Id: 916185

Participatory Educational Research (PER)
Vol.9(2), pp. 416-436, March 2022
Available online at http://www.perjournal.com
ISSN: 2148-6123
http://dx.doi.org/10.17275/per.22.47.9.2

Comparing the Automatic Item Selection Procedure and Exploratory Factor Analysis in Determining Factor Structure

Asiye Şengül Avşar[*]
Measurement and Evaluation in Education, Recep Tayyip Erdoğan University, Rize, Turkey
ORCID: 0000-0001-5522-2514

Article history
Received:
14.04.2021

Received in revised form:
01.10.2021

Accepted:
14.10.2021

Key words:
Construct validity;
Principal component analysis;
Principal axis factoring;
Mokken scale analysis;
Automatic item selection procedure

It is necessary to supply proof regarding the construct validity of the scales. Especially, when new scales are developed the construct validity is researched by the Exploratory Factor Analysis (EFA). Generally, factor extraction is performed via the Principal Component Analysis (PCA) which is not exactly factor analysis and the Principal Axis Factoring (PAF) among EFA methods. Factors may also be determined with different techniques depending on the advances in psychometry. In the context of nonparametric item response theory, the Mokken Scale Analysis (MSA) and the Automatic Item Selection Procedure (AISP) provide significant contributions to researchers in scale development studies. The aim of the current study is to compare the AISP and the EFA methods in determining the factor structures. The Revised Life Orientation Test, whose factor structure was previously known and validated, and the draft Expectation Scale from Academics in Distance Education Scale, which was at the preliminary stage of the scale development process with the unknown factor structure, were considered in this comparison. It was determined that the consistency of the findings obtained from the EFA, and the ones obtained from the AISP provided strong evidence in ensuring the construct validity. The PAF and the AISP produced similar results in this research. The PAF results should be taken into consideration instead of the PCA, especially in scale development studies. It is recommended that the AISP and the PAF results be evaluated together for strong evidence in the investigation of construct validity in scale development studies.

Introduction

Many cognitive and affective characteristics of human beings which are known as latent variables cannot be observed and measured directly. Therefore, indirect measurements are made with various scales developed. It is necessary to determine the psychometric features of these scales that are used to measure the latent variables in order to make correct inferences about individuals. Measurements made with measurement tools that do not give valid and reliable results, will not make valid and consistent measurements for the purpose. Measurements made with these tools will affect the evaluation decisions and cause erroneous decisions about individuals. Researchers are required to use the measurement tools that give

[*] Correspondency: asiye.sengul@erdogan.edu.tr

valid and reliable results in accordance with the feature desired to be measured while measuring features that cannot be observed directly.

There are many features of individuals. Among these features, psychological ones are called constructs. Constructs are the products of theories developed by scientists who have tried to make sense of human behaviours (Crocker & Algina, 1986). Strong evidence should be presented especially regarding the construct validity of the measurement tools used to measure the constructs. At this point, the main concern is the existence of a measurement tool that measures the theoretical construct to be measured validly and reliably. Construct validity is related to the ability of the measurement tools to measure the latent traits that are theoretically stated accurately. In other words, construct validity provides evidence regarding how valid the measurement of the target construct is (Anastasia, 1988; VandenBos, 2015).

There may be a need for new scales according to changing needs, situations, and developments. Scale development studies are empirical studies with scientific processes that need to be followed. These studies are often costly, time-consuming, and tiring. Scale development steps are generally similar in different sources in the literature. These steps are summarized in Figure 1 depending on Crocker and Algina (1986), Cohen and Swerdlik (2009), and DeVellis (2017) sources.

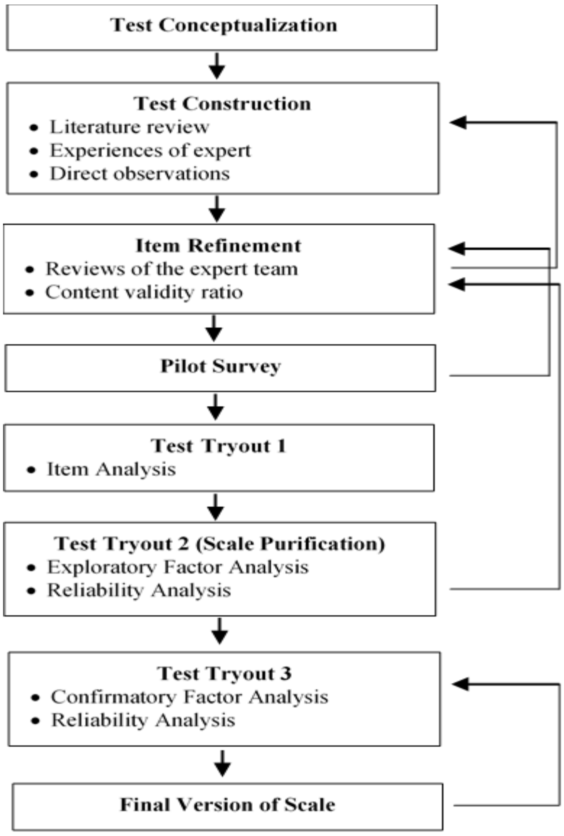

Figure 1. An example of scale development design process

The test conceptualization stage is the phase when the purpose of measurement tools is determined. At this stage, it is clarified whether it is truly necessary to develop the relevant measurement tools or not. The test construction is the phase that the structure to be measured is defined and the items that are indicators of the latent traits to be measured are written. Literature review results, expert opinions, and direct observation results are used for this very purpose. It is also determined how the items will be scored or what type of scale (Thurstone, Likert, Guttman, and alike) will be used.

