.692	0.478	M30	-0.711	0.505
M8	0.801	0.641	M31	-0.721	0.520
M9	-0.870	0.757	M32	0.704	0.496
M10	-0.765	0.586	M33	0.805	0.648
M11	0.594	0.353	M34	-0.581	0.337
M16	0.662	0.438	M36	-0.624	0.389
M20	-0.672	0.451	M41	0.769	0.592
M23	-0.622	0.387	M42	-0.549	0.302
M24	-0.572	0.327	M43	0.805	0.648
M25	-0.702	0.493	M44	0.827	0.684
M26	-0.696	0.485	M45	0.750	0.562

Kaiser Meyer Olkin (KMO) = 0.941
Bartlett = 4737,0, sd = 325, p<.01
Variance Explained by Single Factor = 0.505
Reliability = 0.961

When Table 1 is examined, it is seen that the loading values of the remaining 26 items do not fall below 0.549 and the item total correlation does not fall below 0.327. It has been seen that the items 4, 6, 9, 10, 20, 23, 24, 25, 26, 30, 31, 34, 36 and 42 have negative correlation values. In fact, some of these items are negative or have a negative meaning. Others do not have a negative meaning or are not negative. However, if the scale will become a single structure with all of the remaining materials after the analysis, it is possible that these items will be encoded negatively in order to make them meaningful.

KMO and Bartlett results of the structure in Table 1 show that it is within the accepted values in the literature. The value of reliability is at the desired level that is sought in social sciences. The variance explained by the single factor is within the acceptable boundaries according to some authors (Büyüköztürk, 2013) while it cannot be accepted by others (Özdamar, 2013).

It is seen that Pearson's correlation matrix-based factor analysis results differ from the results based on the polychoric correlation matrix. Pearson's correlation matrix-based factor analysis results are summarized in Table 2.

Table 2.
Exploratory factor analysis and reliability analysis results based on Pearson correlation matrix

Item No	Factor	Item Total Correlation	Item Loadings After Rotation	
			Factor I	Factor II
M11	Factor 1	0,463	0,452	
M27	Factor 1	0,654	0,709	
M28	Factor 1	0,710	0,652	
M33	Factor 1	0,673	0,667	
M41	Factor 1	0,632	0,582	
M43	Factor 1	0,691	0,758	
M44	Factor 1	0,672	0,848	
M45	Factor 1	0,628	0,801	
M12	Factor 2	0,400		0,535
M14	Factor 2	0,479		0,619
M16	Factor 2	0,620		0,734
M22	Factor 2	0,333		0,632
M32	Factor 2	0,623		0,542

KMO = 0,922
Bartlett Sphericity (X2) = 2072,965; sd=78, p<0.01
Variance Explained by Factor 1 = %32,780
Variance Explained by Factor 2 = %20,051
Variance Explained by Both Factors Together= %52,831
Cronbach Alpha = 0,889

When the results in tables 1 and 2 are evaluated, it is seen that they are quite different. In the factor analysis based on the polychoric correlation matrix, items 4, 6, 9, 10, 20, 23, 24, 25, 26, 30, 31, 34, 36 and 42, which are shown with negative correlations, were removed from the scale in the Pearson correlation matrix-based factor analysis. Social gender equality’s explained variance is close in both analyzes. KMO and Bartlett values and reliability values are at the desired level in both analyses (Büyüköztürk, 2013; Özdamar, 2013).

In accordance with the results obtained from the polychoric correlation matrix with the FACTOR software that are shown in table 1, the items 4, 6, 9, 10, 20, 23, 24, 25, 26, 30, 31, 34, 36, and 42 were reverse coded by returning to the same data set in the SPSS program. After this coding, in the SPSS software, additivity analysis was performed for 26 items and single factor structure that FACTOR software revealed. The results are summarized in table 3.

Table 3.
Results of additivity analyses of GES (Structure obtained by polychoric correlation matrix-based factor analysis)

Variance Source	Sum of Squares	Mean of Squares	F	sd	p
Nonadditivity	20.645	20.645	20.711	1	0.000

As a result of the analysis, the non-additivity state is meaningful in the structure consisting of 26 items and one dimension (Tukey Non-additivity p<.05). In this case, additivity is meaningless (Özdamar, 2013). In short, the structure is not in additivity state. The two-dimensional structure obtained by Gözütok et al. (2017) shows additivity. This difference can be explained by the fact that the Pearson correlation matrix accepts the data set at interval level and attempts to reach a scale with an additivity structure. However, the criterion of additivity for a scale structure of an ordinal measurement level is a weak one.

The structure obtained as a result of exploratory factor analysis based on the polychoric correlation matrix for GES is shown in table 1. For this structure, confirmatory factor analysis was applied with the unweighted least square (ULS) method. Unweighted least square (ULS) method is used for confirmatory factor analysis based on polychoric correlation (Katsikatsou, Moustaki, Yang-Wallentin, & Jöreskog, 2012). Confirmatory factor analysis with this method did not confirm the structure.

It has been determined that the two-dimensional structure obtained by Gözütok et al. (2017) has had close results with the fit indices obtained by Gözütok et al. (2017) when using confirmatory factor analysis by unweighted least square-ULS method. The results are summarized in table 4.

Table 4.
Fit indices after CFA

Estimation	χ^2	sd	χ^2/sd	RMSEA	AGFI	RMR	NFI
Pearson Maximum Likelihood	97,01	53	1,83	0,056	0,92	0,062	0,97
Unweighted Least Square	100,22	64	1,57	0,046	0,99	0,067	0,99

When table 4 is examined, it has been determined that the estimations made with the maximum likelihood and unweighted least square methods give very close fit index results. Koğar and Yılmaz Koğar (2015) stated that unweighted least squares (ULS) method used in confirmatory factor analysis of ordinal data is a suitable technique for estimating parameters with a minimum number of repetitions and estimating the parameters. Similar results have also been obtained in this study.

DISCUSSION and CONCLUSION

This study aims to compare the effect of using the Pearson or the polychoric correlation matrix on the analysis results in the process of developing the likert type scales in order to provide evidence of construct validity.

In studies conducted in the fields of social and behavioral sciences, all measurements have a degree of uncertainty/error regardless of precision and accuracy, which causes randomly and/or systematically, and consequently, these problems are reflected in the results. Especially in the analyses carried out while collecting evidence of reliability and validity of Likert type scales developed to be used as data collection instruments in scientific researches or adapted from one culture to another, it was found that using different correlation matrices could increase the error that would cause confusions in the results, and they may lead to a misinterpretation of the construct validity evidence, which is regarded as the most important of the validity types by many researchers (Cronbach & Meehl, 1955; Cronbach, 1984; Messick, 1988, 1993; Şencan, 2005; Urbina, 2014). Moreover, this situation will inevitably effect the results of subsequent studies carried out using the corresponding scales.

Similar to the previous studies carried out with simulated data (Holgado-Tello et al., 2010), it is seen in this real data used study that different correlation matrix based factor analyses results were found to differ from each other. In comparison to the previously held studies carried out with simulated data, this study is based on real data, and it also handles new discussions on number of items, items ‘additivity states. As such, it differs from the previously held studies.

The findings of this study can be summarized as follows: As a result of the analyses carried out using the Polychoric or Pearson correlation matrix,

1) The items showed different "total item correlation" and "loading values".
2) The items could show a different (negative or positive) correlation coefficient.

3) Different factor numbers emerged.
4) Whether or not the resulting sub-dimensions' (factors) possibility of being additive or non-additive states have changed.
5) Different items were removed out of the scale.
6) Being confirmed and not confirmed status of the structure's being tested, has changed.
7) It was determined that estimations made by the maximum likelihood and unweighted least square methods have given very close fit index results.

All of these results clearly show that when the analyses are carried out with the data obtained from the Likert type scales, the results differ greatly from each other when the analyses are conducted by grounding Pearson correlation matrix at the interval measurement level or by grounding the polychoric correlation matrix, in which the data are accepted at the ordinal measurement level. From the perspective of the differences revealed by this study, which has conducted in terms of factor analysis, especially specific to construct validity; for researchers, who will use Likert-type scale or scales in their studies as a means of data collection by developing, and/or adapting, it is recommended to use both Pearson or polychoric correlation matrices for aforementioned analyses, and to compare the results by taking into account the theoretical background of the psychological feature they aim to measure. It is better to decide on which results they will choose, and the final construct they will prefer according to this comparison. Yet, the existence of two different structures will cause the view that the instruments developed and/or adapted for measurement do not actually measure the claimed psychological property, but measure a different property, or cause significant errors in the measurement results.

In the literature, there is a variety of information about the minimum number of participants required for the factor analysis applications in the measurement instrument development process. For example, Kline (2005) recommended at least 100; Hutcheson and Sofroniou (1999) stated at least 150 to 300 and Cattell (1978) claimed the desirable N to be at least 250 participants. Comrey and Lee (1992) described 100 participants as poor, 200 as moderate, 300 as good, 500 as very good and 1000 and more as excellent. Cattell (1978) emphasized that the number of participants in the factor analysis should be 3 to 6 times the number of the items and Gorsuch (1974) stated that it should be at least 5 times. Everitt stated that the number of participants should be at least 10 times the number of items (as cited in Arrindell and van der Ende, 1985, p. 166). The data set used in this study can be said to be obtained from a sufficient number of individuals according to the literature. The results showed that when worked with the proper sample size suggested by the literature, the solutions with Pearson correlation matrix give proper results.

Even though a real data set was used, one limitation of this study is the fact that the results of the factor analysis of the correlation matrices were discussed from a single data set. Performing methodological comparisons over more real data sets will help to make a healthier decision about which of the Pearson or polychoric correlation matrices is proper. Therefore, it is recommended to repeat this methodological comparison with different data sets within the scope of different studies. When performing factor analysis, discussing the results of the data sets which are below the sufficient sample level, can also support to make healthier decisions about the effectiveness of the methods as well. In this sense, it is recommended to repeat similar comparisons over different sized data sets.

This study is only limited to some of the results of factor analysis. Apart from them, other issues that are not taken into account both within the scope of factor analysis (such as cross loadings of items, different estimation methods, different types of rotations, whether different number of categories in the response format differ the results or not, etc.), and within the scope of analyses to obtain evidence of reliability, which is another important psychometric property for psychological measurement tools, and also within the scope of all other correlation-based analyses whether they are effected or not according to the used correlation matrix, can be studied with new studies.

REFERENCES

Albayrak, A.S. (2006). *Uygulamalı çok değişkenli istatistik*. Ankara: Asil Yayın

Amerikan Eğitim Araştırmaları Birliği, Amerikan Psikoloji Birliği, Eğitim Ölçümleri Uluslararası Konseyi (1997). *Eğitimde ve psikolojide ölçme standartları*. (S. Hovardaoğlu and N. Sezgin Trans.). (1st Edition). Ankara: Türk Psikologlar Derneği Yayınları. No 14

Anastasi, A. (1988). *Psychological testing* (6th Ed.). New York: McMillan.

Arrindell, W. A. & Van der Ende, J. (1985). An empirical test of the utility of the observations to variables ratio in factor and components analysis. *Applied Psychological Measurement*, 9(2), 165-178.

Atılgan, H., Kan, A., & Doğan, N. (2006). *Eğitimde ölçme ve değerlendirme*. Ankara: Anı Yayıncılık

Babakus, E., Ferguson, C. E., Jr., & Jöreskog, K. G. (1987), The Sensitivity of Confirmatory Maximum Likelihood Factor Analysis to Violations of Measurement Scale and Distributional Assumptions, *Journal of Marketing Research, 24* (May), 222- 228.

Bowden, S. C. (2004). The role of factor analysis in construct validity: Is it a myth? *Journal of International Neuropsychological Society. 10*, 1018-1019

Bryant, F. B., Yarnold, P. R., & Michelson, E. A. (1999). Statistical methodology: VIII. Using confirmatory factor analysis (CFA) in emergency medicine research. *Academic Emergency Medicine, 6*(1), 54-56

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

Büyüköztürk, Ş. (2013). *Sosyal bilimler için veri analizi el kitabı*. Ankara: Pegem Yayıncılık

Cattell, R. B. (1978). *The scientific use of factor analysis in behavioral and life sciences*. New York: Plenum

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

Crocker, L., & Algina, J. (2008). *Introduction to classical and modern test theory*. USA: Cengage Learning

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*, 281–302

Cronbach, L. (1984). *Essentials of Psychological Testing*. Harper and Row, New York

Çokluk, Ö., Şekercioğlu, G. and Büyüköztürk, Ş. (2010). *Sosyal bilimler için çok değişkenli istatistik SPSS ve LISREL uygulamaları*. Ankara: Pegem Akademi

DiStefano, C. (2002). The impact of categorization with confirmatory factor analysis. *Structural Equation Modeling, 9*, 327-346

Erkus, A. (2003). *Psikometri üzerine yazılar* (1st Edition). Ankara: Türk Psikologlar Derneği Yayınları

Erkuş, A. (2013). *Davranış bilimleri için bilimsel araştırma süreci*. Ankara: Seçkin Yayıncılık.

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*, 272-299

Gorsuch, R. L. (1974). *Factor analysis* (First Edition). Philadelphia: W. B. Saunders Company

Gorsuch, R. L. (1983). *Factor Analysis* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Gözütok, F. D., Toraman, Ç., & Acar Erdol, T. (2017). Toplumsal cinsiyet eşitliği ölçeğinin (TCEÖ) geliştirilmesi. *İlköğretim Online, 16*(3), 1036-1048. doi: 10.17051/ilkonline.2017.330240

Guilley, W., & Uhlig, G. (1993). Factor analysis and ordinal data. *Education, 114*(2), 258-264

Henson, R.K., & Roberts, J.K. (2006). Use of EFA in published research: Common errors and some comment on improved practice. *Educational and Psychological Measurement, 66*, 393-416.

Holgado-Tello, F. P., Chacón-Moscoso, S., Barbero- García, I., & Vila-Abad, E. (2010). Polychoric versus Pearson correlations in exploratory and confirmatory factor analysis of ordinal variables. *Quality and Quantity, 44*, 153-166

Hutcheson, G. & Sofroniou, N. (1999). *The multivariate social scientist, introductory statistics using generalized linear models*. The USA: Sage Publications

Jamieson, S. (2004). Likert Scales: How to (Ab)use them. *Medical Education, 38*, 1217-1218

Jöreskog, K. G., & Moustaki, I. (2001). Factor analysis of ordinal variables: A comparison of three approaches, *Multivariate Behavioral Research, 36*(3), 347-387.

Jöreskog, K. G., & Sörbom, D. (2002). *Prelis 2: User's reference guide*. Lincolnwood: Scientific Software International, Inc

Katsikatsou, M., Moustaki, I., Yang-Wallentin, F., & Joreskog, K. (2012). Pairwise likelihood estimation for factor analysis models with ordinal data. *Computational Statistics and Data Analysis, 56*(12), 4243-4258.

Kieffer, K. M. (1999). An introductory primer on the appropriate use of exploratory and confirmatory factor analysis. *Research in the Schools, 6*, 75-92

Kline, T. J. B. (2005). *Psychological testing, a practical approach to design and evaluation*. The USA: Sage Publications

Koğar, H., & Yılmaz Koğar, E. (2015). Comparison of different estimation methods for categorical and ordinal data in confirmatory factor analysis. *Journal of Measurement and Evaluation in Education and Psychology, 6*(2), 351-364

Lemke, E. & Wiersma, W. (1976). *Principles of Psychological Measurement*. Chicago: Rand McNally.

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

Messick, S. (1988). *The once and future issues of validity*: *Assessing the meaning and consequences of measurement*. In H. Wainer, & H. I. Braun. (Eds.), Test validity. (s. 33-48). Hillsdale, NJ, US: Lawrence Erlbaum Associates, Inc.

Messick, S. (1993). *Foundations of validity: Meaning and consequences in psychological assessment*. Presented at the Second Conference of the European Association of Psychological Assessment. Groningen, Netherlands

Mulaik, S. A. (1987). A brief history of the philosophical foundations of exploratory factor analysis. *Multivariate Behavioral Research, 22*, 267-305

Nunnaly, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (Third Edition). NY: McGraw-Hill, Inc.

Öner, N. (1987). Kültürlerarası ölçek uyarlamasında bir yöntembilim modeli. *Türk Psikoloji Dergisi, 6*(21), 80-83

Özdamar, K. (2013). *Paket programlar ile istatistiksel veri analizi (Volume 1 and 2)* (9th Edition). Eskişehir: Nisan Kitabevi

Pedhazur, E. J., Pedhazer Schmelkin, L. (1991). *Measurement, design and analysis: An integrated approach* (1st Edition). NJ: Lawrence Erlbaum Associate, Inc

Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *Philosophical Magazine, 6*, 559-572

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

Stapleton, C. D. (1997). *Basic concepts in exploratory factor analysis (EFA) as a tool to evaluate score validity: A right-brained approach* (pp. 1–8). Presented at the Southwest Educational Research Association, Austin

Stevens, S. S. (1946). On the theory of scales of measurement. *Science, 103*, 677-680

Şencan, H. (2005). Sosyal ve davranışsal ölçümlerde güvenilirlik ve geçerlilik. (1st Edition). Ankara: Seçkin Yayınları.

Tabachnick, B. G., & Fidell, L. S. (2013). *Using multivariate statistics* (6th ed.). USA: Pearson

Thomas, H. (1982). IQ interval scales, and normal distributions. *Psychological Bulletin, 91*, 198-202

Thompson, B., & Daniel, L. G. (1996). Factor analytic evidence for the construct validity of scores: A historical overview and some guidelines. *Educational and Psychological Measurement, 56*, 197-208

Uebersax, J. S. (2015). *The tetrachoric and polychoric correlation coefficients. Statistical Methods for Rater Agreement*. Retrieved from http://john-uebersax.com/stat/tetra.htm

Urbina, S. (2014). *Essentials of psychological testing* (Second Edition). NJ: Wiley & Sons, Inc.

TÜRKÇE GENİŞLETİLMİŞ ÖZET

Psikolojik ölçme araçları, genelde insan davranışlarını konu edinmiş pek çok sosyal bilim dalında, özellikle psikoloji, eğitim, sosyoloji ve psikiyatride kullanılır. Genel anlamı ile psikolojik testler bireylerin yetenekleri, becerileri, performansları, güdüleri, tutumları hakkında bilgi veren ve bunlardan elde edilecek verilere göre çeşitli kararların alınmasına yardımcı olan sistematik yaklaşımlardır.

Türkiye'de değişik amaçlarla geliştirilmiş ya da uyarlanmış çok sayıda ölçme aracı bulunmaktadır. Yapılacak yeni araştırmalarda kullanılmak üzere, var olanlara ek olarak özellikle farklı psikolojik özellikleri ölçen, farklı örneklemler, farklı yaş grupları vb. için kullanılabilecek yeni ölçme araçlarına da ihtiyaç duyulmaktadır. Bu ihtiyaç ise, yeni ölçme araçlarının geliştirilmesi ya da uyarlanması yolu ile karşılanmaktadır. Her iki yöntemin de avantajlı ve dezavantajlı yönleri bulunmakla birlikte, asıl önemli olan bu araçlar için gerekli geçerlik, güvenirlik ve hatta mümkünse standardizasyon çalışmalarının uygun yöntemlerle ve bu yöntemlerin de doğru süreçler izlenerek yapılmış olmasıdır. Çünkü yapılan bir bilimsel çalışmanın teorik altyapısı ne kadar sağlam olursa olsun veri toplamak amacıyla kullanılan ölçme araçları gerekli psikometrik nitelikleri taşımıyorsa o çalışma ile ulaşılacak bulguların yorumlanmasında güven problemi olacak, bu araçların uygulanmasından elde edilen sonuçlarla yanlış kararlar alınması ise kaçınılmaz olacaktır. Ölçek geliştirme ya da uyarlama çalışmalarında, kuşkusuz en önemli adımlardan biri, geliştirilen ya da uyarlanan aracın psikometrik niteliklerine dair deneysel geçerlilik kanıtlarının ortaya konmasıdır. Bu bağlamda, söz konusu aracın ölçmeyi amaçladığı özelliği tam ve doğru bir şekilde ölçüp ölçmediğine ilişkin bir sorgulamanın yapılması gerekmektedir. Ölçeğe ait diğer psikometrik niteliklerin belirlenmesi bu sorgulamanın ardından yapılabilecek ve ölçeğin uygulanmasından elde edilecek puanlara dair sağlıklı kararlar da ancak, aracın geçerliliğine ilişkin kanıtlar doğrultusunda alınabilecektir.

Her ne kadar kapsam, yordama, zamandaş/eş-zamanlı ve yapı geçerliliği olmak üzere 4 kategori altında tanımlanmış olsa da günümüze kadar süre gelen tartışmalar içerisinde geçerlik kavramının, yapı geçerliği çatısı altında toplanması gerektiğini savunan araştırmacılar olmuştur (Cronbach ve Meehl, 1955; Cronbach, 1984; Messick, 1988, 1993; Şencan, 2005; Urbina, 2014). Birçok araştırmacı tarafından da yapı geçerliği, bu tür geçerliğe ait kanıt elde etmek amacıyla en sık kullanılan yöntem olan faktör analiziyle ilişkilendirilmiş, hatta yapı geçerliliği kavramı yerine faktör geçerliliği kavramını kullanmışlardır (Nunnally ve Bernstein, 1994).

Davranış bilimlerinde ölçek geliştirme ve uyarlama çalışmalarında; yapı geçerliğine ilişkin kanıtların ortaya konması, faktör yapılarının ortaya çıkarılması ya da daha önceden kestirilen faktör yapılarının doğrulanması amacıyla Nunnally ve Bernstein'nin (1994, sf. 111), psikolojik yapıların ölçümünün kalbi olarak tanımladığı, çok değişkenli istatistiksel yöntemlerden biri olan Faktör Analizi (FA) kullanılır. FA, diğer çok değişkenli istatistiklerle birlikte karmaşık araştırma sorularının ele alındığı çok sayıda bilimsel çalışmada kullanılmaktadır. Elbette eğitim, psikoloji, sosyoloji gibi sosyal bilim alanlarında da birden çok değişkenin eşzamanlı çalışıldığı ve tek değişkenli istatistiksel işlemlerin açıklamakta yetersiz kaldığı durumlarda; çok değişkenli varyans analizi, regresyon analizi, diskriminant analizi, çok boyutlu ölçekleme ve faktör analizi gibi çok değişkenli yöntemlere ihtiyaç duyulmaktadır. Bu yöntemlerin bir arada, bilimsel araştırmalarda ele alınan konuların kavramsal temellerinin ve nedensellik ilkelerinin anlaşılmasına açılan kemer şeklinde bir yapı oluşturduğunu varsayarsak, bu yapının kilit taşının "faktör analizi" olduğu söylenebilir. Tüm istatistiksel yöntemlerde olduğu üzere faktör analizi için de analiz öncesinde sorgulanması gereken hususların en başında verilerin hangi ölçek düzeyinde toplandığı gelmektedir. Sürekli değişkenlerle faktör analizi için kuramsal dayanaklar ve kullanılan yöntemler oldukça gelişmiştir. Ancak pratikte, gözlenen ve/veya ölçülen değişkenlerin bir kısmı sıralama ölçeği düzeyindedir ve bu durum sıklıkla göz ardı edilmekte ve doğru olmayan bir şekilde 1, 2, 3, 4 gibi sıralı kategorileri temsil eden sayılara metrik özelliğe sahipmiş gibi davranılmaktadır. Özellikle sosyal bilimler kapsamındaki tüm alanlarda oldukça sık kullanılmakta olan ölçeklerin (likert tipi vb.) madde cevap formatlarından kaynaklı olarak, bu ölçeklerden elde edilen verilerin eşit aralıklı ölçek düzeyinde kabul edilerek buna uygun istatistiksel tekniklerin kullanılması, önceden beri eleştirile gelmiş bir durumdur. Bu eleştirileri ortaya koyan araştırmacılar söz konusu ölçeklerden elde edilen verilerin sıralama ölçeği düzeyinde olduklarını ve bu veriler için uygun istatistiksel tekniklerin kullanılması gerektiğini belirtmektedirler. Sıralama ölçeği düzeyinde değişkenler söz konusu olduğunda bu değişkenler arasındaki ilişkinin tahmin edilmesinde ya da bu değişkenler kullanılarak yapılacak ilişki temelli istatistiksel analizlerde "tetrakorik" (iki kategorili veriler için) ya da "polikorik" (üç veya daha çok kategorili veriler için) korelasyon matrisleri kullanılması gerekliliği vurgulanmaktadır. Faktör analizi kapsamında en tutarlı, en sağlıklı ve en güçlü kestirimlerin polikorik korelasyon matrisi kullanılarak yapılabildiği ortaya konulmuştur. Ancak ölçek geliştirme ve uyarlama süreçlerinde kullanılan hem keşfedici hem de doğrulayıcı faktör analizinin uygulanması sırasında genellikle Pearson korelasyon matrisinden faydalanılmaktadır. Ölçme aracının yapı geçerliğinin test edilmesi amacıyla faktör analizi kullanılacağı zaman, verilerin toplandığı ölçek düzeyinin göz önünde bulundurulması büyük önem taşımaktadır. Diğer taraftan analizlerde Pearson korelasyon matrisinin sıklıkla kullanılması, araştırmacıların konuyla ilgili yeterli bilgiye sahip olmamasından ya da kullanmakta oldukları bilgisayar paket programlarının sınırlılığından kaynaklanmaktadır.

Bu araştırmanın amacı, sosyal bilimlerde geliştirilen likert türü ölçeklere, ölçek geliştirme sürecinde yapı geçerliği kanıtı sağlamak amacıyla uygulanan faktör analizi kapsamında Pearson ya da polikorik korelasyon matrisi temelli analizin gerçekleştirilmesinin, analiz sonuçları üzerindeki etkisini karşılaştırmaktır.

Faktör analizinde Pearson ya da polikorik korelasyon matrisi kullanmanın analiz sonuçları üzerindeki etkisini incelemeyi ve analiz sonuçlarını karşılaştırmayı temel alan bu araştırma betimsel türde bir araştırmadır. Betimsel araştırmalar ilişkiyi ya da farkı merak etmeyen, neyin ne olduğunu saptamaya dönük çalışmalardır. Bu bakımdan betimsel çalışmalar, bilimin betimleme amacına hizmet etmekte ve aynı zamanda sonraki araştırmalar için denence üretmeye yönelik öngörü sağlarlar.

Sosyal ve davranış bilimleri alanlarında yapılan çalışmalarda ölçümlere sıklıkla rastgele ve/veya sistematik hatalar karışmakta, buna bağlı olarak ulaşılan sonuçlara da bu sorunlar yansımaktadır. Özellikle bilimsel araştırmalarda veri toplama aracı olarak kullanılmak üzere geliştirilen veya farklı bir kültürden bir diğerine uyarlanan Likert tipi ölçeklerin geçerlik ve güvenirlik kanıtları toplanırken yürütülen analizlerde, farklı korelasyon matrisi kullanımının sonuçlara karışacak hatayı artırabileceği ve bunun da en başta bir çok araştırmacı tarafından geçerlik türlerinin en önemlisi olarak görülen yapı geçerliği kanıtlarının yanlış yorumlanmasına yol açabilecektir. Dahası bu durum, söz konusu ölçekler kullanılarak yürütülecek sonraki çalışmaların sonuçlarına da kaçınılmaz olarak yansıyacaktır.

Daha önce yapılan ve simülatif verilerle yürütülen çalışmalarda olduğu gibi, gerçek verilerin kullanıldığı bu çalışmada da farklı korelasyon matrisi temelli faktör analizi sonuçlarının birbirinden farklılaştığı görülmüştür. Daha önce yapılan ve simülatif verilerle yürütülen çalışmalarda (Holgado-Tello ve diğerleri, 2010) olduğu gibi, gerçek verilerin kullanıldığı bu çalışmada da farklı korelasyon matrisi temelli faktör analizi sonuçlarının birbirinden farklılaştığı görülmüştür. Bu çalışma kapsamında ulaşılan bulgular şu şekilde özetlenebilir: Polikorik ya da Pearson korelasyon matrisi kullanılarak yürütülen analizler sonucunda,

1) Maddeler farklı "madde toplam korelasyonu" ve "yük değeri" göstermiştir.
2) Maddeler farklı yönde (negatif ya da pozitif) korelasyon değeri gösterebilmiştir.
3) Farklı faktör sayıları ortaya çıkmıştır.
4) Ortaya çıkan alt boyutların (faktör) toplanabilir olup olmama durumları değişmiştir.
5) Farklı maddeler ölçek dışı bırakılmıştır.
6) Test edilen yapının doğrulanıp doğrulanmama durumu değişmiştir.
7) Maximum likelihood ve unweighted least square yöntemi ile yapılan tahminlemelerin birbirine çok yakın uyum indeksi sonuçları verdiği belirlenmiştir.

TURJE
www.turje.org

Volume: 8 Issue: 3

Turkish Journal of Education

DOI: 10.19128/turje.453383
Research Article

Received 14.08.2018
Accepted 13.07.2019

Implications of between-school tracking for Turkish students

Wenke Niehues
Koç University, Department of Psychology, Istanbul, Turkey, wniehues14@ku.edu.tr
Leibniz Institute for Educational Trajectories, Bamberg, Germany, wenke.niehues@lifbi.de
Yasemin Kisbu-Sakarya
Koç University, Department of Psychology, Istanbul, Turkey, ykisbu@ku.edu.tr
Bilge Selçuk
Koç University, Department of Psychology, Istanbul, Turkey, bselcuk@ku.edu.tr

ABSTRACT Previous multilevel analyses for Turkey show that performance differences of students vary more between schools than within schools. These school-disparities might be associated with Turkey's tracking system and related differences in student body and learning environments of school tracks. Since it is not known how Turkey's low-performing vocational, low-performing academic, and high-performing academic school tracks differ regarding students' family background, motivational and behavioral engagement of students, and schools' learning environments, we analyzed the PISA 2012 data to examine these differences. Results indicate that Turkish students which attend high-performing academic schools are more likely to have higher socio-economic status, display higher confidence in their math ability, are less engaged during class and are exposed to a richer learning environment than students attending low-performing academic schools. Policy implications of each finding are discussed in detail.

Keywords: *Between-school tracking, Socio-economic status (SES), Turkish students, Motivational and behavioral engagement, Learning environment*

Türk öğrenciler için okullar arası izleme uygulamaları

ÖZ Çok düzeyli analizler Türkiye'deki okullar arası öğrenci performansı farklılıklarının okul içi performans farklılıklarından daha fazla olduğunu göstermiştir. Bu durum, okullara giriş sistemi ve buna bağlı olarak öğrenci profillerindeki ve de okulların öğrenme ortamlarındaki farklılıklardan kaynaklanabilmektedir. Türkiye'deki düşük performanslı meslek okullarına, düşük performanslı akademik okullara ve yüksek performanslı akademik okullara devam eden öğrencilerin aile geçmişleri, motivasyonel ve davranışsal katılımları ve okulların öğrenme ortamları arasındaki farklar yeteri kadar incelenmemiş olduğundan, bu çalışmada PISA 2012 verisi bu farklılıkları tespit etme amacı ile analiz edilmiştir. Sonuçlar, düşük performanslı akademik okullara giden öğrencilere kıyasla, yüksek performanslı akademik okullardaki Türk öğrencilerinin daha yüksek sosyo-ekonomik statüye sahip olduklarını, matematik becerilerine daha çok güvendiklerini, ders sırasında daha az katılım gösterdiklerini ve daha zengin bir öğrenme ortamına maruz kaldıklarını göstermiştir. Bulgular eğitim politikaları kapsamında tartışılmıştır.

Anahtar Kelimeler: *Okullar arası izleme, Sosyo-ekonomik durum (SED), Türk öğrenciler, Motivasyonel ve davranışsal ilişki, Öğrenme ortamı*

Citation: Niehues, W., Kisbu-Sakarya, Y., & Selcuk, B. (2019). Implications of between-school tracking for Turkish students. *Turkish Journal of Education, 8*(3), 196-216. DOI: 10.19128/turje.453383

INTRODUCTION

Inequality in education is perceived as legitimate in meritocratic societies as long as there is equity in education. Equity in education is present when the academic success of an individual is not associated with their social origin, but with their effort and skills. While many nations aim for equity within their education systems, large-scale international studies show that countries differ greatly in their ability to obtain this goal (OECD, 2013; Özel, Özel, & Thompson, 2013). Turkey, for example, is a developing country with a fast changing and expanding education system which aims to provide high quality formal education for its growing young population (Özdemir, 2016). While the upper-secondary school (i.e., high school) attendance rate in Turkey has increased within the last decade, Turkey is struggling with large performance and socioeconomic status (SES) differences between its upper-secondary schools and students, as well as overall low performance levels in reading, math, and science compared to other OECD countries (Dincer & Uysal, 2010; OECD, 2013). This implies that the current education system in Turkey is not able to ensure equity and high quality education for its whole student population (Özel et al., 2013).

The high performance and SES disparities among Turkish schools are likely to be associated with Turkey's tracking system at the transition to upper-secondary school and college (Maaz, Trautwein, Lüdtke, & Baumert, 2008; Özdemir, 2016). Turkish students compete in a national standardized exam for the admission to a limited number of high-performing upper-secondary schools and colleges (Yavuz, 2009). This selection practice has led to a highly stratified and elite school system within Turkey, in which average school SES and performance level have been associated with Turkish students' general academic performance (Alacaci & Erbas, 2010) and performance on the even more important national university entrance exam (Caner & Okten, 2013). In light of these long term implications of school track attendance for students' later academic career, there is lots of pressure on Turkish students to gain access to one of the scarce prestigious upper-secondary schools in Turkey.

While SES and performance differences between upper-secondary schools in Turkey are well established, until today there is very limited knowledge about how school track attendance in Turkey relates to a wider range of student and school features. The large performance and SES differences between students and schools imply that school track attendance in Turkey does not only relate to students' academic performance, but also to further factors such as SES and possibly also students' motivation, engagement and learning environment at school. To our knowledge, no study yet has examined whether students coming from economically more advantaged families are more likely to attend high-performing schools in Turkey and what role parents' education-related beliefs play in the choosing of upper-secondary school tracks. Moreover, we do not know whether students at high-performing schools are more motivated and exceed more effort during their classes at school. We also do not know how learning environments differ between school tracks in Turkey. Do students at high-performing schools get offered more learning opportunities in terms of a richer offer of extra-curricular activities and a better disciplinary climate during class? Or do high-performing schools have more resources at their disposal to meet the needs of their student body?

Examining differences in school track attendance will help to identify reasons for the large performance differences between schools in Turkey. This is important in order to formulate future policy approaches that aim to dampen performance disparities between schools, and lift academic performance standards particularly for students attending low-performing schools. Furthermore, study results may help to identify which differences between school tracks are important for students' learning in Turkey and which features should be taken into account in future studies examining individual differences in academic achievement of students. Since results help to better understand how students and parents navigate in a highly selective school systems and illustrate the consequences for schools and students of establishing a strict between-school tracking system, findings might also be important for other school systems with similar tracking policies such as in Germany, Austria, Belgium, or France.

The following section gives information about the Turkish education system and the transition process from lower-secondary (i.e., middle school/Ortaokul) to upper-secondary (i.e., high school/lise) school in Turkey. This is followed by a literature review of how family background, students' motivational and behavioral engagement, as well as schools' learning environments may relate to students' school track attendance and choice.

Education in Turkey

Turkey's education system underwent several reforms in the last two decades. One major structural reform was the introduction of the so called "4+4+4"-system in 2012. The "4+4+4"-reform prolonged compulsory education from eight to 12 years in Turkey, since then Turkish students are expected to attend four years each of primary, lower-secondary, and upper-secondary school education. In the public school sector, primary and lower-secondary school admittance takes place according to students' place of residence. At the lower-secondary level students may choose between general and religious schools, and at the upper-secondary level students may choose between different types of schools such as vocational schools with specializations in electricity, accounting, tourism, religion or others and academic-oriented schools such as science, social science, or general upper-secondary schools (Özdemir, 2016). Vocational schools focus on applied skills and assume that their students will start working after graduation. Academic-oriented schools prepare their students for the transition to college with a higher number of academic-oriented classes (Özdemir, 2016). Performance levels between schools differ, with higher performing students often attending academic-oriented schools.

The number of academic-oriented, high-performing, upper-secondary schools is limited in Turkey. Students compete in national standardized tests for the admittance to their school of choice (Özdemir, 2016). If students and parents decide not to take the admission exam or do not obtain sufficient points during the exam for their school of choice, then they get centrally distributed to non-selective schools. These non-selective schools are often vocational or general schools with lower performance levels (Özdemir, 2016). As such, Turkey employs a strict between-school tracking policy in which students get sorted based on their prior academic performance. National high-stake testing also takes place at the transition to college in Turkey and test results are accompanied by a nationwide ranking of students.

Family Background and Transition to Secondary School

Research indicates that SES-related disparities in education systems get amplified at transition points within school systems (Schnabel, Alfeld, Eccles, Köller, & Baumert, 2002). The underlying mechanisms can be traced back to (a) SES-related differences in educational decisions within families and (b) to SES-related differences in students' academic achievement, motivation, and engagement (Maaz et al., 2008).

Studies using the Wisconsin-Modell for status attainment confirm that SES relates to parents' educational decision making. These studies found that lower-SES parents are more employment- and less academically-oriented; thus they tend to send their children to vocational instead of academic schools (Becker, 2010). Additional, research found that higher-SES parents get more actively involved in their child's school choice (Groos, 2016), they possess more knowledge about the education system (Weininger & Lareau, 2003), and they contact teachers more frequently while organizing their child's transition to secondary school (Kleine, 2014). Hence, higher-SES parents seem to be better equipped to manage and ensure a positive outcome of the transition process to secondary school for their children.

Moreover, research found that academic achievement relates to students' SES (Alacaci & Erbas, 2010). Thus, in school systems with between-school tracking based on prior achievement, particularly higher performing, socio-economically advantaged kids are more likely to gain access to more selective schools

(Özdemir, 2016). Additionally, according to the widely used Expectancy-Value-Theory (EVT), a reason why family background relates to students' achievement level is that students develop their education-related beliefs and in turn their academic performance within their sociocultural milieu (Simpkins, Fredericks, & Eccles, 2015). That is, EVT proposes that students coming from more advantaged families will have parents who value academic achievement more. During the socialization process, children are likely to internalize these positive parental education-related beliefs, which in turn is supposed to foster children's academic engagement, achievement and choices, so that these students are more likely to work hard for and apply to more challenging and advanced academic classes and schools (Simpkins et al., 2015).

The positive link between students' motivational beliefs such as value assigned to academic tasks (i.e., academic task value beliefs) and beliefs about one's ability to successfully complete a task (i.e., academic competence beliefs) and students' behavioral engagement and academic achievement as well as choice is well established by EVT-studies (Eccles & Wigfield, 2002). Guo and colleagues (2015) found, for example, that Australian youth who assigned higher value to math and were more confident about their math ability were more likely to choose advanced math classes at upper-secondary school and study Science, Technology, Engineering, and Mathematics (STEM) related subjects at college.