The item refinement step includes the examination of the written items by field experts, psychometrics, psychologists, and language experts, and the removal of items that are similar in content or not required from the candidate scale. The content validation ratio (CVR) and content validation index (CVI) which are recommended by Lawshe (1975) are calculated while determining the suitability of the written items to the content. During the pilot survey phase, the draft items are applied to a small group with similar characteristics to the group intended to apply the scale. At this stage, participants' reactions such as whether there are parts they do not understand and whether they pause while reading the items should be observed and the items should be negotiated. If necessary, the items should be revised, and expert opinions should be obtained on these items again.

Test tryout-1 phase can be considered as a trial phase. Descriptive statistics of items, item-total correlations, etc. can be calculated at this stage. This stage may provide the scale developer some clues before factor analysis. If needed, draft items can be re-examined. Proofs regarding the construct validity of the scores obtained from the scale is collected with the Exploratory Factor Analysis (EFA) in the test tryout-2 (scale purification) stage. At this stage, the factor structure is determined using the EFA methods in accordance with the way the items are scored (binary or polytomous) and the level of information (order or interval) obtained from the scale. Test tryout-1 and test tryout-2 can be combined and so, can be considered as one-stage. However, performing these two steps separately in scale development studies may allow stronger proof for construct validity.

The factor structure of the measurement tool should be confirmed with the Confirmatory Factor Analysis (CFA) on a different sample at the test tryout-3 stage. After all these steps, the scale can be finalized. In addition to these processes, it should be investigated whether the items in the measurement tool have a bias or not with techniques such as Differential Item Functioning and Multi-Group CFA. In addition, the developed scale should be monitored over time, and if necessary, it should be revised by taking into account the appropriate statistical steps.

When Figure 1 is checked, it is seen that there are arrows directing backward. If there is a problem at any stage, the previous step or steps should be checked. As can be understood from Figure 1, scale development steps include complex and intensive processes.

In many fields in the literature, it is observed that many scales have been developed to measure many variables. Researchers present the evidence regarding the construct validity in scale development with the EFA. Generally, the Principal Component Analysis (PCA) method is used while determining factor structures. But different factor extraction methods like the Principal Axis Factoring (PAF) can be used for factor extraction. In fact, the PCA is not seen as a factor analysis in the full sense (Field, 2009; Howard, 2016). The latent variable (factor) directs the items that are the indicators of the latent variable in the EFA. In other words, items written according to a theory are divided into factors according to the theory adopted. The observed variables form the components in the PCA. Here, there is a combination of items that are independent of theory and highly related to each other (Matsunaga, 2010).

Contributions made to test theories depending on the developments in psychometrics affect the scale development processes undoubtedly. In addition to the EFA conducted within the context of the Classical Test Theory (CTT), scale development studies are also carried out according to the modern test theory, in other words, the Item Response Theory (IRT). Within the context of the Parametric Item Response Theory (PIRT), when the model-data fit is provided, the estimation of item and ability parameters independently from each other is expressed as one of the most important advantages of this theory in comparison with the CTT (De Ayala, 2009; Hambleton, Swaminathan, & Rogers, 1991).

With its invariance property, the PIRT is useful in many applications such as scale development, computer adaptive testing applications, bias studies, test equating, and item mapping (Hambleton & Swaminathan, 1985). Reaching measurement tools that give more reliable results with fewer items in the PIRT is one of the most important advantages of this theory (Embretson & Reise, 2000). However, strong assumptions of the PIRT and the model-data fit must be provided to accomplish for these advantages. In addition, it is necessary to collect data from large samples for all these, which may be considered as a limitation. At this point, the psychometric properties of the tests can be determined with fewer items and smaller samples with the Nonparametric Item Response Theory (NIRT) (Stout, 2001). Among the NIRT models, the Mokken Homogeneity Model (MHM) and Double Monotonicity Model (DMM) are the most frequently used models (Sijtsma & Molenaar, 2002).

Investigating other approaches within the scope of the IRT in addition to the classical factor analytical approaches within the scope of the CTT as evidence for construct validity in scale development will allow more robust and diverse evidence for construct validity. According to the literature review, there are studies considering the results of the EFA and the Mokken Scale Analysis (MSA) in the investigation of factor structures. Some of these studies are exemplified below by grouping them by their subject content:

- Construct validity of the scales whose factor structure was previously determined was re-investigated in different groups (de Cock, Emons, Nefs, Pop, & Pouwer, 2011; Emons, Sijtsma, & Pedersen, 2012; Lee, Fu, Liu, & Hung, 2017; Shenkin, Watson, Laidlaw, Starr, & Deary, 2014; Ski, Thompson, Hare, Stewart, & Watson, 2012; Stewart, Allison, Baron-Cohen, & Watson, 2015; Watson, van der Ark, Lin, Fieo, Deary, & Meijer, 2012; Wismeijer, Sijtsma, van Assen, & Vingerhoets, 2008).
- Scales were adapted to different cultures (Bagnasco, Watson, Zanini, Rosa, Rocco, & Sasso, 2015; Bech, Carrozzino, Austin, Møller, & Vassend, 2016).
- New scale development studies were conducted (Chen, Watson, & Hilton, 2016; Gudergan, Mathies, Kyngdon, & Kozicki, 2004; Mooij, 2012).
- Short versions of the scales were developed (Lee, Chen, Jiang, Chu, Chiu, Chen, & Chen, 2016; Yoon, Shaffer, & Bakken, 2015).