Overall these findings suggest that family background and students' motivational and behavioral engagement patterns will positively relate to students' school track attendance such that students coming from higher-SES families and families which value academic achievement more for their children, as well as more motivationally and behaviorally engaged students will attend more academic-oriented and higher performing schools.

Student Body Characteristics Associated with School Track Attendance

The previous section discussed the possibility that family background is, directly or indirectly via students' motivational beliefs, engagement and performance, positively associated with students' school track attendance. However, it is also possible that, besides family background, school factors shape students' motivational beliefs and engagement. This might be particularly true in highly selective school systems, since a strict between-school tracking policy is likely to foster homogeneous learning groups with similar characteristics of the student and teacher body within schools (Maaz et al., 2008). Studies found, for example, higher levels of disruptive behaviors in schools with a concentration of disadvantaged students. These higher levels of disruptive behaviors negatively reflected on behavioral engagement patterns and the motivation of fellow classmates in these studies (Murphy, 2010; Thomas, Hierman, Thompson, & Powers, 2008). Moreover, at low-performing schools compared to high-performing schools, teachers are found to hold lower academic expectations for their students (Gamoran, 2004). Students from low-performing school tracks may adopt these lower performance standards of their teachers and schools, and thus display lower academic expectations and effort (Gamoran, 2004).

Despite this, studies investigating the "Big Fish Little Pond Effect" found that the average performance level of the learning group at school has a dampening effect on students' competence beliefs (Marsh & Hau, 2003). That is, despite comparable academic ability, students in better performing learning groups hold lower competence beliefs in math due to social comparison mechanism with their peers (Marsh & Hau, 2003). However, in highly selective school systems, such as in Turkey, for their social comparison with their peers, students may not refer to their close learning group at school, but rather to their nationwide standing. Turkish students may have information about their nationwide standing due to their school track admittance and ranking on the national admittance test for upper-secondary school. This would be in line with findings from Mann, Legewie, and DiPrete (2015) that, in selective school systems, students' competence beliefs are related to their school admission. Thus, overall it is likely that Turkish students' motivational and behavioral engagement are associated with their school track attendance over and above their SES and parental beliefs.

School Learning Environment and School Track Attendance

Moreover, school tracks are not only likely to differ in terms of characteristics of their student body, but also in terms of the learning environment that they offer to their students. This is important since differences in learning environments are associated with the academic development of students (Maaz et al., 2008). Academically oriented, high-performing schools are supposed to offer a more challenging and cognitively stimulating learning environment to their students than lower-performing or vocational schools (OECD, 2012). In comparison to vocational schools, academic schools implement a more challenging academic curriculum which may result in a higher number of academically oriented classes or extra-curricular activities at school and thus, foster greater familiarization with more demanding academic tasks (Giersch, 2016). Additionally, in high-performing schools, the disciplinary climate during classes is found to be greater than in low-performing schools, which resulted in prolonged instructional time and thus prolonged practice time for students at high-performing schools (Murphy, 2010). Furthermore, research indicates that low-performing schools are less well-equipped than high-performing schools (Roeser, Urdan, & Stephens, 2009). A lack of resources such as deprived facilities, insufficient heating, cooling or number of classrooms is found to be negatively associated with students' achievement levels when it is below a minimum standard (Eccles & Roeser, 2011). In sum, these findings suggest that students may benefit academically from their attendance of academic and high-performing schools due to enhanced learning environments at these schools.

Present Study

Large-scale international data indicates that Turkey is one of the few OECD countries in which the achievement gap between schools is greater than within schools, and in which schools differ greatly in the SES-composition of their student body (OECD, 2013). This puts a threat to equity in Turkey. These disparities among schools are likely to be reinforced by Turkey's selective school system such as its between-school tracking and national high-stake testing policy (Maaz et al., 2008; Özdemir, 2016). Furthermore, the large performance and SES-differences between schools imply that school track attendance in Turkey may not only relate to Turkish students' academic performance, but also to further student characteristics such as SES or possibly also other student and school factors such as student motivation or engagement as well as schools' learning environments.

While performance and SES differences between schools are well researched in Turkey, it is not well understood how school track attendance in Turkey is associated with a wider range of student body and school factors. To our knowledge, no study in Turkey has examined how school track attendance relates to further student body (i.e., math achievement, SES, parental valuing of math, students' motivational and behavioral engagement in math) and school (i.e., disciplinary climate during math classes, extra-curricular activities offered at school, quality of school facilities) characteristics yet. Thus, in order to address this gap in the literature, our study focused on features related to school track attendance in Turkey. Our study did not aim to explain individual differences in Turkish students' academic achievement as such, yet study results may help to identify influential school factors associated with Turkish students learning. Hence, study results may provide information to future studies which aim to examine individual differences in Turkish students' academic achievement by including school factors in their analyses. To our knowledge, studies which include school besides individual factors while explaining individual differences in Turkish students' academic achievement are still very limited.

Overall, the main aim of our study was to better understand the implications of Turkey's tracking policy in order to identify much needed policies that might help to dampen its negative effects (Özel et al., 2013). More specifically, we wanted to learn what kind of students are likely to attend which upper-secondary school track. This would provide us with information about which students are likely to prevail and which students are likely to stay behind in the tough competition for the limited seats at academically successful upper-secondary schools in Turkey. Additionally, we wanted to investigate how students' motivation and engagement levels differ between school tracks, and whether students at high performing schools are more likely to be motivated and engaged as implied by EVT. This would inform us about whether future policies in Turkey should focus on individual factors such as elevating students' motivation and engagement levels in order to combat educational inequalities. Moreover, we wanted to learn how the learning environment between school tracks differ in Turkey, in order to comprehend what kind of learning context and opportunities are offered to students at each attended school track. This information might hold important information for future studies assessing which school factors relate to students' academic achievement in highly selective school systems such as the one found in Turkey. Since the literature indicates that students' academic choices, motivation, and engagement changes with the subject domain, we only focused on one subject in our study, namely mathematics (Guo et al., 2015).

Based on the literature reviewed above and in regard to family and student body characteristics, we hypothesized that students who come from socio-economically more advantaged families, students with parents who value math more, and students who are motivationally and behaviorally more engaged during their math classes would be more likely to attend high-performing and academic-oriented schools than low-performing and vocational schools. In regard to school characteristics, we predicted that classroom-climates in math would be more disciplined, extra-curricular math activities would be offered more, and school facilities would be better at high-performing and academic-oriented schools compared to low-performing and vocational schools.

METHODOLOGY

Participants

Participants were drawn from the Turkish PISA 2012 data-set collected by the OECD. The data is cross-sectional, designed to be representative of 15-year-old students in Turkey, and has a hierarchical structure with students nested within schools. We excluded schools with 10 or fewer students in order to prevent estimation errors for nested data (McNeish & Stapleton, 2016; see Appendix 1 for results of main analyses including all study participants). This resulted in a final sample of 4,742 Turkish students from 152 schools with a mean age of 15.8 years (*SD* = .28) and 49% female participants.

Measures

All measures were collected by the PISA 2012 consortium and were either student or school principal reports. Most scale scores in the data-set (except the categorization into school tracks and parental education-related beliefs) were computed, coded, internationally validated and tested for reliability by the PISA consortium (for more information on the scales please see OECD, 2014). Since the focal point of analysis for PISA 2012 was students' math achievement, the majority of measures were math-specific which was in line with our focus on the math domain.

School Tracks

Based on two measures we categorized schools into four tracks: (1) school's curriculum-orientation (academic-versus vocational-oriented curriculum) and (2) school's average proficiency level in math (low- versus high-proficiency level in math). School principals reported on the curriculum-orientation of their school by indicating whether their school followed a vocational or academic curriculum. Schools with a vocational curriculum emphasized applied, employment-related skills in their teaching. Schools that followed an academic curriculum focused on academic skills.

Based on the performance of students during a standardized math test, the PISA consortium distinguished between seven proficiency levels ranging from 0 to 6 for students' math achievement. The

PISA consortium defined the necessary points on the standardized math test for each proficiency level (OECD, 2014). According to this definition, until level 2 (420.1 to 482.3 points on the standardized math test), students are only capable of making a literal interpretation of math tasks, and starting from level 3 (482.4 to 544.7 points on the standardized math test) students are able to reason and apply problem-solving strategies to math tasks (OECD, 2014). Thus, beginning with level 3 students are able to autonomously apply their math skills to solve basic everyday math tasks. In order to identify schools in which students on average possessed these applied problem-solving skills, in our study we classified schools with an average student performance level in math at or below level 2 as “low-performing”, and schools with an average student performance level above level 2 as “high-performing schools”. Hence, schools with an average score of 482.3 or lower were classified as low-performing schools, and schools with an average score of 482.4 or higher on PISA’s standardized math test were classified as high-performing schools. Since no student in our sample attended high-performing vocational schools, in this study only three school tracks were investigated: Low-performing vocational schools (n = 1,859) (hereafter referred to as vocational schools), low-performing academic schools (n = 1,630), high-performing academic schools (n = 1,253).

Math Achievement

Students took part in a standardized math test. To scale students’ test scores, PISA used a Rasch model. To account for any uncertainty during the Rasch scaling process, PISA provided five estimates (i.e., five plausible values) for each students’ math achievement instead of a single-point estimate. Any analysis including students’ math achievement had to combine estimations across all five plausible values (see OECD, 2009 for information on the use of plausible values). The PISA consortium normed students’ math achievement with 500 points corresponding to the OECD average (SD = 100; OECD, 2014).

Family Background

Family background was measured via students’ SES and parents’ education-related math beliefs. Students’ SES was measured by an index including parents’ education level, parents’ occupational status, and household possessions. The three indices were combined by a principal component analysis by the PISA consortium with an OECD average of 0 (SD = 1; OECD, 2014).

Parents’ education-related math beliefs were measured via students’ perception of their parents’ math norms. On three items, using a 4-point Likert scale (1 = strongly agree, 4 = strongly disagree), students indicated how important their math achievement is for their parents (e.g., “My parents believe it is important for me to study mathematics.”). Our analysis on the Turkish data set revealed acceptable reliability (α = .73) for the three items measuring parents’ education-related beliefs, thus the three standardized items were averaged to obtain parents’ education-related beliefs. We reverse coded the three items so that higher scores indicate more positive parental beliefs.

Motivational Engagement

Students’ motivational engagement was measured via students’ task value and academic ability beliefs as suggested by EVT (Eccles & Wigfield, 2002). Students’ valuing of academic tasks was operationalized via students’ instrumental motivation (i.e., utility value) for math and students’ interest and enjoyment in math. To measure students’ instrumental motivation, students reported on four items (e.g., “Making an effort in mathematics is worth it because it will help me in the work that I want to do later on.”) whether they think that math achievement is important for their future career. To capture students’ interest and enjoyment in math, students specified how much interested they are in math (e.g., “I am interested in the things I learn in mathematics.”) on three items. Both measures used a 4-point Likert scale (1 = strongly agree, 4 = strongly disagree), were Rasch scaled by the PISA consortium with an OECD average of 0 (SD = 1), and had good reliability in our sample (α = .87, α = .89, respectively).

Students' math ability beliefs were operationalized via students' math self-concept and math self-efficacy. While self-concept tapped into students' self-evaluation about their general academic capability in math, students' self-efficacy beliefs referred to their task-specific competence beliefs in math. Since the two concepts tapped into different aspects of students' ability beliefs, both measures were included in the study. Students' math self-concept was captured via five items (e.g., "I learn mathematics quickly.") on a 4-point Likert scale (1 = strongly agree, 4 = strongly disagree). To measure math self-efficacy, students reported their feelings of competence on carrying out eight different math tasks such as "calculating how much cheaper a TV would be after a 30% discount" on a 4-point Likert scale (1 = very confident, 4 = not at all confident). Both indices were Rasch scaled by the PISA consortium with an OECD average of 0 (*SD* = 1) and had good reliability for the Turkish data (α = .84, α = .82, respectively).

Behavioral Engagement

The behavioral engagement was measured via students' engagement during math classes and students' openness to problem-solving. To assess students' engagement during math classes, students reported on nine items (e.g., "I pay attention in mathematics class.") whether they are attentive during math classes and complete their class work. A 4-point Likert scale (1 = strongly agree, 4 = strongly disagree) was used for this measure. To measure students' openness to problem-solving, students reported on five items their effort and persistence during tasks (e.g., "When confronted with a problem I give up easily."). A 5-point Likert scale (1 = very much like me, 5 = not at all like me) was used for this measure. Both measures were Rasch scaled by the PISA consortium with an OECD average of 0 (*SD* = 1) and had good reliability in our sample (α = .91, α = .78, respectively).

Learning Environment at School

Three variables measured schools' learning environments: Disciplinary climate during math classes, math-related extra-curricular activities at school, and quality of school facilities. The disciplinary climate during math classes tapped into the quality of the learning environments at the classroom level, while the offer of extra-curricular activities and the quality of school facilities were indices at the school level.

To capture the disciplinary class-climate, students indicated on five items (e.g., "Students don't start working for a long time after the lesson begins.") the degree to which classmates displayed disruptive behaviors during math classes. The index used a 4-point Likert scale (1 = every lesson, 4 = never or hardly ever), was Rasch scaled by the PISA consortium with an OECD average of 0 (*SD* = 1), and had a good reliability in the Turkish data (α = .86).

To assess schools' extra-curricular math activities, school principals reported on four items whether the school offered additional math classes, math/computer clubs, or participation in math competitions. Scores were combined into a composite score of extra-curricular math activities offered at school by the PISA consortium, with higher scale scores indicating an enriched extra-curricular math-related school environment (OECD, 2014). The quality of school facilities was measured by the availability of school building and grounds, heating/cooling and lightening system, and classroom space. School principals evaluated on a 4-point Likert scale (1= not at all, 4 = a lot) and three items (e.g., "Is your school's capacity to provide instructions hindered by any of the following issues? Shortage or inadequacy of school buildings and grounds.") whether their school's facilities are sufficient. The index was Rasch scaled by the PISA consortium with an OECD average of 0 (*SD* = 1) and had an acceptable reliability for the Turkish data (α = .75).

Analytic Strategy

The preliminary analysis examined whether there were mean differences on study variables across school tracks, using low performing academic schools as the reference category. Following recommendations of the PISA consortium, regressions were conducted with 80 replicate weights using SPSS 22 (OECD, 2009).

In the main analysis, we conducted a multinomial logistic regression via Mplus 7.4 (Muthen & Muthen, 1998-2012) to examine the role of family background, students' motivational and behavioral engagement, and experienced learning environments for school track attendance in Turkey. We incorporated student weights into our analyses to account for the sampling error (OECD, 2014) and used the Mplus "type = complex" option with school ID as the cluster variable in order to account for the hierarchical data structure (i.e., students nested within schools). Thus, we did not estimate a two level model in order to account for the nested data structure, but with using the "type = complex" option of Mplus we employed a design-based approach to correct for reduced standard errors due to the nested data structure (Stapleton, McNeish, & Seung Yang, 2016). The design-based approach as described by Stapleton et al. 2016 allowed us to estimate a model only at the student level while taking the hierarchical data structure into account. This approach is in line with our research question and aim to identify differences in individual students' school track attendance (Stapleton et al., 2016). We chose this approach, since with the multinomial logistic regression we wanted to explain what kind of students are likely to attend which school track and what kind of learning environment students are likely to experience at each school track. Thus, during the multinomial logistic regression, we were not interested in between school differences as such, therefore we did not model them. A logit function and the robust maximum likelihood estimator (MLR) was used for model estimation. For model assessment a robust likelihood-ratio test, the Bayesian information criterion (BIC), McFadden's R^2 and a classification table was used (Agresti, 2007; Satorra & Bentler, 1999; Tabachnick & Fidell, 2007). The equations for the logistic regression were as follows:

$$ln\left(\frac{P\ (vocational\ school\ track)}{P\ (low-performing\ academic\ school\ track)}\right) =$$

b100 + b101(SES) + b102(parent math beliefs) + b103(student instrumental motivation) + b104(student interest in math) + b105(student math self-concept) + b106(student math self-efficacy) + b107(student behavioral engagement in math classes) + b108(student openness to problem solving) + b109(disciplinary climate during math classes) + b120(extra-curricular math activities) + b121(quality of school facilities)

$$ln\left(\frac{P\ (high-performing\ academic\ school\ track)}{P\ (low-performing\ academic\ school\ track)}\right) =$$

b200 + b201(SES) + b202(parent math beliefs) + b203(student instrumental motivation) + b204(student interest in math) + b205(student math self-concept) + b206(student math self-efficacy) + b207(student behavioral engagement in math classes) + b208(student openness to problem solving) + b209(disciplinary climate during math classes) + b210(extra-curricular math activities) + b211(quality of school facilities)

When collecting data, PISA 2012 used a rotated design for their student questionnaire in order to increase the covered content by the student questionnaire (OECD, 2014). That means that three different booklets with varying items were randomly distributed to students and therefore, most missing data in the data-set was missing by design and missing completely at random (MCAR) (Graham, 2009). Checking for item-level missing data revealed that less than 2% of data was missing on each item. This data can be assumed to be missing at random (MAR) and thus, does not pose any threat to our parameter estimation (Graham, 2009).

RESULTS

Preliminary Analysis

On average Turkish students performed 51 points less than the OECD average on the standardized math test (Table 1). Since the PISA 2012 consortium estimated that on average 41 points in math achievement referred to a learning progress within one school year, Turkish students were behind for over one year of schooling compared to the OECD average (OECD, 2013). Additionally, the SES level in our sample was low. With a mean of -1.44, mean SES of Turkish students was close to one and a half standard deviations below the OECD mean. Turkish students' value beliefs (ranging from M = -.05 to M = .05) were comparable to the value beliefs held by the average OECD student, except for Turkish students' higher interest in math. Turkish students were close to half a standard deviation more interested in math (M = .42), and they also reported higher average engagement levels at school (M = .21 and M = .24) compared to their peers from OECD countries. Regarding the learning environment at school, Turkish students reported a slightly lower level of disciplinary classroom climate (M = -.09) than the average OECD student. And Turkish school principals indicated poorer quality of school facilities (M = -.24) compared to the OECD average.

Table 1.
Descriptive statistics for study variables

	N	*M*	*SD*	Min	Max
Student math achievement	4,783	449	90.9	175	795
SES	4,742	-1.44	1.09	-4.61	1.94
Parent math beliefs	3,123	0.01	0.80	-2.73	1.14
Student Value Beliefs					
Student instrumental motivation	3,136	0.05	0.99	-2.30	1.59
Student interest in math	3,141	0.42	1.06	-1.78	2.29
Student Ability Beliefs					
Student math self-concept	3,141	-0.05	0.97	-2.18	2.26
Student math self-efficacy	3,146	-0.02	0.93	-3.75	2.27
Student Behavioral Engagement					
Student behavioral engagement in math classes	3,146	0.24	1.12	-3.45	2.72
Student openness to problem solving	3,151	0.21	0.95	-3.63	2.45
School Learning Environment					
Disciplinary climate during math classes	3,137	-0.09	0.91	-2.48	1.85
Extra-curricular math activities offered at school	4,783	1.75	1.32	0.00	5.00
Quality of school facilities	4,783	-0.24	0.97	-2.76	1.31

Preliminary analysis comparing means across school tracks in Turkey revealed more favorable student body characteristics and learning environments for high-performing academic compared to low-performing academic schools (Table 2). On average, students attending high-performing academic schools had significantly higher math achievement, came from more advantaged families, were more motivated and behaviorally more engaged, had parents with higher math beliefs, and reported more favorable learning environments at school than students attending low-performing academic schools. Students from high-performing academic schools in Turkey did not significantly differ from students from low-performing schools in their interest in math and behavioral engagement during math classes (Table 2).

Low-performing academic schools did not differ from vocational schools in terms of students' math achievement, SES, instrumental math motivation, and extra-curricular school activities. Students attending low-performing academic schools in Turkey reported higher interest in math, higher ability beliefs, higher behavioral engagement, higher disciplinary class-climate, better-equipped school facilities, and their parents had more positive school-related beliefs compared to those attending vocational schools (Table 2).

Table 2.
Means and mean differences across school tracks with low-performing-academic school track as the reference category

	Low-performing academic track (n = 1,630)	Vocational track (n = 1,859)			High-performing academic track (n = 1,253)		
	M(SD)	*M(SD)*	$M_{difference}$	95% CI	*M(SD)*	$M_{difference}$	95% CI
Student math achievement	415(65.0)	409(64.6)	-5.51	[-15.3,4.32]	555(66.9)	141*	[125,156]
SES	-1.65(1.03)	-1.70(.90)	-.05	[-.18,.08]	-.77(1.17)	.88*	[.71,1.06]
Parent math beliefs	.00(.84)	-.10(.82)	-.10*	[-.17,-.02]	.18(.69)	.18*	[.09,.27]
Student Value Beliefs							
Student instrumental motivation	.04(1.00)	-.02(.95)	-.05	[-.14,.03]	.16(1.02)	.13*	[.02,.23]
Student interest in math	.47(1.10)	.32(1.04)	-.15*	[-.25,-.04]	.53(1.01)	.07	[-.04,.18]
Student Ability Beliefs							
Student math self-concept	-.09(.98)	-.18(.92)	-.10*	[-.18,-.02]	.18(.98)	.26*	[.15,.37]
Student math self-efficacy	-.16(.86)	-.28(.88)	-.12*	[-.20,-.04]	.53(.86)	.69*	[.57,.81]
Student Behavioral Engagement							
Student behavioral engagement in math classes	.30(1.16)	.15(1.15)	-.15*	[-.20,-.05]	.32(1.01)	.02	[-.09,.13]
Student openness to problem solving	.22(.98)	.08(.99)	-.15*	[-.22,-.07]	.39(.83)	.16*	[.07,.27]
School Learning Environment							
Disciplinary climate during math classes	-.16(.89)	-.25(.91)	-.09*	[-.18,-.00]	.24(.86)	.40*	[.28,.52]
Extra-curricular math activities offered at school	1.48(1.18)	1.46(1.09)	-.02	[-.51,.47]	2.58(1.47)	1.10*	[.57,1.63]
Quality of school facilities	-.17(.90)	-.60(.96)	-.43*	[-.78,-.07]	.19(.89)	.36*	[.05,.68]

Note. Results of regression analyses with 80 replicate weights. $M_{difference}$ = Difference between means for vocational/high-performing-academic track versus low-performing-academic track. 95% CI = 95% Confidence interval for the mean difference. *p < .05.

Main Analysis

In preparation for the logistic regression, we tested for multi-collinearity, screened for possible outliers and assessed linearity. Across all independent variables students' interest in math (tolerance = 0.29, VIF = 3.44) had the lowest tolerance and highest variance inflation factor (VIF). Since none of the tolerance values were below 0.25 and none of the VIFs were above 5.00, we concluded that no problematic multi-collinearity issues existed (Urban & Mayerl, 2006). The screening for outliers revealed that nine students reported lower math self-efficacy beliefs and additionally five students reported lower openness to problem solving scores than 3.29 standard deviations units away from the mean of the respective variables. Since for these students no unusual response pattern on further study variables could be detected, we refrained from deleting these cases from our dataset. The linearity in the logit was assessed with the Box-Tidwell (1962) procedure. Results indicated that only the relationship between the quality of the school building and the logit transformation of the dependent variable might not be linear. Yet, since alternative modelling of the relationship between the quality of the school building and the dependent variable did not result in any better fit, we kept the linear modelling for the final model.

In order to examine what kind of students are likely to attend which school track and what kind of learning environment students are likely to experience at each school track in Turkey, we estimated a multinomial logistic regression with school tracks (i.e., vocational, low-performing academic, high-performing academic schools) as the dependent variable. Low-performing academic schools were used as the reference category in our estimations. For student body characteristics, SES, parents' education-related math beliefs, students' valuing of math, students' math-competence beliefs, and their behavioral engagement were predictor variables in the logistic regression. For school factors, disciplinary climate during math classes, extra-curricular activities, and quality of school facilities served as predictor variables in the logistic regression. A significant likelihood-ratio test indicated that the addition of the predictors to an intercept only model significantly improved the fit between model and data, $\chi 2(22, N = 1507) = 187.14$, $p < .001$. Moreover, results showed that the variance accounted for by school track attendance in the final model was satisfactory (McFadden's R^2 =.19). When a cut-off value of .5 was applied, the classification table indicated that our model predicted 54.6% of cases precisely, compared to 33.6% in the original data (Table 3).

Table 3.
Classification for the multi-nominal logistic regression

Observed	Predicted			
	Vocational track	Low-performing academic track	High-performing academic track	Percent correct
Vocational track	405	158	25	68.9%
Low-performing academic track	243	213	50	42.1%
High-performing academic track	120	75	218	52.8%
Overall Percentage	51.0%	27.2%	21.8%	54.6%

Overall, the logistic regression indicates that vocational and low-performing academic schools are similar in their student body characteristics. When comparing vocational with low-performing academic school tracks, only parental educational beliefs, students' math self-concept, quality of school facilities, and classroom environment were significant predictors of school track attendance (Table 4). Yet, students' SES, instrumental motivation, interest in math, math self-efficacy beliefs, behavioral engagement, and extra-curricular activities at school were non-significant predictors in the logistic regression. Specifically, given a one-unit increase in parental academic beliefs, disciplinary classroom-climate, and quality of school facilities, the odds of attending a vocational school relative to a low-performing academic school decreased by .77, .84, and .60 times, respectively. Given a one-unit increase in math self-concept scores, the odds of attending a vocational school relative to a low-performing academic school were 1.29 times more likely.

Comparing low-performing academic to high-performing academic school track attendance, students' SES, math self-efficacy, behavioral engagement, students' classroom environment and extra-curricular math activities at school were significant predictors, while parental math beliefs, students' valuing of academic tasks, students' math self-concept, and school resources were non-significant predictors. That is, given a one-unit increase in students' SES, math self-efficacy beliefs, classroom-climate, and extra-curricular school activities, the odds of attending a high-performing academic school relative to attending a low-performing academic school, were 1.90, 2.30, 1.47, and 1.73 times more likely, respectively. On the other hand, given a one-unit increase in students' math engagement and persistence in problem-solving, the odds of attending a high-performing academic school relative to attending a low-performing academic school were reduced by .78 and .74 times, respectively.

Table 4.
Results of multinomial logistic regression predicting school track attendance with low-performing academic school track as reference category (N = 1,507)

	Vocational track			High-performing academic track		
	B(SE)	*Exp(B)*	*p*	*B(SE)*	*Exp(B)*	*p*
SES	.08(.09)	1.08	.411	.64(.09)	1.90	<.001
Parent math beliefs	-.27(.08)	.77	.001	-.05(.11)	.95	.640
Student value beliefs						
Student instrumental motivation	.02(.12)	1.02	.883	-.01(.12)	1.00	.967
Student interest in math	-.08(.12)	.92	.510	-.08(.13)	.93	.572
Student ability beliefs						
Student math self-concept	.26(.10)	1.29	.014	.03(.13)	1.03	.802
Student math self-efficacy	-.15(.10)	.86	.117	.83(.12)	2.30	<.001
School behavioral engagement						
Student behavioral engagement in math classes	-.05(.09)	.95	.581	-.25(.10)	.78	.015
Student openness to problem solving	-.10(.07)	.91	.179	-.30(.09)	.74	.001
School learning environment						
Disciplinary climate during math classes	-.17(.09)	.84	.044	.38(.13)	1.47	.003
Extra-curricular math activities offered at school	.03(.16)	1.03	.846	.55(.19)	1.73	.004
Quality of school facilities	-.51(.21)	.60	.018	.17(.29)	1.18	.565
Log-likelihood	-1,325					
Likelihood-ratio $\chi2(22)$	187		<.001			
BIC	2,825					
McFadden's R^2	.19					

Note. $n_{low\text{-}performing\ academic\ school\ track}$ = 506, $n_{vocational\ school\ track}$ = 588, $n_{high\text{-}performing\ academic\ school\ track}$ = 413.

DISCUSSION

With the aim to understand implications of Turkey's tracking system and to identify policy approaches to dampen its negative implications, the current study assessed how school track attendance related to (1) student and family as well as (2) school characteristics in Turkey. For our estimations, we used the PISA 2012 data which provided us with a large representative sample of 15-year-old students from 152 different schools throughout Turkey. Our results confirmed that Turkish students coming from economically more advantaged families were more likely to attend high-performing schools. At high-performing schools, Turkish students were more likely to trust their own math ability and experience more favorable learning environments compared to students at low-performing academic schools. Contrary to our expectations, students at high-performing schools were just as much likely to value math and were less likely to be behaviorally engaged during their math classes as their peers from low-performing academic schools. Learning environments were likely to be particularly poor at vocational schools. For lower-SES students, school track attendance was associated with parental beliefs instead of SES. That is, when students reported that their parents assign a higher value to their math abilities, then students were more likely to attend low-performing academic schools instead of vocational schools. Overall, the analyses highlighted the importance of school track attendance in Turkey's selective school system. Implications of each finding are discussed in more detail below.

Social Segregation between School Tracks

In line with previous research (Alacaci & Erbas, 2010; Dincer & Uysal, 2010), our results documented a channeling of higher-SES students into high-performing academic schools in Turkey. Lower-SES students concentrated in Turkey's low-performing academic or low-performing vocational schools. While previous studies on Turkey already documented high social and academic disparities among Turkish schools, our study is the first to verify the increased likelihood of higher-SES students attending high-performing upper-secondary schools in Turkey. This is an interesting finding, since it implies that the admission process to Turkish upper-secondary schools is socially biased, despite being mostly based on test results from a national standardized exam. That is, students coming from more advantaged families are more likely to be successful during the admission process to high-performing upper-secondary schools in Turkey. This finding confirms our expectations and is in line with other studies from Turkey or France indicating the positive association between standardized admission exam results and SES (Caner & Okten, 2013; OECD, 2012; Özdemir, 2016). It is also in line with studies suggesting that SES-related educational differences are likely to increase at transition points in selective school systems such as between-school tracking (Schnabel et al., 2002).

Different reasons exist why students coming from economically more advantaged families are more likely to prevail in the tough competition to gain access to high-performing upper-secondary schools in Turkey. Higher-SES students are, for example, found to have more resources at their disposal in order to prepare themselves for the admission process (e.g., by taking extra classes in preparation for standardized exams; Caner & Okten, 2013; Özdemir, 2016). It is also possible that higher-SES students perform better on the admission exams since their overall academic achievement relates to their SES (Alacaci & Erbas, 2010). Another reason might be that higher-SES parents are more successful in navigating their child through the admission process, since higher-SES parents are found to possess more knowledge about the education system, get more actively involved during transition times and more frequently contact teachers during transition times (Groos, 2016; Kleine, 2014; Weininger & Lareau, 2003). Overall, our results support the findings that between-school tracking, even when placement is mostly based on standardized test results, is likely to increase social segregation among upper-secondary schools in Turkey, and thus, it is likely to put a threat to equity in education in Turkey.

Moreover, interestingly, while SES mattered for the access to high-performing instead of low-performing academic schools in Turkey, SES did not matter for the access to low-performing schools in our study. Comparing low-performing school track attendance with each other (i.e., vocational versus low-performing academic school track attendance) revealed that academic-oriented parents would send their child to an academic instead of a vocational low-performing school. It is possible that children of these education-oriented parents actually tried to gain access to better performing schools, but due to the high competition within the Turkish education system they were not able to, and thus, they attended low-performing academic schools. In line with studies on school choice, these results underscore the importance of parental cognitions besides SES for students' school track choice (Becker, 2010). Moreover, these results highlight variation in parents' education-related beliefs among lower-SES Turkish parents. This is important, since school track attendance is likely to hold important implications for students' future academic development, with students attending vocational schools being less likely to attend college (Caner & Okten, 2013; Özdemir, 2016). Furthermore, these results are important, since they suggest that there is variation and possibly room for modification of education-related parental beliefs of lower-SES parents in Turkey. Thus, interventions, which aim to tackle SES-related disparities in school track attendance in Turkey, may also take family characteristics such as education-related beliefs of parents into account.

Comparable Task Value, Lower Ability Beliefs, and Higher Behavioral Engagement Levels at Low-performing Academic Schools

The EVT posits that more positive motivational beliefs result in higher achievement and engagement levels of students, as well as their enrolment into more challenging academic classes and school tracks (Eccles & Wigfield, 2002). Thus, we expected that motivational beliefs and engagement levels would be highest among students attending high-performing schools. However, contrary to our expectations from EVT, the results of the logistic regression revealed that school track attendance was not associated with students' interest and instrumental motivation (i.e., value beliefs), but with their confidence in their academic abilities and behavioral engagement.

Surprisingly, students at low-achieving academic schools were less likely to trust their own math ability, but put in more effort during their math classes than students from high-performing schools. Hence, it is possible that students at low-performing academic schools may not be able to benefit academically from their enhanced work for their math classes in Turkey. At the same time, Turkish students' higher academic performance and higher confidence in their math ability at high-performing academic schools is not likely to be associated with more effort of these students. Higher performance and confidence beliefs of students at high-performing academic schools might rather be due to school factors, such as the learning environment at school.

Also surprisingly, students attending low-performing academic schools were likely to hold lower ability beliefs than students from vocational or high-performing academic schools. A possible explanation for these low ability beliefs of students at low-performing academic schools is that in Turkey's selective school system, students' competence beliefs get shaped by (1) school admittance and (2) students' personal standards. Admission to selective upper-secondary schools in Turkey is regulated by test scores from a national examination. Students attending low-achieving academic schools are likely to have been aiming to attend better schools, but due to their low exam scores, they got assigned to less successful schools. Based on this admission and their exam scores, their ability beliefs might have suffered. This would be in line with findings from Mann and colleagues (2015) who argue that in selective school systems students' ability beliefs are associated with students' school admission. Moreover, students from low-performing academic schools were more likely to hold lower ability beliefs than students at vocational schools. This might be because students at vocational schools did not strive to attend prestigious upper-secondary schools. Thus, their ability beliefs did not suffer from their admission to non-prestigious vocational schools. Other studies from Turkey also indicate that rather than students' value beliefs and behavioral engagement, particularly students' competence beliefs and school track attendance are important for their academic development (Özel, Caglak, & Erdogan, 2013; Senler & Sungur, 2009). Yildirim (2012), for example, found in a mediational model for 15-year-old Turkish students that, while students' task value beliefs predicted their learning-related behavior, the learning-related behavior in turn was not significantly associated with students' science achievement. The author found that only students' ability beliefs predicted their academic performance. These findings hold important implications for future studies aiming to explain individual differences in Turkish students' achievement as well as EVT, since it highlights the associations between school track attendance and competence beliefs in the Turkish education system. Thus, in addition to the family background, as suggested by EVT, also features of the school system, in such selective school systems as the one found in Turkey, are likely to shape students' competence beliefs and consequently also their academic achievement (Mann et al. 2015, Simpkins et al., 2015). Thus, future studies explaining individual differences in competence beliefs and academic achievement in Turkey should take features of the school system such as the performance level of schools into account.

Enhanced Learning Environments at High-performing Schools

In line with our hypothesis, results of the present study indicated that students at high-performing compared to low-performing academic schools experienced a more favorable learning environment with more positive disciplinary classroom-climates and a richer offer of extra-curricular math activities at school. These features of high-performing academic schools may result in prolonged instructional and practice time, and thus pronounced cognitive stimulation and self-efficacy beliefs for students at high-performing compared to low-performing schools in Turkey (OECD, 2012). This effect may be reinforced by an overall more academically demanding curriculum in high-performing compared to low-performing academic or vocational schools in Turkey (Giersch, 2016).

Additionally, study results emphasized that vocational schools are particularly disadvantaged in Turkey: These schools had the lowest disciplinary climate during math classes and poor resources. This is in line with previous findings that low-performing schools are often poorly equipped but are attended by students with the highest needs (Muijs, Harris, Chapman, Stoll, & Russ, 2004). Hence, researchers are calling for a relocation of resources towards these schools in order to enable them to meet the high needs of their student body (OECD, 2012; Windle, 2014).

To sum up, study results suggest that, in Turkey, admittance to high-performing academic schools and the associated experience of a more stimulating learning environment at school may supersede the positive effect of individual students' interest, instrumental motivation and behavioral engagement on their performance. This implies that in order to foster students' academic achievement, policies should focus on enhancing learning environments and students' academic confidence, especially at low-performing schools in Turkey, instead of focusing on fostering students' valuing of academic tasks or effort (OECD, 2012; Murphy, 2010). Particularly, vocational schools seem to be poorly equipped to meet the needs of its often disadvantaged students.

Moreover, our results imply that ability beliefs of students are sensitive to competition within the school system and the attainment of personal goals (Mann et al., 2015). For the Turkish school system this means that its between-school tracking has detrimental effects on students' competence beliefs with unrealized goals. In regard to EVT, it means that in addition to family background also the school system has the potential to shape students' ability beliefs, particular in countries with highly selective school systems.

LIMITATIONS and CONCLUSION

The results should be evaluated in the context of some limitations. First, the PISA data-set is cross-sectional. Therefore, no inferences about causality can be made. It is possible that the present findings reflect a self-selection bias of students with high achievement and ability beliefs into high-performing academic schools, that high-performing academic schools foster subsequent student characteristics, or that there are reciprocal effects between students' school track attendance, student body, and school characteristics. In order to disentangle these effects, longitudinal studies are needed. Another limitation of the present study is that most measures were student-reports and math-specific, which may reduce the generalizability of the results. Thus, future studies should employ additional methods of assessment such as teacher report, parent report or observational measures, and include additional academic domains.

Despite these limitations, the present study has a host of strengths. It uses a large, representative sample including a large number of schools. Moreover, it extends the literature by estimating school-track attendance in a highly selective school system such as the one found in Turkey. To our knowledge studies that take school features into account are still very scarce in Turkey. This focus on school tracks allowed us to reveal three key findings with important implications for Turkey's selective school system: First, our study is the first to confirm that lower-SES students get channeled into low-performing schools in Turkey. Thus, the Turkish between-school tracking system with its allocation system via standardized national exams to prestigious upper-secondary schools is likely to reinforce socio-economic inequalities and puts a threat to equity. Therefore, future school allocation policies should aim to combat the social segregation between upper-secondary schools in Turkey. Second, Turkish students at high-performing schools are likely to benefit from enriched learning environments at their schools and less so from enhanced behavioral engagement patterns or interest and instrumental motivation of students. Thus, in order to enhance student achievement in Turkey, educational policies should aim to improve learning environments, especially at low-performing schools in Turkey. Moreover, how school features related to gains in students' academic achievement is an important question for further longitudinal studies in Turkey. Third, ability beliefs of students are likely to be sensitive to personal goal settings and features of the school system. Hence, ability beliefs are likely to suffer for students with disappointed ambitions in selective school systems. These results suggest that besides family factors, as proposed by EVT, features of the school system, such as tracking, are also associated with students' motivational beliefs. Overall, findings of this study are likely to be applicable to other countries with between-school tracking based on prior achievement, yet, further research in other countries with such selective school systems is needed.