The MSA is suitable for the exploratory dimensionality analysis (Emons et al., 2012). In general, when the PCA is compared to the MSA, the MSA has important advantages to the PCA. According to Emons et al. (2012), these advantages are:

- The MSA is one of the NIRT models which is a modern test theory. It can be easily used in determining the factor structures of the scales or the scale dimensionality, like the PCA, which is accepted from the EFA scope by most researchers.
- The MSA tests the assumptions underlying IRT models that can support the evaluation of psychometric properties of scales. The PCA calculates eigenvalues based on an inter-item correlation matrix. However, it cannot test underlying

dimensionality models such as IRT. In this context, the MSA can determine the number of underlying dimensions by choosing unidimensional Mokken scales in providing proof for the construct validity of the scale.

- The PCA is more appropriate for the analysis of continuous data while the MSA is more appropriate for the Likert-type discrete questionnaire data.

The MSA seems to have important advantages over both the PIRT models and the EFA in exploring the dimension structure in scale development. In the MSA, the determination of dimensionality is made by the Automatic Item Selection Procedure (AISP). The AISP determines different sets of items (or item clusters) that measure different structures (Emons et al., 2012). These item clusters provide information about the dimensionality of the scale (Sijtsma & Molenaar, 2002). The general aim of the current study is to compare the results obtained from the EFA, which is frequently used in scale development, with the results obtained from the AISP.

In line with the determined general purpose, two different data sets were examined, and answers were sought for the following research questions. The first data set is the Revised Life Orientation Test (LOT-R) data obtained from the "Dutch Longitudinal Internet Studies for the Social Sciences (LISS) panel (www.lissdata.nl)". The other data set was obtained from the application of the draft-Expectation Scale from Academics in Distance Education (ESADE), which is in the preliminary scale development process. Accordingly, the following questions were answered.

(1) What is the factor structure of the LOT-R, whose scale structure is predetermined, according to the PAF and the AISP methods?
(2) What is the factor structure of the ESADE, which is in the preliminary stage in scale development, according to the PCA, the PAF, and the AISP methods?
(3) Are the findings obtained from the PAF and the AISP methods consistent with each other?

Method

Participants

The first dataset consisted of 5,859 individuals who fully responded to the LOT-R, which was obtained from the LISS panel 2020. Individuals aged at least 16-year-old were included in the LISS application. The second dataset is the dataset regarding the ESADE, which is in the preliminary phase of the scale development process within the scope of the relevant research. The ESADE was applied to 313 participants, but 24 of them who did not provide the assumptions were excluded from the dataset. The demographic information of the participants whose data were evaluated for ESADE is given in Table 1.

Table 1. Demographic characteristics of the participants for ESADE

Variables		N	%
Gender	Female	217	75.10
	Male	72	24.90
University	Recep Tayyip Erdoğan University	280	96.90
	Other	9	3.10
Faculty	Education	123	42.60
	Theology	155	53.60
	Other	11	3.80
Grade	1	14	4.80
	2	49	17.00
	3	163	56.40
	4	63	21.80
Total		289	100.00

There are 289 participants who responded to the ESADE. It is seen that most of the participants (75.10%) are female. Almost all the participants (96.90%) are from Recep Tayyip Erdoğan University, and most of the participants are (96.20%) attending the faculty of education and theology. Also, more than half of the students (56.40%) are at the 3rd-grade level.

Data Collection Tools

The Revised Life Orientation Test (LOT-R)

The first measurement tool used in this research is the LOT-R. The LOT-R, which was developed by Scheier, Carever, and Bridges (1994), is a five-point Likert type scale, whose four items are fillers and therefore are not included in the scoring. The LOT-R consists of two factors, which are Optimism and Pessimism. There are three items in each factor. It is necessary to analyze the items in the Optimism scale by re-coding them.

The Expectation Scale from Academics in Distance Education (ESADE)

A new draft scale whose factor structure was unknown, and which was still under development was considered in line with the aim of the current study. In this study, the steps up to the test tryout-3, as stated in the scale development steps given in Figure 1, were followed and these steps followed are described below in order.

(1) Test Conceptualization: Is this scale really needed? As it is known, the COVID-19 pandemic had significant effects on education. There has been a transition from face-to-face education to online education. Determining students' expectations from academicians and measuring their ideal academician perceptions takes an important place in the literature (Aljubaily, 2010). Since no scales developed specifically for the distance education process have been encountered that measure students’ expectations from academicians in distance education in the literature review at present and it is considered important to measure students’ expectations to increase the quality of education, developing the ESADE has become a necessity.

(2) Test Construction: A group of volunteer students was asked to write an essay about their expectations from academicians in distance education, and these essays were examined by the researcher. The literature was examined and the items that were included in the previously developed scales and that could be used in the research were determined. First, 25 items were prepared.

(3) Item Refinement: A team of five experts, including the fields of classroom education, language education, psychological counselling and guidance, computer and instructional technology education, and measurement and evaluation, examined the items written. The CVR and the CVI values were calculated by considering the evaluations of the experts regarding the examination of the items. According to these values, 18 items remain on the draft scale. Sample items can be found in the appendix.

(4) Pilot Survey: Ready-to-use draft items were applied to a group of 10 volunteer students. The students were consulted about whether there was a part in the items they did not understand, and it was determined that there was no problem with the items.

(5) Test Tryout-1 and Test Tryout-2: This study was analyzed at the meeting held by Recep Tayyip Erdoğan University Senate Ethics Commission on 04 January 2021 and was found ethically appropriate (Letter dated 04.01.2021 and numbered 2021/186). After getting the necessary ethical permissions, the draft scale was applied online, completely on a voluntary basis. The analysis process has been carried out in this step.

The procedure

Exploratory Factor Analysis (EFA)

One of the most important steps in scale development studies is the EFA where the factor structure is determined. The EFA makes the dataset understandable by grouping the variables related to each other to define and summarize the data (Tabachnick & Fidell, 2014). The structure between the variables that are analyzed is revealed in the EFA (Hair, Black, Babin, & Anderson, 2019).

Various factor extraction methods are performed with the EFA. Since it is easy to interpret, the PCA which is not actually a complete factor analysis is frequently performed (Field, 2009, p. 638; Howard, 2016). The PCA is the appropriate method when the main purpose is data reduction, in other words, to explain the maximum variance with the minimum factor. That is why it is ideal to summarize the data (Hair et al., 2019). Components are produced as a result of the PCA (Tabachnick & Fidell, 2014).

One of the factor analytical techniques which is in the EFA that can be used in scale development is the PAF. The main purpose of the PAF is to reproduce the correlation matrix with a few orthogonal factors from the dataset (Tabachnick & Fidell, 2014, p. 688). Factors are produced as a result of the PAF. The factors obtained from the PAF represent the basic structures that summarize or explain the observed variables (Hair et al., 2019).

There are differences between the PCA and the PAF statistically. While the PCA analyses the variance, the PAF analyses the covariance (Tabachnick & Fidell, 2014, p. 688). While all variances in the observed variables are analyzed in the PCA, only shared variance is analyzed in the PAF. The main purpose of the PCA is to extract the highest variance from a dataset with a few orthogonal components.

It is not easy to interpret the PCA or the PAF results alone. Therefore, rotation is needed. There are different rotation techniques such as orthogonal rotation and oblique rotation (Field, 2009). Varimax method which is one of the orthogonal rotation methods enables a highly fewer number of items to be loaded on each factor (Field, 2009, p.644) and researchers reach the factor sets that are easy to interpret (Howard, 2016). The correlation between factors is allowed in the oblique rotation. While promax method provides faster results in large samples, direct oblimin is a rotation technique that can be used in small samples (Field, 2009).

It is assumed that there is no relationship between factors in the orthogonal rotation. The results are easier to interpret than the ones obtained in the oblique rotations. However, this assumption is not considered realistic in social sciences (Tabachnick & Fidell, 2014). If researchers in social sciences adopt the oblique rotation technique to reveal the real factor structures, it will help them reach statistically correct results.

Mokken Scale Analysis (MSA) and Automatic Item Selection Procedure (AISP)

Mokken Models among the NIRT models are divided into two as the MHM and the DMM. Mokken models order the individuals based on their total scores when the assumptions are provided (Sijtsma & Molenaar, 2002).

Unidimensionality, local independence, and monotonicity of item characteristic curve (ICC) are the assumptions of the MHM (Sijtsma & Molenaar, 2002). There is an additional assumption for the DMM. The DMM requires non-intersecting ICC. In Mokken models, the *H* scalability coefficient is interpreted in determining whether the data fit the model or whether the items order the individuals correctly.

The minimum value of the *H* is zero, the maximum value of the *H* is one (Mokken, 1997; Sijtsma & Molenaar, 2002). The criteria for the *H* coefficients are stated as $.30 \leq H < .40$ weak, $.40 \leq H < .50$ medium, and $H \geq .50$ high. In a test scaled according to the Mokken model, the *H* coefficient values, and the criteria mentioned above for them are used in the item selection (Meijer & Banake, 2004; Mokken, 1997). The *H* coefficients can also be interpreted as the item discrimination index. The higher the values of the *H* coefficient are, the better- the more correctly individuals are ordered (Sijtsma & Molenaar, 2002).

Mokken analyses also include the AISP. AISP method is a procedure that creates various item sets or item clusters. Item sets/clusters and item selections in the AISP are performed according to the determined cut-off (*c*) value (Sijtsma & Molenaar, 2002). These item sets/clusters created according to AISP provide information about the number of dimensions of the scale (Emons et al., 2012; Sijtsma & Molenaar, 2002).

The first step in the AISP is determining the *c* coefficient. The AISP selects the items that meet the $H \geq c$ criteria. If the item pair with the largest *H* coefficient in the scale meets the *c* criterion, the other items are included in the cluster according to the size of their *H* value. In this way, items are selected for the first cluster. New clusters are determined with the remaining items. This process is continued until an item that does not meet the *c* criterion remains (Sijtsma & Molenaar, 2002).

The higher the *c* coefficient used in item selection in the AISP is, the more strictly the process is carried out and the fewer item clusters are determined. The variable values of *c* in the AISP studies were compared with the selected item clusters (Chou, Lee, Liu, & Hung, 2017; Emons et al., 2012; Vaughan & Grace, 2018). A higher *c* criterion lower bound means that it is a powerful scale that orders individuals more accurately according to their total scores (Sijtsma & Molenaar, 2002). Although there is no exact ideal *c* value, researchers are recommended to examine the results starting from .30 to .55 (Emons et al., 2012).

Evidence regarding that the scale or scales obtained as a result of the AISP provide the monotonicity assumption for MHM should be presented (de Cock et al., 2011). The crit values are examined for this. Reference ranges have been defined for the interpretation of the crit values obtained. According to this definition, crit<40 is interpreted as suitable, 40≤crit<80 is interpreted as suspicious, and crit>80 is interpreted as seriously incompatible (Crişan, Tendeiro, & Meijer, 2021; Emons et al., 2012).

The AISP generally offers exploratory models to determine the dimensionality of scales. However, the AISP is a mathematically helpful tool in determining the dimensionality of scales mechanically. It should also be noted that the AISP is not a method that guarantees construct validity (Sijtsma & Molenaar, 2002, p. 87). Within this context, evaluating the AISP and EFA results together will be effective in determining the factor structures of scales.