REFERENCES

Agresti, A. (2007). *An Introduction to categorical data analysis* (2nd ed.). New Jersey: Wiley.

Alacaci, C., & Erbas, A.K. (2010). Unpacking the inequality among Turkish schools: Findings from PISA 2006. *International Journal of Educational Development, 30,* 182-192.

Becker, B. (2010). *Bildungsaspirationen von Migranten: Determinanten und Umsetzung in Bildungsergebnisse* [Educational aspirations of immigrants: Determinants and application on educational outcomes]. MZES Workingpapers, 137.

Box, G., & Tidwell, P. (1962). Transformation of the independent variables. *Technometrics, 4,* 531-550.

Caner, A., & Okten, C. (2013). Higher education in Turkey: Subsidizing the rich or the poor? *Economics of Education Review, 35,* 75-92.

Dincer, M.A., & Uysal, G. (2010). The determinants of student achievement in Turkey. *International Journal of Educational Development, 30,* 592–598.

Eccles, J.S., & Roeser, R.W. (2011). Schools as developmental contexts during adolescence. *Journal of Research on Adolescence, 21,* 225-241.

Eccles, J.S., & Wigfield, A. (2002). Motivational beliefs, values, and goals. *Annual Review of Psychology, 53,* 109-132.

Gamoran, A. (2004). Classroom organization and instructional quality. In M. Wang & J. Walberg (Eds.) *Can unlike students learn together? Grade retention, tracking, and grouping* (pp. 141-155). Greenwich:Information Age.

Giersch, J. (2016). Academic tracking, high-stakes tests, and preparing students for college: How inequality persists within schools. *Educational Policy*, 1-29.

Graham, J.W. (2009). Missing data analysis: Making it work in the real world. *Annual Review of Psychology, 60,* 549–576.

Groos, T. (2016). *Gleich und gleich gesellt sich gern, zu den sozialen Folgen freier Grundschulwahl* [Birds of feather flock together, social implications of free primary school choice]. (Workingpaper No. 5). Gütersloh:Bertelsman Stiftung.

Guo, J., Parker, P.D., Marsh, H.W., & Morin, A.J. (2015). Achievement, motivation, and educational choices: A longitudinal study on expectancy and value using a multiplicative perspective. *Developmental Psychology, 51*, 1163-1176.

Kleine L.(2014). *Der Übergang in die Sekundarstufe I. Die Bedeutung sozialer Beziehungen für den Schulerfolg and die Formation elterlicher Bildungsentscheidungen* [Transition into secondary school. Importance of social relations and the development of parental educational decisions]. Bamberg:University of Bamberg Press.

Maaz, K., Trautwein, U., Lüdtke, O., & Baumert, J. (2008). Educational transitions and differential learning environments: How explicit between-school tracking contributes to social inequality in educational outcomes. *Child Development Perspectives, 2*, 99-106.

Mann, A., Legewie, J., & DiPrete, T.A. (2015). The role of school performance in narrowing gender gaps in the formation of STEM aspirations: A cross-national study. *Frontier in Psychology, 6*, 1-11.

Marsh, H.W., & Hau, K.T. (2003). Big-Fish--Little-Pond effect on academic self-concept: A cross-cultural (26-country) test of the negative effects of academically selective schools. *American Psychologist, 58,* 364-376.

McNeish, D., & Stapleton, L. (2016). The effect of small sample size on tow level model estimates: A review and illustration. Educational *Psychology Review, 28*, 295-314.

Muijs, D., Harris, A., Chapman, C., Stoll, L., & Russ, J. (2004). Improving schools in socioeconomically disadvantaged areas – a review of research evidence. *School Effectiveness and School Improvement, 15*, 149-175.

Murphy, J. (2010). The educator’s handbook for understanding and closing achievement gaps. Corwin:Thousand Oaks.

Muthen, B.O., & Muthen, L.K. (1998-2012). *Mplus user’s guide*. Los Angeles, CA:Muthen and Muthen.

OECD. (2009). *PISA data analysis manual. SPSS second edition.* Paris: OECD Publishing.

OECD. (2012). *Equity and quality in education: Supporting disadvantaged students and schools.* Paris: OECD Publishing.

OECD. (2013). *PISA 2012 results: What makes schools successful? Resources, policies and practices (Volume IV).* Paris: OECD Publishing.

OECD. (2014). *PISA 2012 technical report*. Paris:OECD Publishing.

Özdemir, C. (2016). Equity in the Turkish education system: A multilevel analysis of social background influences on the mathematics performance of 15-year-old students. *European Educational Research Journal, 15*, 193-217.

Özel, E., Özel, S., & Thompson, B. (2013). SES-related mathematics achievement gap in Turkey compared to European Union countries. *Education and Science, 38*, 179-193.

Özel, M., Caglak, S., & Erdogan, M. (2013). Are affective factors a good predictor of science achievement? Examining the role of affective factors based on PISA 2006. *Learning and Individual Differences, 24,* 73–82.

Roeser, R.W., Urdan, T.C. & Stephens, J.M. (2009). School as a context of motivation and development. In K.R. Wentzel & A. Wigfield (Eds.) *Handbook of Motivation at School* (pp. 381-410). New York: Routledge.

Satorra, A., & Bentler, P. M. (1999). A scaled difference chi-square test statistic for moment structure analysis (Working paper No. 412). Barcelona: Universitat Pompeu Fabra, Department of Economics.

Schnabel, K.U., Alfeld, C., Eccles, J.S., Köller, O., & Baumert, J. (2002). Parental influence on students’ educational choices in the United States and Germany: Different ramifications – Same effect? *Journal of Vocational Behavior, 60*, 178-198.

Senler, B., & Sungur, S. (2009). Parental influences on students' self-concept, task value beliefs, and achievement in science. *The Spanish Journal of Psychology, 12*, 106-117.

Simpkins, S.D., Fredericks, J., & Eccles, J.S. (2015). The role of parents in the ontogeny of achievement-related motivation and behavioral choices [Monograph]. *Monographs of the Society for Research in Child* Development, 317.

Stapleton, L., M., McNeish, D., M., & Seung Yang, J. (2016). Multilevel and single-level models for measured and latent variables when data are clustered. *Educational Psychologist, 53*, 317–330.

Tabachnick, B.G., & Fidell, L.S. (2007). *Using multivariate statistics, 5th ed*. Boston: Pearson.

Thomas, D.E., Hierman, K.L., Thompson, C., & Powers, C.J. (2008). Double jeopardy: Child and school characteristics that predict aggressive-disruptive behavior in first grade. *School Psychology Review, 37*, 516-532.

Weininger, E., & Lareau, A. (2003). Translating Bourdieu into American context: The question of social class and family-school relations. *Poetics, 31,* 375–402.

Windle, J. (2014). The rise of school choice in education funding reform: An analysis of two policy moments. *Educational Policy, 28,* 306-324.

Yavuz, M. (2009). Factors that affect mathematics-science (MS) scores in the secondary education institutional exam: An application of structural equation modeling. *Educational Sciences: Theory & Practice, 9,* 1557-1572.

Yildirim, S. (2012). Teacher support, motivation, learning strategy use, and achievement: A multilevel mediation model. *The Journal of Experimental Education, 80*, 150-172.

APPENDIX 1.

Results of multinomial logistic regression predicting school track attendance with low-performing academic school track as reference category including schools with less than 10 students (N = 1,520)

	Vocational track			High-performing academic track		
	B(SE)	*Exp(B)*	*p*	*B(SE)*	*Exp(B)*	*p*
SES	.10(.09)	1.10	.261	.67(.09)	1.97	<.001
Parent math beliefs	-.27(.08)	.77	.001	-.06(.11)	.94	.584
Student value beliefs						
Student instrumental motivation	.02(.11)	1.02	.863	-.01(.12)	1.00	.981
Student interest in math	-.10(.12)	.90	.404	-.10(.13)	.91	.455
Student ability beliefs						
Student math self-concept	.28(.10)	1.32	.008	.06(.13)	1.06	.638
Student math self-efficacy	-.16(.10)	.85	.084	.82(.12)	2.27	<.001
School behavioral engagement						
Student behavioral engagement in math classes	-.07(.09)	.93	.441	-.27(.10)	.77	.009
Student openness to problem solving	-.08(.07)	.93	.286	-.28(.09)	.76	.003
School learning environment						
Disciplinary climate during math classes	-.17(.08)	.84	.038	.39(.13)	1.47	.003
Extra-curricular math activities offered at school	.01(.16)	1.01	.962	.52(.19)	1.69	.006
Quality of school facilities	-.49(.21)	.61	.019	.18(.29)	1.20	.531
Log-likelihood	-1,337					
Likelihood-ratio $\chi 2(22)$	189		<.001			
BIC	2,894					
McFadden's R^2	.19					

Note. $n_{\text{low-performing academic school track}}$ = 519, $n_{\text{vocational school track}}$ = 588, $n_{\text{high-performing academic school track}}$ = 413.

TÜRKÇE GENİŞLETİLMİŞ ÖZET

Türkiye'deki okulların performans açısından değerlendirme çalışmaları yapılmış olsa da öğrenmeyi arttırmak için sağladıkları imkanlar ve bünyelerindeki öğrenci profilleri açısından farklılıklar yeterince incelenmemiştir. Bu makalede, okul merkezli bir yaklaşım izlenerek ve PISA 2012 verileri incelenerek, Türkiye'deki nispeten düşük performanslı mesleki ve akademik okullar ile yüksek performanslı okullar, öğrencilerin aile geçmişleri, motivasyonları ve okulun öğrencilere sağladığı imkanlar göz önünde bulundurularak incelenmiştir.

Araştırmalar ailelerin sosyo-ekonomik durumlarının öğrencilerin başarı ve motivasyon seviyelerini etkilediğini ortaya koymuştur. Düşük gelirli aileler mesleki okulları tercih ederken, yüksek gelirli olanlar çocuklarının akademik başarısı yüksek olan liselere devam etmeleri için çaba sarf etmeyi tercih etmektedirler. Aynı zamanda yüksek sosyo-ekonomik durumdaki aileler ülkedeki eğitim ve sınav sistemi ile ilgili daha fazla bilgi sahibi olmakta ve okul ile daha sıkı iletişim kurmaktadırlar. Bu ailelerin çocukları da düşük gelirlilere kıyasla daha başarılı ve daha motive olmaktadır. Beklenti-değer teorisine göre çocukların eğitime bakış açıları yetiştikleri ortamdan etkilenmektedir. Okul başarısını özendiren ve buna değer veren ailelerin çocukları, yüksek kalitede eğitim almak için daha motive olmaktadırlar.

Aile etkisi dışında, okulların sağladıkları ve sundukları da öğrencilerin başarı ve motivasyon seviyelerini ve kararlarını etkileyen bir diğer faktördür. Okulların fiziki ortamları ve örneğin matematik eğitimine olan bakış açıları öğrencilere sunulan imkanları ve dolayısıyla başarılarını etkilemektedir. PISA sonuçlarına göre başarılı olan okullarda daha yoğun bir akademik eğitim görülmekte ve fiziksel şartlar genel olarak daha iyi olmaktadır. Öğrencilere ayrılan okul-sonrası aktiviteler daha çeşitlidir. Verilen ödev ve görevler daha zorlayıcı olabilmektedir.

Bu araştırma okullar arasında öğrenci performansına etki edecek farklılıkları incelemek amacıyla öğrenci (matematik başarısı, ailenin matematik algısı, sosyo-ekonomik durum, vb.) ve okul (matematik dersi esnasındaki atmosfer, imkanların kalitesi ve çeşitliliği, vb.) alt başlıkları altındaki değişkenleri incelemektedir. Hangi profildeki okullarda hangi profildeki öğrencinin olduğu araştırmanın odak noktasıdır. Bu seçimlerdeki farklılıklar ve sebepleri daha iyi anlaşıldıkça, genel başarıyı ve motivasyonu yükseltecek gerek bireysel gerekse toplumsal adımlar atılmasında önemli bir yer tutar.

Araştırma OECD tarafından yapılan PISA 2012 Türkiye örneklemi verisini kullanmaktadır. 10 veya daha az öğrenciden veri elde edilmiş okullar analizlere dahil edilmemiştir. Analizlerin gerçekleştirildiği örneklem ortalama 15.8 yaşında ve %49'u kız olmak üzere 4742 öğrenciden oluşmaktadır.

Verdikleri müfredata ve PISA matematik başarılarına göre okullar kategorize edilmiş ve düşük başarılı mesleki, düşük başarılı akademik ve yüksek başarılı akademik olarak üçe ayrılmıştır. Matematik skoruna göre yüksek başarılı mesleki okullar bulunmadığı için o kategori değerlendirmeye alınmamıştır. Başarı seviyeleri PISA değerlendirmesine göre 0'dan 6'ya kadardır. Seviye 3'ten itibaren problem çözme becerisi dahil olur. Bu çalışmada seviye 2 ve altı düşük, seviye 2 üstü yüksek başarılı olarak adlandırılmıştır.

Türkiye ve OECD ülkeleri ortalamaları karşılaştırıldığında, Türkiye'deki öğrenciler OECD ortalamasına kıyasla matematik testinde 51 puan geride kalmışlardır. Türkiye'deki ailelerin sosyo-ekonomik durumları da ortalamanın yaklaşık 1.5 standart sapma altındadır. Türk öğrenciler OECD ortalamasına kıyasla matematik alanına daha fazla ilgi duyarken, dersteki disiplin atmosferini değerlendirdiklerinde OECD ortalamasına yakın cevaplar vermişlerdir. Okul müdürlerince değerlendirilen okulların fiziksel koşulları ise OECD ortalamasının altında kalmıştır.

Analiz sonuçları mesleki ve düşük başarılı akademik okulların öğrenci profili bakımından benzer olduğunu göstermektedir. Düşük başarılı akademik okullar mesleki okullar ile karşılaştırıldığında ise, düşük başarılı akademik okullarda matematik derslerindeki disiplin ortamı, okul fiziksel koşulları ve ailelerin eğitimle ilgili görüş değişkenlerindeki skorlar daha yüksektir. Mesleki okullarda ise öğrencilerin matematik benlik algılarının daha yüksek olduğu görülmektedir.

Yüksek başarılı akademik okullar ve düşük başarılı akademik okullar karşılaştırıldığında ise, öğrenci matematik öz-yeterliliği, matematik derslerindeki disiplin ortamı ve okulda ders sonrası matematik aktiviteleri yapılması değişkenlerindeki skorlar daha yüksektir. Fakat öğrencilerin matematiğe ilgisi ve motivasyonu ve okul fiziksel koşulları değişkenleri açısından fark bulunmamıştır. Analizler sosyo-ekonomik olarak daha avantajlı ailelerin çocuklarının yüksek başarılı okullara gitme olasılığının daha fazla olduğunu da doğrulamıştır. Aynı zamanda, bu okullardaki öğrenciler matematik alanında kendilerine daha fazla güvenmektedir ve öğrenme ortamları olumlu yönde farklılıklar göstermektedir. Fakat beklediğimizin aksine öğrencilerin matematik dersine kendini verme değişkeninde okullar arasında anlamlı bir farklılık görülmemiştir. Başarılı okullardaki öğrenciler matematik konusunda kendilerine daha fazla güveniyorken, düşük başarılı okullarda olanlar matematik derslerine kendilerini daha çok vermektedirler.

Bu farklılıklar sistemin yarattığı sıralamanın bir sonucu olabilir. Matematik başarısının hali hazırda düşük olmasının yanı sıra, zaten başarılı bir okulda değilim algısı öğrencinin motivasyonunu ve kendine güvenini etkileyebilir. Örneğin, matematik konusunda en düşük kendine güven meslek okulu öğrencilerindedir. Meslek okulları, matematik dersi işlenişi ve dersteki disiplin seviyesinde de son sırayı almaktadır. Eğitimlerine devam etmeyecekleri düşüncesi de bu durumda etkili olabilir. Düşük sosyo-ekonomik seviyedeki öğrencilerin matematik seviyeleri ailelerinin matematiğe bakışlarından çok etkilenebilmektedir. Matematiğe ve eğitime değer veren ailelerin çocukları meslek okulları yerine genel liseleri tercih edebilmektedir. Sadece okul ortamı değil, ailelerin eğitimle ilgili inanışları da son derece etkileyici olabilmektedir.

Bu bilgilerin ışığında, öğrenciler ve okullar arasındaki bu farklılıkları giderecek, okulları ve öğrenme ortamlarını iyileştirici, imkanları daha ulaşılabilir hale getiren, öğrencilerin ruh sağlığına zarar vermeyen politikalar geliştirmek bizleri eğitim alanında eşitliğe bir adım daha yaklaştıracaktır.

TURJE
www.turje.org

Volume: 8 Issue: 3

Turkish Journal of Education

DOI: 10.19128/turje.545798
Research Article

Received 28.03.2019
Accepted 17.07.2019

Evaluating Turkish science curriculum with PISA scientific literacy framework

Nurcan Cansız
Artvin Çoruh University, Faculty of Education, Artvin, Turkey, nurcansiz7911@gmail.com
Mustafa Cansız
Artvin Çoruh University, Faculty of Education, Artvin, Turkey, mustafacansiz@gmail.com

ABSTRACT Any society needs more scientifically literate citizens even if they do not follow a career in science. In the 2015 PISA assessment, Turkey ranked 34th among 35 OECD countries based on science literacy scores. The relatively unsuccessful results of Turkey from international level examinations like PISA has necessitated the questioning of various components of science education. One of these components is surely the science curriculum. Being aware of this, we investigated the primary and middle school Turkish science curriculum for the balance of science literacy aspects based on the PISA 2015 science literacy framework. This framework defines scientific literacy under four aspects, namely contexts, knowledge, competencies, and attitudes. The results revealed that the Turkish science curriculum does not adequately reflect all dimensions of science literacy and is dominated by the pure knowledge of the content of science. The curriculum developers should consider these two points in future curriculum revisions to increase our success in international examinations like PISA and to help raise scientifically literate students.

Keywords: *Science literacy, Science curriculum, PISA 2015, Science education.*

Türkiye'de uygulanan fen bilimleri dersi öğretim programının PISA fen okuryazarlığı çerçevesiyle değerlendirilmesi

ÖZ Her toplumun -fen bilimleri alanında kariyer yapmayacak olsa bile- fen okuryazarı bireylere ihtiyacı vardır. Türkiye 2015 PISA uygulamasında fen okuryazarlığı puanına göre 35 OECD üyesi ülkeler arasında 34. sırada yer almıştır. PISA gibi uluslararası düzeyde uygulanan sınavlarda alınan görece başarısız sonuçlar, Türkiye'de fen eğitiminin farklı bileşenlerinin sorgulanmasını gerekli kılmıştır. Sınıf içi uygulamalara dönük bu bileşenlerden biri de kuşkusuz fen bilimleri dersi öğretim programıdır. Bu noktadan hareketle, bu çalışmada Türkiye'de uygulanan fen bilimleri dersi öğretim programının fen okuryazarlığı boyutlarını hangi ölçüde yansıttığı PISA 2015 Fen Okuryazarlığı Değerlendirme Çerçevesi kullanılarak araştırılmıştır. Bu çerçeve fen okuryazarlığını bağlamlar, bilgi, yeterlikler ve tutumlar olmak üzere dört boyutuyla tanımlamaktadır. Bulgular mevcut programın fen okuryazarlığın dört boyutunu dengeli bir şekilde vurgulamada yetersiz kaldığını ve programının daha çok içerik bilgisine yoğunlaştığını ortaya koymuştur. Gelecekte yapılacak program güncelleme ve geliştirme çalışmalarında bu iki noktanın göz önünde bulundurulması, hem PISA gibi uluslararası sınavlarda başarımızı artıracak, hem de fen okuryazarı öğrenciler yetiştirmede mesafeleri daha hızlı kat etmemize olanak sağlayacaktır.