Data Analysis

In line with the purpose of the research, in this study, the EFA and the AISP results were compared for the scores obtained from the LOT-R and the ESADE. While Jamovi 1.6.1 program was used for the EFA and the descriptive statistics, the AISP analyses were performed using Mokken Package 3.0.6 in RStudio 4.0.5 (van der Ark, 2020).

Results

The PAF Results for the Dimensionality of the LOT-R

In determining the factor structure of the LOT-R, the data were analyzed using the PAF, which is one of the EFA methods. The findings obtained are given below respectively. In order to conduct the EFA, data sets should not have missing values and outliers. These assumptions were provided in the data set. Analyses were carried out on the data of 5,859 individuals for the LOT-R. The descriptive statistics for the items of the LOT-R are given in Table 2.

Table 2. Descriptive Statistics of the LOT-R items

	Mean	Median	Standard deviation	Skewness	Kurtosis
Item 1	3.40	3	0.83	-.43	-.01
Item 3	3.34	3	0.95	-.27	-.42
Item 4	3.51	4	0.85	-.49	.01
Item 7	3.28	3	0.95	-.23	-.48
Item 9	3.52	4	0.96	-.31	-.46
Item 10	3.62	4	0.79	-.58	.53

Note. Item numbers were given according to the original scale.

It is seen that the skewness coefficients take values between -.58 and -.23, and the kurtosis coefficients take values between -.48 and .53 in Table 2. The fact that these values are between -2.00 to 2.00, and mean and median values are close to each other show that the item scores are normally distributed (George & Mallery, 2016).

Having a correlation value of .90 and above between variables is a multicollinearity problem and this is not desired in factor analysis (Hair et al., 2019). The correlation values between items were ranged between .15 and .50, and the determinant of the correlation matrix was calculated as .28. When the correlation values and the determinant value [if the determinant of correlation matrix>.00001, data shows no multicollinearity (Field, 2009)] are examined together, it can be concluded that there was no multicollinearity.

An important assumption in the EFA is the adequacy of the sample size. For this, Kaiser-

Meyer-Olkin (KMO) and Bartlett test of sphericity results were examined for the homogeneity of the variances. The findings showed that the data were suitable for the EFA (KMO=.78; χ^2= 7310.00, p<.00). The promax rotation technique was used for the PAF and the results obtained are given in Table 3.

Table 3. Factor loadings of the LOT-R items

	Factor 1	Factor 2
Item 7	0.72	
Item 9	0.71	
Item 3	0.65	
Item 4		0.72
Item 1		0.63
Item 10		0.42

Item 7, Item 9, and Item 3 are clustered under Factor 1 (Pessimism) while Item 4, Item 1, and Item 10 are clustered under Factor 2 (Optimism). This finding obtained is compatible with the factor structure of the LOT-R in the literature. The lowest factor loading is .42 and the highest one is .72. Factor loadings between .30 and .59 indicate moderately high, and the factor loadings which are .60 and above indicate high-level loading (Kline, 1994, p. 6).

Parallel analysis result was also used in determining the factors. Parallel analysis is a graphical method in which components with an eigenvalue greater than 1 are determined. In this method, which is carried out in three stages (Tabachnick & Fidell, 2014, p. 698):

(1) The data suitable for the data set whose factor structure is investigated is simulated.

(2) The PCA is repeated on all simulated data sets and eigenvalues are calculated.

(3) These eigenvalues are averaged and compared with the real data set. In the real data set only components whose eigenvalues exceed the mean eigenvalues of the simulated data are taken into account.

The graph obtained from the parallel analysis is given in Figure 2. The construct validity of the LOT-R, whose two-factor theoretical structure was known in the literature and whose factor structure was found to be compatible in accordance with the findings of this research, was provided according to both PAF and parallel analysis.

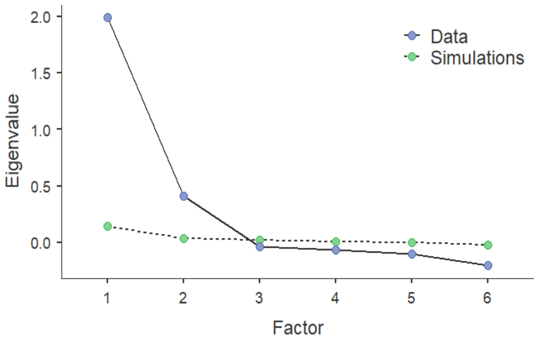

Figure 2. Result of parallel analysis of the LOT-R

The AISP Results for the Dimensionality of the LOT-R

The factor structure of the LOT-R was determined with the AISP. To determine the factors of the LOT-R according to the AISP, scales were obtained in changing values of *c* starting from .30 (the value was increased by .01) parallel to the literature. These scales and the numbers of the items in these scales are given in Table 4.

Table 4. The AISP results of the LOT-R

c	Number of scales	Items
0.30-0.32	1	3, 4, 7, 9, 10
0.33-0.37	2	3, 7, 9, 10; 1, 4
0.38-0.39	2	3, 7, 9; 1, 4, 10
0.40-0.44	2	3, 7, 9; 1, 4
0.45-0.47	1	3, 7, 9
0.48-0.55	1	7, 9
0.56	0	-

The number of items that can be scaled in the increasing values of *c* according to the Mokken model generally decreases. Two scales are proposed at .38 and .39 values of *c* in accordance with the PAF results and the theoretical structure of the LOT-R. In this study, the *c* cut-off point was accepted as .39. Two factors determined according to this cut-off point are evaluated as two separate scales in the MSA. These scales determined according to the AISP must meet monotonicity assumption of MHM. In accordance with this, the results obtained regarding the assumptions of the MHM are given in Table 5.