Anahtar Kelimeler: *Fen okuryazarlığı, Fen bilimleri öğretim programı, PISA 2015, Fen eğitimi.*

Citation: Cansız, N., & Cansız, M., (2019). Evaluating Turkish science curriculum with PISA scientific literacy framework. *Turkish Journal of Education, 8*(3), 217-236. DOI: 10.19128/turje.545798

INTRODUCTION

The result of Programme for International Student Assessment (PISA) provided evidence that Turkish 15-year-old students are not scientifically literate enough to meet the needs of the knowledge society in which individuals are prepared as self-sufficient participants to create scientific and technological knowledge. Based on PISA 2015 statistics, most of the 15-year-old students (96.7%) attend high schools in Turkey (OECD, 2016). That is, they successfully completed the middle school and had covered the elementary and middle school curricula already. The major goal of Turkish science curriculum is to educate scientifically literate students (Ministry of National Education [MoNE], 2004, 2013, 2017). Therefore, we expect those students to perform with a high degree of competency in PISA which evaluates the science literacy of 15-year-olds along with mathematical and reading literacy. In the 2015 PISA assessment, Turkey ranked 34th among 35 Organization for Economic Cooperation and Development (OECD) countries based on science literacy scores. Moreover, PISA 2015 result indicated that the percent of top performing students who can show advance scientific thinking and reasoning skills is less than 1% of all 15-year-old in Turkey (OECD, 2016; Yorulmaz, Çolak, & Ekinci, 2017). PISA describes six levels of proficiency for science literacy and identifies students as top performers at or above level 5 (OECD, 2017). The levels are arranged hierarchically in such a way that the use of content knowledge, cognitive operations, and complexity in scientific reasoning increase toward the upper levels.

It is possible to offer a number of reasons for the low success of Turkey in PISA such as the deficiency of parental support (Şad, 2012), the shortage of educational resources (OECD, 2013), low job satisfaction of teachers (Blandford, 2000), the disparity between schools (OECD, 2016), and exam-oriented education (Acat, Anılan, & Anagun, 2010). Another main reason for the low success of Turkey may be the science curriculum that does not fully meet what is required to be scientifically literate. Since a curriculum is among the major sources available to teachers (Kesidou & Roseman, 2002), they make use of curriculum in many different ways such as to find out objectives, contents, activities and the limitations for a specific topic. Moreover, it is a guide for a teacher to decide on how to teach, what to teach, when to teach, where to teach, and even why to teach. Therefore, the science curriculum should be analyzed to address inadequacies for better science education (Cansiz & Turker, 2011; Kesidou & Roseman, 2002). This research is significant and necessary to explore the reason behind the low scientific literacy performances of Turkish 15-year-old students although they completed a science curriculum from grade 3 to 8 which states the scientific literacy as its major goal. Moreover, it may reveal important results from which curriculum developers and educators draw conclusions which may trigger a fundamental change for achieving scientific literacy. Considering this critical issue, we investigated whether Turkish science curriculum, released in 2017, has the potential to prepare scientifically literate students based on PISA science literacy framework.

Science Literacy

For more than five decades, many countries, especially developed ones, have attached particular importance to science literacy. Hurd (1958), for example, questioned the education system of United States by referring to the term science literacy for the first time after Soviet Union had launched the Sputnik I -the world's first artificial satellite- into Earth’s orbit in 1957. From then on, the stakeholders of science education have focused on the important question of why people should be scientifically literate. Some researchers provided several reasonable arguments to this question (e.g., Durant, Evans, & Thomas, 1989; Millar, 1996). These researchers defended that science literacy may have a profound impact on the wealth of a nation, becoming informed users of scientific knowledge in everyday life, and the use of science in public decision-making. These arguments have shaped the definition of science literacy and the characteristics of scientifically literate individuals. Although there is not a unique definition of science literacy, previous studies underlined that public should have a general sense of science in order to be informed and critical users of science (e.g., Kolstø, 2001; Miller, 1995, 1998). In line with this perspective, Durant (1993) underlined that science literacy refers to scientific knowledge that the societies should know to maintain their daily life. The general public does not have direct access to scientific research but in a scientifically and technologically complex culture, they ought to know something about science (Durant, 1993). National Research Council (NRC, 1996) defined a scientifically literate person as the one who can read and understand popular science articles, evaluate the trustworthiness of such articles, express opinion about socio-scientific issues, and differentiate facts from fictions.

Regarding the concerns about the level of public science literacy, the stakeholders of education have consistently focused on the science education in schools (e.g., Carlton, 1963; Collins, 1998; Fensham, 2008; Gallagher, 1971; Hurd, 1958; Rudolph & Horibe, 2015; Yager, 1986). They have questioned if the science curriculum includes knowledge and skills to prepare students for the special needs of the times. It is specifically suggested that schools should focus more on teaching the essentials of science literacy rather than covering further content (Rutherford & Ahlgren, 1990). Considering these warnings and recommendations, a number of reform movements were initiated to improve science education that prepares students to a “real” world. In order to improve science education, countries changed the way science is taught (Van Driel, Beijaard, & Verloop, 2001). Instead of traditional science teaching -which includes introducing concepts, facts, theories and memorizing them- an inquiry-based science teaching was embraced including hands-on activities (Van Driel et al., 2001). It was argued that this change in science education helps students discover science topics, develop higher-order thinking skills, and be prepared for science and technological issues of the 21st century (Van Driel et al., 2001).

In 1985, the American Association for the Advancement of Science (AAAS) provided a definition for science literacy and recommendations for what students should know in science when they graduate from high school as scientifically literate individuals. Based on AAAS's (1990) definition, science literacy does not just mean knowing scientific concepts and principles; rather it has many aspects such as being aware of the environment; understanding the complex relationship between science, technology, and mathematics; having a basic understanding of science; and appreciating the subjective elements of scientific knowledge.

Almost all recent definitions of science literacy are also based on similar aspects. OECD (2016), for example, highlighted that scientifically literate society should be intellectual in that they need to approach issues scientifically. OECD intends to measure 15-year-old students’ science literacy by preparing diverse assessment tools based on this definition. OECD’s assessment of science literacy has awakened countries to rethink whether the science education in their schools has the standards to raise students as scientifically literate citizens. Based on such international assessments, countries are rethinking their education systems so that their students become more prepared to succeed in the 21st century.

A Quick Look at the Science Curriculum in Turkey

Since 2012-2103 school year, the compulsory education in Turkey has been 12 years with three stages. The first stage is the primary school which includes grades 1 to 4. It covers children of 66 months to 10 years old (Eurydice, 2018). The middle school is the second stage including grades 5 to 8 and covers children of 10 to 14 years old. The high school, grades 9 to 12, is the last stage before higher education. Children aged 14 to 18 years old attend high school. The type of schools in middle and high school may change. However, all students take the same science education until the end of middle school. Science education starts in the 3rd grade of primary school and continues until the 8th grade. In each grade level, all students should follow the same national curriculum. Until the recent one, the organization of the topics was based on a spiral curriculum, i.e. the topics are covered in each grade with increasing complexity by reinforcing previous learning. With the 2017 curriculum, the order of the topics has been changed in such a way that topics progress from universe to human body (see 2017 science curriculum).

The science curriculum has been updated several times since the 2000s with the most recent revisions in 2017. In 2000, a major paradigm shift arose in many disciplines including science. Following the international reforms in science education, the national educational paradigm has shown a shift from the traditional teacher-centered approach to a contemporary student-centered approach as well. In line with this philosophy, the process of active construction of knowledge based on personal experience was underlined rather than passively acquiring it. With the changing philosophy of curriculum, education systems aim to prepare students with skills such as discovering knowledge, testing hypothesis and evaluating results, arguing and making evidence-based decisions (MoNE, 2004, 2013, 2017). In 2013, the science curriculum was revised again to include socio-scientific issues which include controversial issues related to science, technology, and society. The continuous reforms in science education resulted in new trends such as science, technology, engineering, and mathematics (STEM). The current science curriculum (i.e., 2017) focuses on STEM education for the first time. Values education is implicitly included in the curriculum, and the role of the teacher is highlighted in values education (MoNE, 2017).

METHODOLOGY

Document analysis, a type of qualitative research methods, was adopted in this study. Bowen (2009) underlined that in document analysis researchers interpret various resources (e.g., books, curriculum materials, and lesson plans) to give meaning to them. Bowen added that one of the frequent use of document analysis is to use rubric to score documents, which is what the researchers utilized within the scope of this study.

The aim of this study is to analyze the Turkish science curriculum from grades 3 to 8 to determine the emphasis given to the following aspects of science literacy: (1) contexts, (2) knowledge, (3) competencies, and (4) attitudes. The curriculum has been released in 2017 and is now being implemented in all schools in Turkey. The subject areas and the corresponding number of objectives in each grade were given in Table 1.

Table 1.
The content knowledge categories and number of objectives in 2017 Turkish science curriculum

Subject Area	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
Earth and the Universe	5	5	9	5	10	3
Life and Living Things	11	8	8	20	17	28
Physical phenomena	16	20	14	19	27	16
Matter and Its Nature	4	10	6	13	16	17
Science and Engineering Practices	-	3	3	4	4	4
Total	*36*	*46*	*40*	*61*	*74*	*68*

Instrument: Science Literacy Framework of PISA 2015

For the purpose of this study, the objectives of science curriculum from grades 3 to 8 were analyzed and categorized using the PISA 2015 science literacy framework. PISA has introduced the four aspects of science literacy as the contexts, the knowledge, the competencies, and the attitudes (OECD, 2016). The framework in Table 2 presents the aspects of science literacy and the explanations for each aspect based on PISA 2015 Assessment and Analytical Framework.

Table 2.
PISA 2015 science literacy assessment framework

Science Literacy Aspect	Description
Contexts	Personal, local/national and global issues, both current and historical, which demand some understanding of science and technology
Knowledge	An understanding of the major facts, concepts and explanatory theories that form the basis of scientific knowledge. Such knowledge includes knowledge of both the natural world and technological artefacts (content knowledge), knowledge of how such ideas are produced (procedural knowledge), and an understanding of the underlying rationale for these procedures and the justification for their use (epistemic knowledge)
Competencies	The ability to explain phenomena scientifically, evaluate and design scientific enquiry, and interpret data and evidence scientifically.
Attitudes	A set of attitudes towards science indicated by an interest in science and technology, valuing scientific approaches to enquiry where appropriate, and a perception and awareness of environmental issues.

The four dimensions presented in Table 2 have also subdimensions. Moreover, each subdimension includes several practices that a scientifically literate person is capable of doing. For example, the competencies aspect has three subdimensions as *explain phenomena scientifically, evaluate and design scientific enquiry, and interpret data and evidence scientifically*. The subdimension *explain phenomena scientifically* has also several practices and one of them is *recall and apply appropriate scientific knowledge*. That is, a scientifically literate person should have scientific competencies and one of which is to be able to explain phenomena scientifically. To achieve this, he/she should recall and apply appropriate scientific knowledge. Each dimension including subdimensions and practices are provided in Appendix 1.

The Analysis Procedure

We analyzed the objectives of each unit from Grade 3 to Grade 8 using the framework in Table 2. Two independent researchers examined each science curriculum to determine the distribution of the four aspects of science literacy. Before the analysis, they examined the science literacy framework for clarifications and scoring process. They analyzed a number of objectives together to use the same analysis criterion and scoring procedure. They performed analysis by assigning each objective to one of the four aspects of the framework. After researchers completed the analysis of each curriculum independently, they compared their findings to reconcile their decisions. At the beginning, there were 41 inconsistencies out of 432 decisions. Therefore, the interrater reliability was calculated as 90.51%. Then Cohen's Kappa (Cohen, 1960) was administered to test the significance of this interrater reliability. The interrater reliability was statistically significant (Cohen's Kappa = .78, $p < .001$). Based on Landis and Koch's (1977) criteria, the level of agreement was substantial. In the end, full reconciliation occurred in the analysis of the objectives. It is worth mentioning that one objective can be assigned to more than one aspect of science literacy. Therefore, the total frequencies given in the results section should not necessarily be equal to the total number of objectives for each grade in the curriculum.

A step by step procedure was presented below to exemplify the analysis of objectives. An objective was focused, e.g. "students discover the materials that are attracted to magnets by doing experiments" (Objective number: F.4.3.2.2).

Each researcher read it carefully and decided whether it emphasizes content, competency, or attitude aspect. Two researchers decided that it emphasizes competency aspect.

It was then decided to which subdimension it belongs to under competency aspect. This objective requires that students interpret data and evidence scientifically (a subdimension of competency aspect) since they will experiment with different materials and magnets. As a result, they will have data on which materials are attracted or not attracted to magnets.

After reaching consensus on subdimension, the researchers focused on which practice the objective belongs to under the subdimension. This objective was assigned to the practice "analyze and interpret data and draw appropriate conclusions" because students experiment with different kind of materials and magnets, collect data on which of them are attracted to the magnets and draw conclusions about the materials that are attracted to the magnets.

This objective was also assigned to the content aspect of scientific literacy. It was in physical science subdimension.

Finally, it was decided that it is written free from a context.

A similar pattern was followed during the analysis of all objectives. The Figure 1 summarizes the data analysis process of objectives.

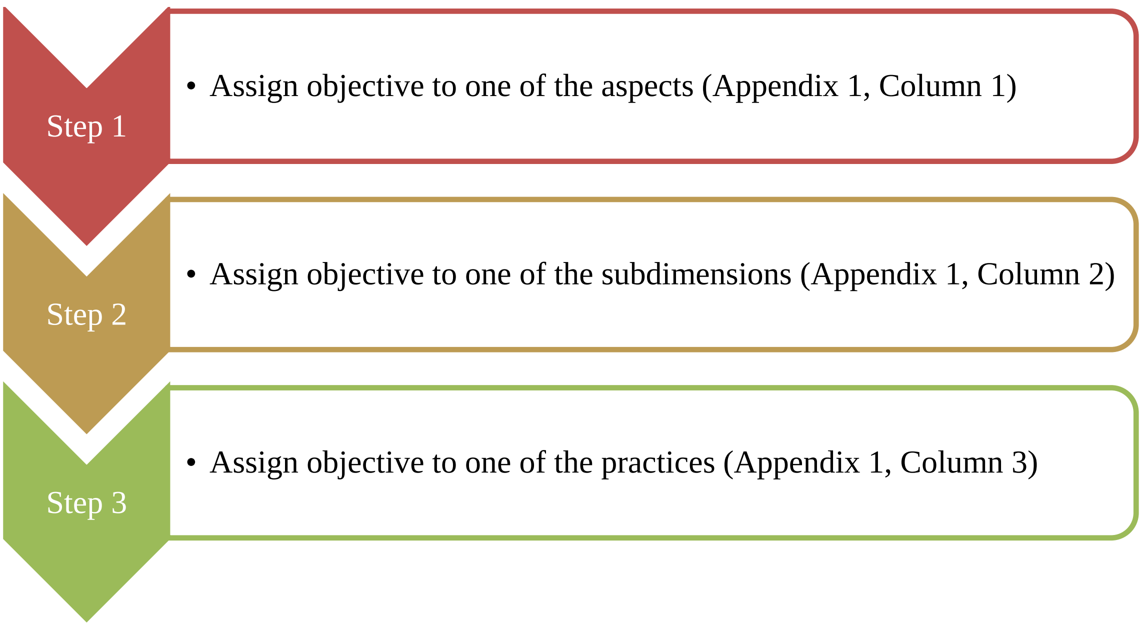


Figure 1. The data analysis process

RESULTS

The present study was designed to determine the extent to which Turkish science curriculum emphasizes the PISA 2015 science literacy aspects. With this in mind, we classified the objectives in Turkish science curriculum for grades 3 to 8. Table 3 shows the results of data analysis which reveals the frequencies and percentages in each aspect of science literacy.

Table 3.
The distribution of objectives in each aspect of science literacy from grade 3 to 8

	Grade 3		Grade 4		Grade 5		Grade 6		Grade 7		Grade 8	
	f	%	f	%	f	%	f	%	f	%	f	%
Contexts												
Personal	7	19.4	6	13.0	-	-	3	4.9	1	1.4	1	1.5
Local/National	1	2.8	4	8.7	2	5.0	5	8.2	4	5.4	8	11.8
Global	-	-	7	15.2	3	7.5	2	3.3	-	-	3	4.4
Total	*8*	*22.2*	*17*	*36.9*	*5*	*12.5*	*10*	*16.4*	*5*	*6.8*	*12*	*17.7*
Knowledge												
Knowledge of the Content of Science	22	61.1	15	32.6	10	25.0	27	44.3	36	48.6	31	45.6
Procedural Knowledge	10	27.7	6	13.0	13	32.5	8	13.1	6	8.1	9	13.2
Epistemic knowledge	1	2.8	-	-	-	-	-	-	2	2.7	-	-
Total	*33*	*91.6*	*21*	*45.6*	*23*	*57.5*	*35*	*57.4*	*44*	*59.4*	*40*	*58.8*
Competencies												
Explain phenomena scientifically	10	27.7	16	34.8	10	25.0	18	29.5	19	25.7	17	25.0
Evaluate and design scientific inquiry	1	2.8	6	13.0	3	7.5	-	-	1	1.4	4	5.9
Interpret data and evidence scientifically	3	8.3	10	21.7	5	12.5	4	6.6	6	8.1	3	4.4
Total	*14*	*38.8*	*32*	*69.5*	*18*	*45.0*	*22*	*36.1*	*26*	*35.2*	*24*	*35.3*
Attitudes												
Interest in science	-	-	-	-	-	-	-	-	-	-	-	-
Valuing scientific approaches to enquiry	2	5.6	1	2.2	1	2.5	4	6.6	-	-	5	7.4
Environmental awareness	4	11.1	8	17.4	5	12.5	2	3.3	6	8.1	5	7.4
Total	*6*	*16.7*	*9*	*20.2*	*6*	*15.0*	*6*	*9.9*	*6*	*8.1*	*10*	*14.8*

Table 3 shows that objectives in all grades rarely include personal, local, or global issues. In grade three, only eight objectives (out of 36) are based on a context. These numbers vary in different grades: in fourth grade, 17 objectives (out of 46); in fifth grade, five objectives (out of 40); in sixth grade, 10 objectives (out of 61); in seventh grade, five objectives (out of 74); and in eighth grade, 12 objectives (out of 68) include personal, local, or global context. These contexts are important since how students use knowledge and competencies in these specific contexts is the main idea underlying science literacy.

The analysis of the objectives in terms of the knowledge aspect of science literacy revealed that the content knowledge is represented more than other two knowledge categories (see Table 3). The least emphasis is on epistemic knowledge. Only in grade five, the emphasis on procedural knowledge is higher than content and epistemic knowledge.

The percentage distribution based on scientific competencies showed that the most emphasis is on explain phenomena scientifically (see Table 3). The other two competencies -evaluate and design scientific inquiry- are emphasized least in the curriculum for all grades except grade eight.

Attitudes aspect of science literacy is given the least emphasis among other aspects. Moreover, objectives mostly focus on environmental awareness among other dimensions of attitude. Table 3 gives an overall distribution of objectives based on four aspects of science literacy. We provided in-depth results including typical objectives for each aspect in the following sections.