Table 5. Assumptions of the MHM for subscales of the LOT-R

	Scale 1 (Optimism)			Scale 2 (Pessimism)		
	Item 1	Item 4	Item 10	Item 3	Item 7	Item 9
Item *H*	.40	.44	.39	.48	.50	.53
SE_H	.01	.01	.01	.01	.01	.01
#ac	40	40	24	60	40	60
#vi	0	0	0	0	0	0
maxvi	0	0	0	0	0	0
sum	0	0	0	0	0	0
zmax	0	0	0	0	0	0
#zsig	0	0	0	0	0	0
crit	0	0	0	0	0	0

Note. SE_H: Standard error of *H*, #ac: active comparisons, #vi: violations of manifest monotonicity, maxvi: the largest violation of manifest monotonicity, sum: sum of all violations, zmax = maximum *z*-value, #zsig: significant violations.

The *H* coefficients of the items for both Scale 1 and Scale 2 are high, the SE_H values are low, and the monotonicity assumptions are met. The graphs on the monotonicity of the ICCs that belong to the items are given in Figure 3.

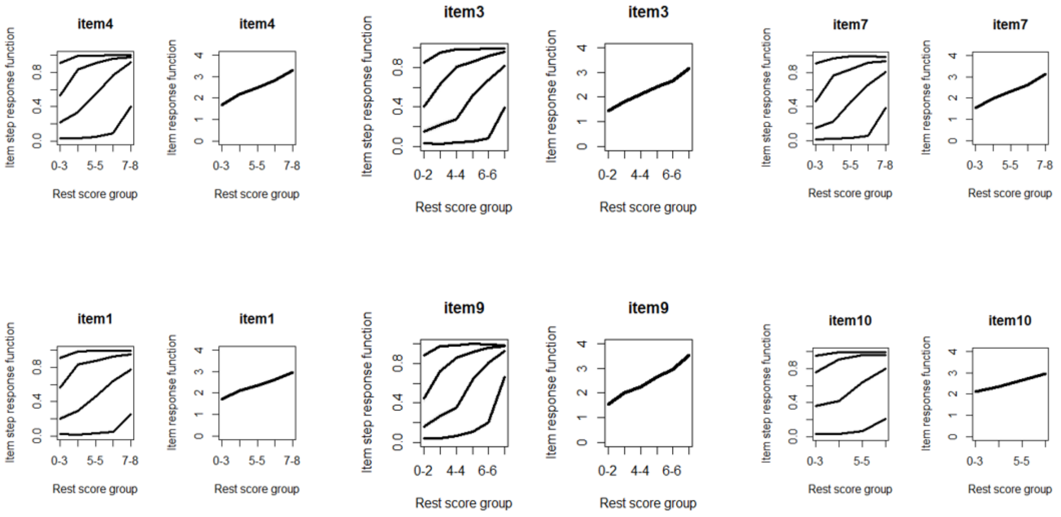

Figure 3. Monotonicity plots of the LOT-R items

Scale 1 and Scale 2 selected according to the AISP are compatible with the factor structure of the LOT-R as a result of the PAF. In addition, the *H* and SE_H values for the whole LOT-R were calculated as .36 and .01, respectively. These values show that the LOT-R is scaled according to the MHM. Table 6 below represents the results regarding the reliability coefficients.

Table 6. Reliability estimations of the LOT-R

	MS	*α*	λ_2
Optimism	0.66	0.65	0.65
Pessimism	0.74	0.72	0.73
All scale	0.75	0.74	0.75

MS: Molenaar-Sijtsma Statistic, *α*: Cronbach's Alpha, λ_2: Guttman lambda-2.

The reliability of the scores obtained from the LOT-R is relatively low but generally acceptable. The number of items influences the reliability estimates. As the number of items increases to a certain point, the estimated *Cronbach alpha* (*α*) value also increases (Cortina, 1993; Field, 2009). In accordance with this, the scores obtained from the LOT-R, which consists of a small number of items, can be evaluated as reliable.

The PCA and The PAF Results for the Dimensionality of the ESADE

The PCA and the PAF were taken into consideration in determining the factor structure of the ESADE, which was in the preliminary stage of the development phase. The PCA is not actually a factor analysis in its full sense as stated before. Furthermore, although the PCA is suggested to be used in the preliminary step process of the scale development studies (Hair et al., 2019), it is reported by many researchers as the main factor analysis alone. In the current study, both methods were used to determine the different or similar results of the PCA with the PAF.

First, 24 individuals in the dataset were determined as outliers, and the data analysis was performed on 289 individuals in total by deleting those outliers. There was no missing data in the dataset. The descriptive statistics for the items of the ESADE are given in Table 7.

Table 7. Descriptive statistics of the ESADE items

	Mean	Median	Standard deviation	Skewness	Kurtosis
Item 1	4.69	5	0.57	-1.70	1.88
Item 2	4.77	5	0.54	-2.32	4.29
Item 3	4.86	5	0.37	-2.47	5.40
Item 4	4.38	5	0.75	-0.84	-0.36
Item 5	4.49	5	0.75	-1.52	2.27
Item 6	4.64	5	0.63	-1.80	2.99
Item 7	4.91	5	0.28	-2.96	6.79
Item 8	4.71	5	0.56	-1.84	2.36
Item 9	4.18	4	0.88	-0.87	0.36
Item 10	4.46	5	0.72	-1.18	0.80
Item 11	4.75	5	0.48	-1.75	2.22
Item 12	4.64	5	0.58	-1.35	0.84
Item 13	4.24	5	0.96	-1.18	0.78
Item 14	4.81	5	0.43	-2.24	4.41
Item 15	4.73	5	0.52	-1.84	2.53
Item 16	4.81	5	0.46	-2.45	5.44
Item 17	4.38	5	0.80	-1.08	0.47
Item 18	3.73	4	1.26	-0.73	-0.46

Item scores of items 2, 3, 7, 14, and 16 are platykurtic, and the scores of the other items do not much deviate from the normal distribution. The correlation values between items ranged between .10 and .46, and the determinant of the correlation matrix was calculated as .04. When the correlation values and the determinant value are examined together, it can be concluded that there was no multicollinearity. KMO and Bartlett test of sphericity results were examined, and the data were suitable for the EFA (KMO=.80; χ^2= 928.00, p<.00).

Item 11 was excluded as a result of the PCA analysis made with varimax rotation, and the findings obtained are given in Table 8. Also, the reliability values obtained for the factor can be seen in Table 8.

Table 8. Factor loadings of the ESADE items and reliability estimations (item 11 excluded)

	Factor 1	Factor 2	Factor 3	Factor 4	Variance (%)	Cronbach's Alpha (α)
Item 12	0.68				13.6	0.65
Item 17	0.64					
Item 9	0.63					
Item 8	0.57					
Item 18	0.54					
Item 10		0.69			11.7	0.58
Item 4		0.61				
Item 5		0.60				
Item 13		0.60				
Item 3			0.63		10.9	0.51
Item 2			0.61			
Item 7			0.60			
Item 6			0.57			
Item 1			0.45			
Item 16				0.74	10.5	0.67
Item 15				0.74		
Item 14				0.62		
Total					46.8	0.76

Four factors were determined by the PCA. The lowest factor loading is .45, the highest one is .74. The reliability of the scores obtained from the ESADE is generally quite low in the factors. Figure 4 presents a scree plot graph for the PCA.

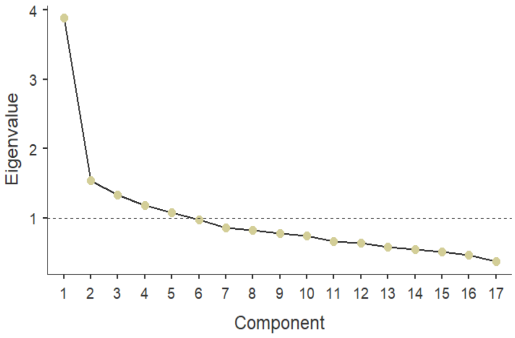

Figure 4. Scree plot of the ESADE

Figure 4 was used for deciding on the number of factors for the ESADE. When Figure 4 is examined, it can be said that there are four factors determined by the PCA for the ESADE.

The PAF analysis was performed with the oblique rotation-direct oblimin technique while determining the factor structure of the ESADE. It was confirmed that the data were suitable to be able to perform the PAF (KMO=.77; χ^2= 654.00, p<.00). According to the PAF analysis, items 1, 8, 9, and 18 should be excluded. The PAF was performed again after excluding these items. The factors and the factors loadings of the items obtained from the PAF are given in Table 9. Also, the reliability values obtained for the factor can be seen in Table 9.

Table 9. Factor loadings of the ESADE items and reliability estimations (items 1, 8, 9, and 18 excluded)

	Factor 1	Factor 2	Factor 3	Factor 4	Variance (%)	Cronbach’s Alpha (α)
Item 15	0.71					
Item 14	0.58				9.30	0.67
Item 16	0.55					
Item 10		0.69				
Item 5		0.47			8.94	0.58
Item 4		0.43				
Item 13		0.39				
Item 3			0.55			
Item 7			0.46			
Item 6			0.46		7.99	0.54
Item 2			0.41			
Item 11			0.33			
Item 17				0.68	7.79	0.62
Item 12				0.58		
Total					34.03	0.73

Four factors were determined by the PAF. The lowest factor loading is .33, and the highest one is .71. The reliability of the scores obtained from the ESADE is generally quite low in the factors. These reliability estimates are close to the reliability values obtained as a result of the PCA. Figure 5 presents both the scree plot graph and the graph obtained for the parallel analysis.

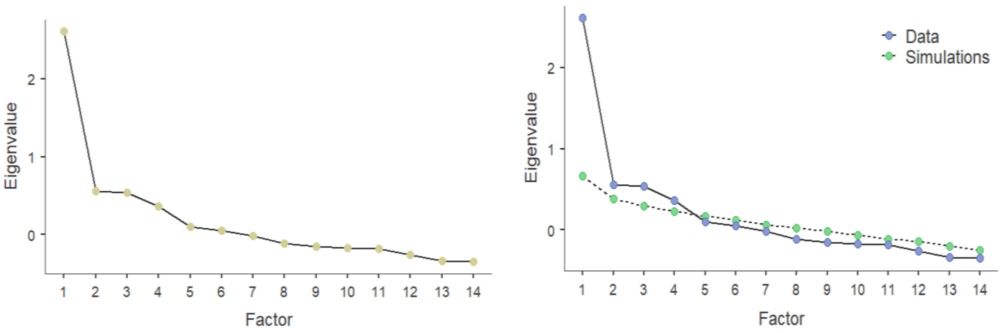

Figure 5. Scree plot and parallel analysis of the ESADE

Figure 5 was used for deciding on the number of factors for the ESADE. Figure 5 shows that there is just one factor that is greater than eigenvalue 1 in the first graphic. The number of factors may be four as a result of parallel analysis in the second graph.

The AISP Results for the Dimensionality of ESADE

The factor structure of the ESADE was determined with the AISP. To determine the factors of the ESADE according to the AISP, scales were obtained in changing values of *c* starting from .30 (the value was increased by .01) parallel to the literature. The AISP results for the ESADE are given in Table 10.