Contexts in the Science Curriculum

Regarding the context aspect of science literacy, the percentage distribution was presented in Table 4. The minus (-) sign in the cells of tables refers to the fact that the curriculum does not include any objectives corresponding to that cell. The result is notable in that there are not adequate context-based objectives in Turkish science curriculum. That is, there are few objectives including personal, local and global contexts. Among them, health issues at personal level are represented more than others. At global level, on the other hand, environmental quality is emphasized more than others. Natural resources are emphasized at local level most. However, the frequencies for different contexts at each level are very low as compared to the total number of objectives. Typical objectives for contexts dimension are provided in Appendix 2.

Table 4.
The frequencies of the context of objectives from grade 3 to 8

	Grade 3		Grade 4		Grade 5		Grade 6		Grade 7		Grade 8	
	f	%	f	%	f	%	f	%	f	%	f	%
Personal												
Health and Disease	4	11.1	4	8.7	-	-	3	4.9	-	-	1	1.5
Natural Resources	-	-	-	-	-	-	-	-	-	-	-	-
Environmental Quality	1	2.8	2	4.3	-	-	-	-	-	-	-	-
Hazards	2	5.6	-	-	-	-	-	-	-	-	-	-
Frontiers of Science and Technology	-	-	-	-	-	-	-	-	-	-	-	-
Total	*7*	*19.5*	*6*	*13.0*	-	-	*3*	*4.9*	-	-	*1*	*1.5*
Local												
Health and Disease	-	-	1	2.2	-	-	2	3.3	-	-	1	1.5
Natural Resources	-	-	2	4.3	-	-	1	1.6	1	1.4	3	4.4
Environmental Quality	1	2.8			1	2.5	1	1.6	1	1.4	2	2.9
Hazards	-	-	-	-	-	-	-	-	-	-	-	-
Frontiers of Science and Technology	-	-	1	2.2	-	-	1	1.6	2	2.7	2	2.9
Total	*1*	*2.8*	*4*	*8.7*	*1*	*2.5*	*5*	*8.1*	*4*	*5.5*	*8*	*11.7*
Global												
Health and Disease	-	-	-	-	-	-	-	-	-	-	-	-
Natural Resources	-	-	-	-	-	-	1	1.6	-	-	-	-
Environmental Quality	-	-	7	15.2	3	7.5	1	1.6	-	-	-	-
Hazards	-	-	-	-	-	-	-	-	-	-	3	4.4
Frontiers of Science and Technology	-	-	-	-	-	-	-	-	-	-	-	-
Total	-	-	*7*	*15.2*	*3*	*7.5*	*2*	*3.2*	-	-	*3*	*4.4*

Knowledge Aspect of Objectives

Table 5 displays the distribution of the knowledge aspect of science literacy. As mentioned before, the knowledge aspect has three subdimensions as content, procedural, and epistemic knowledge. The objectives in the content dimension only cover the facts, principles or theories rather than apply them in contexts. For example, one of the objectives in grade five is as follows: students "explain the main differences between heat and temperature" (MoNE, 2017, p. 28). The emphasis in this objective is only on the facts of science. There is no reference to procedural or epistemic knowledge of science. In all grades, the subdimension procedural knowledge is observed less in the objectives as compared to the content knowledge. There are more objectives including measurement issues compared to others under procedural knowledge. The other more frequent procedural knowledge is the use of control variables and identifying possible causal mechanisms. However, the objectives very rarely include the epistemic dimension of scientific knowledge. Only two objectives emphasize the reasoning based on data (see Appendix 3 for typical objectives for knowledge dimension).

Table 5.
The frequencies of the knowledge aspects of objectives from grade 3 to 8

	Grade 3		Grade 4		Grade 5		Grade 6		Grade 7		Grade 8	
	f	%	f	%	f	%	f	%	f	%	f	%
Knowledge of the Content of Science												
Physical systems	12	33.3	8	17.4	4	10.0	9	14.8	17	23.0	17	25.0
Living System	6	16.7	2	4.3	1	2.5	15	24.6	12	16.2	11	16.2
Earth and Space System	4	11.1	5	10.9	5	12.5	3	4.9	7	9.5	3	4.4
Total	*22*	*61.1*	*15*	*32.6*	*10*	*25.0*	*27*	*44.3*	*36*	*48.7*	*31*	*45.6*
Procedural Knowledge												
The concept of variables	-	-	-	-	2	5.0	2	3.3	1	1.4	1	1.5
Concepts of Measurements	7	19.4	5	10.9	8	20.0	3	4.9	5	6.8	4	5.9
Ways of assessing and minimizing uncertainty	-	-	-	-	-	-	-	-	-	-	-	-
Mechanisms to ensure the replicability of data	-	-	-	-	-	-	-	-	-	-	-	-
Common ways of abstracting and representing data	-	-	-	-	-	-	1	1.6	-	-	2	2.9
The control-of-variables	3	8.3	1	2.2	3	7.5	2	3.3	-	-	2	2.9
The nature of an appropriate design for a scientific question	-	-	-	-	-	-	-	-	-	-	-	-
Total	*10*	*27.7*	*6*	*13.1*	*13*	*32.5*	*8*	*13.1*	*6*	*8.2*	*9*	*13.2*
Epistemic knowledge												
The nature of scientific observations, facts, hypotheses, models, and theories	-	-	-	-	-	-	-	-	-	-	-	-
The purpose and goals of science as distinguished from technology	-	-	-	-	-	-	-	-	-	-	-	-
The values of science	-	-	-	-	-	-	-	-	-	-	-	-
The nature of reasoning	-	-	-	-	-	-	-	-	-	-	-	-
How scientific claims are supported by data	1	2.8	-	-	-	-	-	-	1	1.4	-	-
The function of different forms of empirical enquiry	-	-	-	-	-	-	-	-	-	-	-	-
How measurement error affects the degree of confidence	-	-	-	-	-	-	-	-	1	1.4	-	-
The use and role of abstract models and their limits	-	-	-	-	-	-	-	-	-	-	-	-
The role of collaboration and critique	-	-	-	-	-	-	-	-	-	-	-	-
The role of scientific knowledge, along with other forms of knowledge	-	-	-	-	-	-	-	-	-	-	-	-
Total	*1*	*2.8*	-	-	-	-	-	-	*2*	*2.8*	-	-

Scientific Competencies

PISA 2015 framework introduces scientific competencies as an aspect of science literacy including three subdimensions (OECD, 2017). In the objectives, the most emphasized competency is *explaining phenomena scientifically*. The other competencies are seldom emphasized in the objectives as seen in Table 6. In terms of subdimension interpret data and evidence scientifically, the emphasis in the objectives is only on the ability to analyze and interpret data and draw appropriate conclusions. Considering the subdimension evaluate and design scientific inquiry, the emphasis in the objectives is only on the ability to propose a way of exploring a given question scientifically. Typical objectives for scientific competencies dimension are provided in Appendix 4.

Table 6.
The frequencies of the objectives in competencies aspect from grade 3 to 8

	Grade 3		Grade 4		Grade 5		Grade 6		Grade 7		Grade 8	
	f	%	f	%	f	%	f	%	f	%	f	%
Explain phenomena scientifically												
Recall and apply appropriate scientific knowledge.	4	11.1	4	8.7	1	2.5	6	9.8	7	9.5	6	8.8
Identify, use and generate explanatory models and representations	2	5.6	-	-	7	17.5	3	4.9	5	6.8	2	2.9
Make and justify appropriate predictions	1	2.8	1	2.2	2	5.0	6	9.8	3	4.1	4	5.9
Offer explanatory hypotheses	2	5.6	-	-	-	-	1	1.6	-	-	-	-
Explain the potential implications of scientific knowledge for society	1	2.8	11	23.9	-	-	2	3.3	4	5.4	5	7.4
Total	*10*	*27.9*	*16*	*34.8*	*10*	*25.0*	*18*	*29.4*	*19*	*25.8*	*17*	*25.0*
Evaluate and design scientific enquiry												
Identify the question explored in a given scientific study.	-	-	-	-	-	-	-	-	-	-	-	-
Distinguish questions that could be investigated scientifically	-	-	-	-	-	-	-	-	-	-	-	-
Propose a way of exploring a given question scientifically	1	2.8	6	13.0	3	7.5	-	-	1	1.4	4	5.9
Evaluate ways of exploring a given question scientifically	-	-	-	-	-	-	-	-	-	-	-	-
Describe and evaluate how scientists ensure the reliability of data and the objectivity and generalizability of explanations	-	-	-	-	-	-	-	-	-	-	-	-
Total	*1*	*2.8*	*6*	*13.0*	*3*	*7.5*	-	-	*1*	*1.4*	*4*	*5.9*
Interpret data and evidence scientifically												
Transform data from one representation to another	-	-	-	-	-	-	-	-	-	-	-	-
Analyze and interpret data and draw appropriate conclusions	3	8.3	10	21.7	5	12.5	4	6.6	6	8.1	3	4.4
Identify the assumptions, evidence, and reasoning in science-related texts.	-	-	-	-	-	-	-	-	-	-	-	-
Distinguish between arguments that are based on scientific evidence and theory and those based on other considerations.	-	-	-	-	-	-	-	-	-	-	-	-
Evaluate scientific arguments and evidence from different sources	-	-	-	-	-	-	-	-	-	-	-	-
Total	*3*	*8.3*	*10*	*21.7*	*5*	*12.5*	*4*	*6.6*	*6*	*8.1*	*3*	*4.4*

Attitudes Aspect

Table 7 gives the frequencies and percentage distribution of objectives for attitudes aspect of science literacy. The objectives mostly focus on environmental awareness with a high percentage in grade four. There is no explicit reference to the aspect interest in science in any grade (see Appendix 5 for typical objectives for attitude dimension).

Table 7.
The frequencies of the objectives in attitude aspect from grade 3 to 8

	Grade 3		Grade 4		Grade 5		Grade 6		Grade 7		Grade 8	
	f	%	f	%	f	%	f	%	f	%	f	%
Interest in science	-	-	-	-	-	-	-	-	-	-	-	-
Valuing scientific approaches to enquiry	2	5.6	1	2.2	1	2.5	4	6.6	-	-	5	7.4
Environmental awareness	4	11.1	8	17.4	5	12.5	2	3.3	6	8.1	5	7.4
Total	*6*	*16.7*	*9*	*19.6*	*6*	*15.0*	*6*	*9.9*	*6*	*8.1*	*10*	*14.8*

DISCUSSION

This research was designed to explore the extent to which 2017 Turkish science curriculum supports the development of scientifically literate students. The objectives from grade 3 to 8 were closely scrutinized for their emphasis given on the aspects of science literacy as defined by PISA. Since science curriculum is the major source that provides the outcomes about what is taught in science classrooms to a great extent, we expect that this study offers some important contribution to both teacher educators and curriculum developers.

There are two main overall findings of this study. One of the most significant findings to emerge from this study is that Turkish science curriculum reflects the aspects of science literacy in varying degrees. That is, the curriculum is not adequate to reflect each of the four dimensions in a balanced manner. The results of this study revealed that the context aspect of the science literacy framework of PISA is almost not included in the objectives of Turkish science curriculum. However, active engagement of students with real-world contexts that affect their lives is seriously highlighted in the definition of science literacy. In other words, the definition of science literacy highlights context-driven curriculum. This context-driven curriculum in science teaching aims to engage students with issues that they are highly likely to come across as citizens (Roberts, 2007). The science curriculum that is contextualized around real-world problems is needed to help students develop a more realistic understanding of the world around them by adapting to a wider social and cultural reality in science classrooms. Lack of a real-world context may lead to a feeling that the concepts learned in schools are not related to daily life. The context-driven science helps students to find out the underlying science concepts for real-world issues (Fensham, 2009). Moreover, choosing the contexts compatible with students' daily lives results in the generation of intrinsic interests (Fensham, 2009). Especially personal and local contexts are valuable to make sense of the world around us and global context to understand others' world.

The second important finding of this study was about the emphasis given on the knowledge aspect. The three subdimensions of knowledge aspect are represented in varying degrees. The most emphasis is given to the content knowledge while the epistemic knowledge is emphasized least. The objectives mostly include the memorizing of the scientific facts and principles. The emphasis on the science content may be important but it should be balanced with procedural and epistemic knowledge. The science curriculum is particularly deficient in providing students with experiences for procedural and epistemic knowledge. In order to meet the needs of the 21st century, the curriculum should be readjusted to underlie these two types of knowledge. The industry world in the 21st century requires people with diverse skills such as creating testable hypothesis, design experiments to test the hypothesis, manipulate variables, and collect data (Duggan & Gott, 2002). Citizens can gain these skills through practicing during K-12 education, primarily in science classrooms. Especially the reforms in science education underline the process of science and students' understanding of how scientific knowledge is produced (e.g., AAAS, 1989, 1993; NRC, 1996). Students should develop certain abilities for scientific inquiry. These abilities cannot be separated from science content and yet there is no need to choose skills over content (Rillero, 1998). It is obvious that procedural knowledge is necessary to do science. This, in turn, will lead to produce first-hand science knowledge. Students cannot link the procedural knowledge and content knowledge if there is less emphasis on procedural knowledge. The objectives of the curriculum are also problematic in terms of epistemic knowledge. Students who experience the Turkish science curriculum do not have adequate epistemic knowledge of science until Grade 8 because it is not highlighted in the curriculum until then. Epistemic knowledge is related to the nature and characteristics of scientific knowledge. As content and procedural knowledge, it is an eminent aspect of science literacy. The reforms in science education have brought the idea of inquiry-based science teaching in order to introduce students to the procedural knowledge and skills in science. An inquiry is defined as complex activity that contains numerous skills such as asking questions, observing the environment, reviewing literature about what is previously known, collecting data, interpreting the evidence, and evaluating alternative solutions (NRC, 1996). Inquiry refers to the way how scientists work and to the methodology of science teaching and learning (Carlson, Humphrey, & Reinhardt, 2003). Therefore, the science curriculum should include objectives promoting inquiry-based practices if it aims to promote science literacy.

Regarding scientific competencies, the Turkish science curriculum emphasizes mostly explain phenomena scientifically and seldom focus on other two aspects -evaluate and design scientific inquiry, interpret data and evidence scientifically. As discussed in the knowledge aspect above, the three competencies should share equal importance. Therefore, students need to experience all of them with similar emphasis. In addition to having knowledge of scientific concepts, science literacy requires the ability to carry out a scientific inquiry. The discussions on science literacy also confirm this viewpoint. For example, when Roberts (2007) suggested the two visions of science literacy, he aimed to emphasize both contents of science (vision I) and the use of these contents (vision II). Therefore, there is a need for a science curriculum that provides students with opportunities to conduct scientific inquiry in which they utilize content, procedural, and epistemic knowledge together. In this way, they can evaluate their results and reach meaningful conclusions.

Among others, the most neglected aspect of science literacy is attitudes aspect. Regarding three subdimensions of this aspect, only the fourth-grade science curriculum includes objectives emphasizing environmental awareness while science curriculum from six to eight grade emphasizes neither of them. The attitudes towards science are generally neglected in science curriculum because the cognitive gains are given more emphasis than affective ones. However, having positive attitudes toward science bring about other outcomes such as science achievement. Moreover, Kirk (2018) underlined that if the learning environment only offers students cognitive opportunity to learn, then it is highly likely that such environment may inhibit students' learning at a certain point in time. If students appreciate science and scientific way of thinking, they become more scientifically literate. A key policy priority should, therefore, be to consider the ways of promoting students' attitudes toward science which, in turn, promotes science literacy.

CONCLUSION and RECOMMENDATIONS

The present study was designed to evaluate the extent to which Turkish science curriculum emphasizes the PISA 2015 science literacy aspects. Overall, the investigation of objectives has shown that there are certain drawbacks of current science curriculum in terms of raising scientifically literate children. The most significant finding of this study is that Turkish science curriculum includes objectives fostering content knowledge more. That is, the curriculum is dominated by the pure knowledge of the content of science. This is definitely an essential element of science education but there is a need to focus on the process of science in classrooms.

The Turkish science curriculum underlines the importance of educating every young person as scientifically literate. However, the evidence from this study indicated that this is not totally reflected in the objectives. Although current science curriculum emphasizes science literacy and include objectives fostering it to some degree, international assessments show that Turkish students (especially 15-year-old ones) are not well-equipped with the elements of science literacy. Therefore, Turkish science curriculum in each grade needs to be redesigned to provide students with opportunities to raise them science literacy in all aspects. It should include objectives fostering each aspect of science literacy in a balanced manner. Lack of such curriculum might be one of the reasons why 15-year-old students in Turkey "fail" in PISA assessments.

REFERENCES

Acat, M. B., Anılan, H., & Anagun, S. S. (2010). The problems encountered in designing constructivist learning environments in science education and practical suggestions. *The Turkish Online Journal of Educational Technology*, *9*(2), 212-220.

American Association for the Advancement of Science [AAAS]. (1989). *Science for all Americans*. New York: Oxford University Press.

American Association for the Advancement of Science [AAAS]. (1990). *Science for all Americans*. New York: Oxford University Press.

American Association for the Advancement of Science [AAAS]. (1993) *Benchmarks for science literacy*. New York: Oxford University Press.

Blandford, S. (2000). *Managing professional development in schools*. London: Routledge.

Bowen, G. A. (2009). Document analysis as a qualitative research method. *Qualitative Research Journal*, *9*(2), 27-40. DOI: https://doi.org/10.3316/QRJ0902027

Carlson, M. O. B., Humphrey, G. E., & Reinhardt, K. (2003). *Weaving science inquiry and continuous assessment: Using formative assessment to improve learning*. California: Corwin Press.

Carlton, R. (1963). On scientific literacy. *NEA Journal*, *52*(4) 33-35.

Cansiz, M., & Turker, N. (2011). Scientific literacy investigation in science curricula: The case of Turkey. *Western Anatolia Journal of Educational Sciences, Special Issue*, 359-366.

Collins, A. (1998). National science education standards: A political document. *Journal of Research in Science Teaching*, *35*(7), 711-727. DOI: https://doi.org/10.1002/(SICI)1098-2736(199809)35:7<711: AID-TEA3>3.0.CO;2-O

Duggan S., & Gott, R. (2002). What sort of science education do we really need? *International Journal of Science Education*, *24*(7), 661-679. DOI: https://doi.org/10.1080/09500690110110133

Durant, J. R. (1993). What is scientific literacy? In J. R. Durant & J. Gregory (Eds.), *Science and culture in Europe* (pp. 129– 137). London: Science Museum.

Durant, J. R., Evans, G., & Thomas, G. P. (1989). The public understanding of science. *Nature*, *340*(6228), 11–14. DOI: https://doi.org/10.1038/340011a0

Eurydice (2018). Basic characteristics of education system: Turkey overview. Retrieved September, 10, from https://eacea.ec.europa.eu/national-policies/eurydice/content/turkey_cs.

Fensham, P. J. (2008). *Science education policy-making*. Paris: UNESCO.

Fensham, P. J. (2009) Real world contexts in PISA science: Implications for context-based science education. *Journal of Research in Science Teaching*, *46*(8), 884–896. DOI: https://doi.org/10.1002/tea.20334

Gallagher, J. (1971). A broader base for science teaching. *Science Education*, *55*(3), 329-338. DOI: https://doi.org/10.1002/sce.3730550312

Hurd, P. D. (1958). Science literacy: Its meaning for American schools. *Educational Leadership*, *16*(1), 13-16.

Kesidou, S., & Roseman, J. E. (2002). How well do middle school science programs measure up? Findings from project 2061's curriculum review. *Journal of Research in Science Teaching*, *39*(6), 522-549. DOI: https://doi.org/10.1002/tea.10035

Kirk, K. (2018). What is the affective domain anyway? Retrieved May 8, 2018, from https://serc.carleton.edu/NAGTWorkshops/affective/intro.html

Kolstø, S. D. (2001). Scientific literacy for citizenship: Tools for dealing with controversial socio-scientific issues. *Science Education*, *85*(3), 291-310. DOI: https://doi.org/10.1002/sce.1011

Millar, R. (1996). Toward a science curriculum for public understanding. *School Science Review*, *77*(280), 7-18.

Miller, J. D. (1995). Scientific literacy for effective citizenship," in R. E. Yager (ed.), *Science/Technology/Society as reform in science education*, pp. 185-204. New York: State University of New York Press.

Miller, J. D. (1998). The measurement of civic scientific literacy. *Public Understanding of Science* *7*(3), 203-223. DOI: https://doi.org/10.1088/0963-6625/7/3/001

Ministry of National Education [MoNE] (2004). *Fen ve teknoloji dersi programı, ilköğretim 4. –5. sınıf [Science and technology curriculum, 4th and 5th grades]*. Ankara.

Ministry of National Education [MoNE] (2013). *Fen bilimleri dersi öğretim programı, 3. 4. 5. 6. 7. ve 8. sınıflar) [Science curriculum, 3rd, 4th, 5th, 6th, 7th, and 8th grades]*. Ankara.

Ministry of National Education [MoNE] (2017). *Fen bilimleri dersi öğretim programı, 3. 4. 5. 6. 7. ve 8. sınıflar) [Science curriculum, 3rd, 4th, 5th, 6th, 7th, and 8th grades]*. Ankara.

National Research Council [NRC]. (1996). *National science education standards*. Washington, DC: National Academy Press.

Organisation for Economic Cooperation and Development [OECD]. (2013). PISA 2012 results: What makes schools successful? Resources, policies and practices (Volume IV). Paris: OECD Publishing. DOI: http://dx.doi.org/10.1787/9789264201156-en

Organisation for Economic Cooperation and Development [OECD]. (2016). *PISA 2015 results (Volume I): Excellence and equity in education*. Paris: OECD Publishing. DOI: https://dx.doi.org/10.1787/9789264266490-en

Organisation for Economic Cooperation and Development [OECD]. (2017). *PISA 2015 Assessment and analytical framework: Science, reading, mathematic, financial literacy and collaborative problem solving.* Paris: OECD Publishing. DOI: https://doi.org/10.1787/9789264281820-en

Rillero, P. (1998). Process skills and content knowledge. *Science Activities*, *35*(3), 3-4. DOI: https://doi.org/10.1080/00368129809600910

Roberts, D. A. (2007). Scientific literacy/science literacy. In S. K. Abell & N. G. Lederman (Eds.), *Handbook of Research on Science Education* (pp. 729-780). London: Lawrence Erlbaum Associates.

Rutherford, F. J., & Ahlgren, A. (1990). *Science for all Americans*. New York: Oxford University Press.

Rudolph, J. L., & Horibe, S. (2015). What do we mean by science education for civic engagement? *Journal of Research in Science Teaching, 53*(6), 805–820. DOI: https://doi.org/10.1002/tea.21303

Şad, S. N. (2012). Investigation of parental involvement tasks as predictors of primary students' Turkish, math, and science & technology achievement. *Eurasian Journal of Educational Research*, *48*, 135-154.

Van Driel, J., Beijaard, D., & Verloop, N. (2001). Professional development and reform in science education: The role of teachers' practical knowledge. *Journal of Research in Science Teaching, 38*(2), 137-158. DOI: https://doi.org/10.1002/1098-2736(200102)38:2<137: AID-TEA1001>3.0.CO;2-U

Yager, R. (1986). Searching for excellence. *Journal of Research in Science Teaching, 23*(3), 209-217. DOI: https://doi.org/10.1002/tea.3660230305

Yorulmaz, Y. İ., Çolak, İ., & Ekinci, C. E. (2017). An evaluation of PISA 2015 achievements of OECD countries within income distribution and education expenditures. Turkish Journal of Education, 6(4), 169-185. DOI: 10.19128/turje.329755

APPENDIX 1.

Aspects, subdimensions, and practices in PISA scientific literacy framework

Aspects	Subdimensions	Practices
Contexts	Personal Local Global	Health and Disease Natural Resources Environmental Quality Hazards Frontiers of Science and Technology
Knowledge	Knowledge of the Content of Science	Physical systems Living System Earth and Space System
	Procedural Knowledge	The concept of variables Concepts of Measurements Ways of assessing and minimizing uncertainty Mechanisms to ensure the replicability of data Common ways of abstracting and representing data The control-of-variables The nature of an appropriate design for a scientific question
	Epistemic Knowledge	The nature of scientific observations, facts, hypotheses, models, and theories The purpose and goals of science as distinguished from technology The values of science The nature of reasoning How scientific claims are supported by data The function of different forms of empirical enquiry How measurement error affects the degree of confidence The use and role of abstract models and their limits The role of collaboration and critique The role of scientific knowledge, along with other forms of knowledge
Competencies	Explain phenomena scientifically	Recall and apply appropriate scientific knowledge. Identify, use and generate explanatory models and representations Make and justify appropriate predictions Offer explanatory hypotheses Explain the potential implications of scientific knowledge for society
	Evaluate and design scientific enquiry	Identify the question explored in a given scientific study. Distinguish questions that could be investigated scientifically Propose a way of exploring a given question scientifically Evaluate ways of exploring a given question scientifically Describe and evaluate how scientists ensure the reliability of data and the objectivity and generalizability of explanations
	Interpret data and evidence scientifically	Transform data from one representation to another Analyze and interpret data and draw appropriate conclusions Identify the assumptions, evidence, and reasoning in science-related texts. Distinguish between arguments that are based on scientific evidence and theory and those based on other considerations. Evaluate scientific arguments and evidence from different sources
Attitudes	Interest in science Valuing scientific approaches to enquiry Environmental awareness	

APPENDIX 2.

Typical objectives in the context aspect of science literacy

Contexts	Typical Objectives
Personal	
Health and Disease	Discuss what should be done to protect the health of the sense organs (Objective number: F.3.2.1.3).
Natural Resources	No objectives
Environmental Quality	Be careful to be efficient in the use of resources (Objective number: F.4.6.1.1).
Hazards	Discuss the dangers of moving objects in daily life (Objective number: F.3.3.2.3).
Frontiers of Science and Technology	No objectives
Local	
Health and Disease	Assume responsibility to reduce smoking in the close vicinity (Objective number: F.4.2.1.6.).
Natural Resources	Discuss the separation of mixtures in terms of their contribution to national economy and effective use of resources (Objective number: F.4.4.5.3.).
Environmental Quality	Discuss the effect of battery waste on environment and what should be done about this (Objective number: F.3.7.2.2.).
Hazards	No objectives
Frontiers of Science and Technology	Explain his/her ideas on the new applications of magnets (Objective number: F.4.3.2.4).
Global	
Health and Disease	No objectives
Natural Resources	Discuss the importance of density of solid and liquid water for the livings
Environmental Quality	Explain the negative effects of light pollution on natural life and observation of celestial bodies (Objective number: F.4.5.3.2).
Hazards	Discuss the causes and possible consequences of global climate change (Objective number: F.8.6.3.5)
Frontiers of Science and Technology	No objectives

APPENDIX 3.

Typical objectives in the knowledge aspect of science literacy

Knowledge	Typical Objectives
Knowledge of the Content of Science	
Physical systems	Classifies the substances according to their state (Objective number: F.3.4.2.1).
Living System	Explain the basic functions of sensory organs. (Objective number: F.3.2.1.2).
Earth and Space System	Explains the events that occurs as a result of Earth's motion (Objective number F.4.1.2.2).
Procedural Knowledge	
The concept of variables	Tests and predicts the variables that affects bulb brightness in an electrical circuit. (Objective number F.5.7.2.1).
Concepts of Measurements	Compares the mass and volume of different substances by measuring them (Objective number F.4.4.2.1)..
Ways of assessing and minimizing uncertainty	No objectives
Mechanisms to ensure the replicability of data	No objectives
Common ways of abstracting and representing data	Interpret the factors affecting the rate of photosynthesis by drawing graphs (Objective number F.8.6.2.3)
The control-of-variables	Discover by doing experiment that the heat energy of a matter depends on the type, mass, and temperature of the matter (Objective number F.8.4.5.1)
The nature of an appropriate design for a scientific question	No objectives
Epistemic knowledge	
The nature of scientific observations, facts, hypotheses, models, and theories	No objectives
The purpose and goals of science as distinguished from technology	No objectives
The values of science	No objectives
The nature of reasoning	No objectives
How scientific claims are supported by data	Questions how the ideas about the concept of atom has changed from past to present (Objective number F.7.4.1.2)
The function of different forms of empirical enquiry	No objectives
How measurement error affects the degree of confidence	Discusses the ideas about the structure of the cell from past to present by associating with technological developments. (Objective number F.7.2.1.2)
The use and role of abstract models and their limits	No objectives
The role of collaboration and critique	No objectives
The role of scientific knowledge, along with other forms of knowledge	No objectives

APPENDIX 4.

Typical objectives in the scientific competencies aspect of science literacy

Scientific Competencies	Typical Objectives
Explain phenomena scientifically	
Recall and apply appropriate scientific knowledge.	Prepares solution by using the solutes and solvents from daily life substances (Objective number F.7.4.3.2).
Identify, use and generate explanatory models and representations	Designs an imaging tool using mirrors or lenses (Objective number F.7.5.3.5)
Make and justify appropriate predictions	Predicts and tests the environments in which sound can propagate. (Objective number F.6.5.1.1)
Offer explanatory hypotheses	Demonstrates the relationship between the sense of smell and taste by designing experiment (Objective number F.6.6.2.2)
Explain the potential implications of scientific knowledge for society	Discuss the separation of mixtures in terms of their contribution to the national economy and the effective use of resources. (Objective number F.4.4.5.3)
Evaluate and design scientific enquiry	
Identify the question explored in a given scientific study.	No objectives
Distinguish questions that could be investigated scientifically	No objectives
Propose a way of exploring a given question scientifically	Offers solutions to prevent acid rain (Objective number F.8.4.4.7)
Evaluate ways of exploring a given question scientifically	No objectives
Describe and evaluate how scientists ensure the reliability of data and the objectivity and generalizability of explanations	No objectives
Interpret data and evidence scientifically	
Transform data from one representation to another	No objectives
Analyze and interpret data and draw appropriate conclusions	Discusses the importance of freshness and naturalness of foods for a healthy life based on research data. (Objective number F.4.2.1.3.)
Identify the assumptions, evidence, and reasoning in science-related texts.	No objectives
Distinguish between arguments that are based on scientific evidence and theory and those based on other considerations.	No objectives
Evaluate scientific arguments and evidence from different sources	No objectives

APPENDIX 5.

Typical objectives in the attitudes aspect of science literacy

Attitudes	Typical Objectives
Interest in science	No objectives
Valuing scientific approaches to enquiry	Discuss the consequences of consanguineous marriages. (Objective number F.8.2.2.3)
Environmental awareness	Takes an active role in the cleaning of the environment. (Objective number F.3.6.2.2)

TÜRKÇE GENİŞLETİLMİŞ ÖZET

Uluslararası Öğrenci Değerlendirme Programı (PISA) sonuçları, Türkiye'de eğitim gören 15 yaşındaki öğrencilerin bilgi toplumunun ihtiyaçlarını karşılayacak düzeyde fen okuryazarı olmadıklarına dair kanıtlar sağlamıştır. Türkiye'de 15 yaşındaki öğrencilerin büyük bir çoğunluğu (%96,7) ortaöğretime devam etmektedir. Yani, Türkiye'de 15 yaşındaki öğrencilerin yaklaşık %97'si, ilkokul ve ortaokul fen programının gereksinimlerini tamamlamıştır. Halen uygulanmakta olan Fen Bilimleri Dersi Öğretim Programı'nın (ilkokul ve ortaokul 3, 4, 5, 6, 7 ve 8. sınıflar) temel amaçlarından biri fen okuryazarı bireyler yetiştirmektir. Bu nedenle bu programı başarıyla tamamlamış öğrencilerden, fen okuryazarlık düzeyini değerlendirme amacı taşıyan PISA'da yüksek performans göstermeleri beklenir. Bu beklentinin aksine, 2015 PISA uygulamasında Türkiye fen okuryazarlık puanına göre 35 OECD ülkesi arasında 34. sırada yer almıştır. Ayrıca aynı sınavda ileri düzey bilimsel düşünme ve akıl yürütme becerileri gösterebilen öğrencilerin oranı, toplam öğrencilerin % 1'inden daha azdır.

Türkiye'nin PISA'da elde ettiği görece başarısız sonuçlara bir takım makul açıklamalar getirmek mümkündür. Bunların bazıları eğitim-öğretimde ebeveyn desteğinin eksikliği, eğitime özgü kaynakların yetersizliği, okullar arası ciddi başarı farklılıkları ve öğrencileri ulusal sınavlara hazırlama kaygısı olarak sıralanabilir. Bunların yanında görece başarısızlığın bir nedeni de, fen okuryazarı öğrenci yetiştirmek için gereken bileşenleri tam karşılayamayan öğretim programları olabilir. Öğretmenler öğretim programlarını kazanımları belirlemek, içerikleri hazırlamak, etkinliklere karar vermek gibi farklı amaçlarla kullanırlar. Ayrıca öğretim programları bir öğretmenin ne öğreteceğine, nasıl öğreteceğine, ne zaman öğreteceğine, nerede öğreteceğine ve hatta neden öğreteceğine karar vermesi adına bir rehber niteliği taşır. Bu sebeple öğretim programlarının çeşitli yönleriyle analiz edilmesi önemlidir. Bu noktadan hareketle ortaya koyulan bu çalışmanın amacı, 2017 yılında yayımlanan ve halen uygulanmakta olan fen bilimleri öğretim programının fen okuryazarı öğrenci yetiştirme potansiyelini, PISA Fen Okuryazarlığı Değerlendirme Çerçevesi kullanarak analiz etmektir.

Fen okuryazarlığı, PISA Fen Okuryazarlığı Değerlendirme Çerçevesinde en genel haliyle bağlamlar, bilgi, yeterlikler ve tutumlar olmak üzere dört boyutta ele alınmaktadır. Her bir alt boyut, ilgili yapıyı ortaya koyacak şekilde detaylandırılmıştır. Örneğin bu çerçeve; yeterlikler boyutunu "olguları bilimsel olarak açıklama", "bilimsel sorgulama yöntemi tasarlama ve değerlendirme" ve "verileri ve bulguları bilimsel olarak yorumlama" olarak açıklamıştır. Bu çalışmada bu çerçeve kullanılarak 3. sınıftan 8. sınıfa kadar uygulanmakta olan fen bilimleri öğretim programının bütün kazanımları analiz edilmiştir.

PISA Fen Okuryazarlığı Değerlendirme Çerçevesinde bağlamlar; kişisel, yerel/ulusal ve küresel olmak üzere üç farklı kategoride ele alınmıştır. Fen bilimleri öğretim programı bu açıdan incelendiğinde, programda yeterli miktarda bağlam temelli kazanım bulunmadığı dikkat çekmektedir. Öğretim programında bağlamı kişisel, yerel/ulusal ve küresel ölçekte sorunlar olan az sayıda kazanım vardır. Bu kazanımlar içinde kişisel bağlamda sağlık sorunlarını içeren kazanımlar, diğer kazanımlara kıyasla daha fazladır. Öte yandan küresel bağlamın alt boyutlarından çevresel kalite, kazanımlarda diğerlerinden daha fazla vurgulanmaktadır. Doğal kaynaklar ise en çok yerel/ulusal bağlamda vurgulanmaktadır.

PISA Fen Okuryazarlığı Değerlendirme Çerçevesinde bilgi; içerik bilgisi, süreçsel bilgi ve epistemik bilgi olmak üzere üç farklı kategoride ele alınmıştır. Fen bilimleri öğretim programı bilgi yönünden incelendiğinde 3. sınıftan 8. sınıfa kadar bütün sınıflarda süreçsel bilgi, içerik bilgisine kıyasla kazanımlarda daha az vurgulanmıştır. Süreçsel bilgi boyutunda yer alan ölçüm konuları, kazanımlar içerisinde diğer becerilere göre kendisine daha fazla yer bulabilmiştir. Diğer yandan kazanımlar epistemik bilgi yönünden oldukça yetersiz kalmıştır. Fen bilimleri dersi öğretim programında epistemik bilgiyi doğrudan geliştirmeye yönelik hiçbir kazanıma 4, 5, 6 ve 8. sınıf düzeylerinde rastlanmamıştır.

PISA Fen Okuryazarlığı Değerlendirme Çerçevesinde yeterlikler; olguları bilimsel olarak açıklama, bilimsel sorgulama yöntemi tasarlama ve değerlendirme, verileri ve bulguları bilimsel olarak yorumlama olmak üzere üç farklı kategoride ele alınmıştır. Kazanımlarda en çok vurgulanan yeterlik, olguları bilimsel olarak açıklama şeklinde karşımıza çıkmaktadır. Diğer iki yeterlik öğretim programının kazanımlarında nadiren vurgulanmaktadır.

PISA Fen Okuryazarlığı Değerlendirme Çerçevesinde tutumlar; fen bilimlerine duyulan ilgi, bilimsel sorgulama yöntemlerine verilen değer ve çevresel farkındalık olmak üzere üç farklı kategoride ele alınmıştır. Program kazanımları bu üç kategori arasında en fazla çevresel farkındalığa odaklanmıştır. Kazanımlarda tutumlar açısından gözlemlenen önemli noktalardan biri de, farklı sınıf düzeylerinde fen bilimlerine duyulan ilgiye yönelik doğrudan vurgu yapan hiçbir kazanım bulunmamasıdır.

Özetle, bu çalışmanın iki genel bulgusu vardır: Ortaya konulan önemli bulgulardan ilki Türkiye'de uygulanan Fen Bilimleri Dersi Öğretim Programı'nın fen okuryazarlığının dört boyutunu dengeli bir şekilde yansıtmada yetersiz kaldığıdır. İkinci önemli bulgu ise fen bilimleri dersi öğretim programının daha çok içerik bilgisine yoğunlaştığıdır. İçerik bilgisi elbette herhangi bir programın önemli bileşenlerinden biridir. Fakat süreçsel ve epistemik bilgiye de içerik bilgisi ile kıyaslanabilecek ölçüde programda yer verilmelidir. Fen bilimleri dersi öğretim programı öğrencilere süreçsel ve epistemik bilgiyi de kazandırabilecek şekilde tasarlanmalıdır. Bu iki bilgi içerik bilgisiyle birlikte 21. yüzyılın ihtiyaç duyduğu becerilere sahip bireyleri yetiştirmede önem arz etmektedir. Günümüz sanayi dünyası hipotez kurabilen, bu hipotezleri sınayabilecek deneyler tasarlayabilen, amaca yönelik veri toplayabilen bireylere ihtiyaç duymaktadır. Bireyler bu becerileri en iyi zorunlu eğitim kademesi boyunca, özellikle fen bilimleri dersinde, ilk elden deneyimleyerek kazanabilirler. Bu deneyimleri kazandırmayı hedefleyen bir fen bilimleri dersi öğretim programının olması, fen okuryazarı bireyler yetiştirmede atılacak önemli adımlardan biridir. Sonuç olarak, gelecekte yapılacak program güncelleme ve geliştirme çalışmalarında, fen okuryazarlığın dört boyutunu dengeli bir şekilde yansıtmak ve içerik bilgisinin yanında süreçsel ve epistemik bilgiye de benzer vurgular yapmak, PISA benzeri uluslararası sınavlarda başarımızı artıracağı gibi fen okuryazarı öğrenciler yetiştirmede mesafeleri daha hızlı kat etmemize olanak sağlayacaktır.