Table 10. The AISP results of the ESADE

c	Number of scales	Items
0.30	4	9, 12, 14, 15, 16, 17, 18; 6, 7; 4, 5, 10; 3, 11
0.31	4	9, 12, 14, 15, 16, 17; 6, 7; 4, 5, 10; 3, 11
0.32-0.33	5	9, 12, 14, 15, 16, 17; 6, 7; 5, 10; 13, 14; 3, 11
0.34-0.35	5	12, 15, 16, 17; 6, 7; 3, 8, 9; 5, 10; 13, 14
0.36-0.37	3	12, 15, 16, 17; 6, 7; 3, 8, 9
0.38-0.41	4	12, 17; 14, 15, 16; 6, 7; 3, 9
0.42-0.45	3	12, 17; 15, 16; 6, 7
0.46- 0.53	2	12, 17; 15, 16
0.54- 0.56	1	12, 17
0.57-0.60	0	-

The AISP results differed considerably from the PCA and the PAF results. The *c* value was accepted as .41 and this value can be considered as a high value for the AISP (de Cock et al., 2011). As a result of the AISP, a lot of items were excluded from the scale. The most important reason for this is that many items of the ESADE, which was in the preliminary stage cannot be scaled according to the MHM. The four factors determined by the AISP are considered as four separate scales in the MSA. These scales are required to meet monotonicity assumption of MHM. In accordance with this, the results obtained regarding the assumptions of the MHM are given in Table 11.

Table 11. Assumptions of the MHM for subscales of the ESADE

	Scale 1		Scale 2			Scale 3		Scale 4	
	Item 12	Item 17	Item 14	Item 15	Item 16	Item 6	Item 7	Item 3	Item 9
Item *H*	0.57	0.57	0.42	0.53	0.43	0.46	0.46	0.42	0.42
SE_H	0.07	0.07	0.06	0.07	0.07	0.09	0.09	0.08	0.08
#ac	2	2	2	2	2	0	1	3	0
#vi	0	0	0	0	0	0	0	0	0
maxvi	0	0	0	0	0	0	0	0	0
sum	0	0	0	0	0	0	0	0	0
zmax	0	0	0	0	0	0	0	0	0
#zsig	0	0	0	0	0	0	0	0	0
crit	0	0	0	0	0	0	0	0	0

Note: SE_H: Standard error of *H*, #ac: active comparisons, #vi: violations of manifest monotonicity, maxvi: the largest violation of manifest monotonicity, sum: sum of all violations, zmax = maximum *z*-value, #zsig: significant violations.

The MHM assumptions were met for Scale 1 to 4. Although the numbers of items in the scales are low, the *H* coefficients are high. The *H* and SE_H values of the scale consisting of all the selected items were calculated as .30 and .03, respectively. The scale formed from the items generally selected according to these values was scaled to the MHM. Table 12 shows the estimation results for the reliability coefficients.

Table 12. Reliability estimations of the ESADE

	MS	α	λ_2
Scale 1	0.67	0.62	0.62
Scale 2	0.68	0.67	0.67
Scale 3	0.56	0.40	0.40
Scale 4	0.51	0.33	0.33
All scale	0.73	0.70	0.72

MS: Molenaar-Sijtsma Statistic, α: Cronbach's Alpha, λ_2: Guttman lambda-2.

The reliability of the scores obtained from the ESADE is low. Among the reasons for this can be shown that the low number of items in the scales. Also, it must be underlined that strong construct validity of the scales are not provided.

Discussion

The aim of the current study is to compare the EFA and the AISP for determining factor structure or dimensionality. In line with the specified purpose, the LOT-R data and the ESADE were evaluated within the scope of the research.

The comparison of the results of the EFA and the AISP first started with the LOT-R whose validity was known. The results of the PAF, and the results of the AISP which was carried out by determining a high cut-off point such as .39 showed that the factor structure of the LOT-R was in accordance with the literature. The PAF and the AISP results, which are compatible with each other, can be considered as strong evidence for the construct validity. There are studies in the literature with similar results on different scales (de Cock et al., 2011; Shenkin et al., 2014). However, it should not be forgotten that CFA should be performed in order to investigate the construct validity of scales whose factor structure is known beforehand, in new samples. Additionally, test-try out 3 step is given in Figure 1 should be followed in order to provide proof of construct validity in scale development studies. In this step, the structure of the scale is confirmed on a different sample by performing the CFA.

When the steps in Figure 1 are examined, it will be seen that EFA should be done at the stage of scale purification in the scale development studies. This study focuses on the use of the AISP and the EFA together in the research of factor structure of scales in scale development research. In other words, the scales developed according to a theory and whose factor structure is desired to be explored are the focus of the current study. The similarity of the AISP and PAF results for LOT-R suggests that the use of both methods in scale development studies will benefit researchers for searching construct validity.

One of the points targeted of the current study is to compare the EFA and the AISP results for a new scale whose factor structure is unknown and is in the development process and to determine the gain that the AISP will provide to researchers. For this purpose, scores obtained from the ESADE data which was in the preliminary stage at the development process were analyzed.

The factor structure of the ESADE was determined with the PCA and the PAF within the scope of the EFA. A four-factor structure was revealed according to both methods, and two of these factors included the same items in both methods. However, different results were reached in both methods in general. If the PCA and the PAF results are evaluated individually, it can be thought that the construct validity of the ESADE which is in the preliminary stage is provided at the first glance. However, the differentiation of the items excluded from the ESADE according to both methods and the inclusion of different items under the factors determined according to both methods point to an important issue. Although the main factor analysis is the PAF, it is also stated that the results obtained from the PCA and the PAF do not differ in many experimental studies (Hair et al., 2019, p.140). Accordingly, it can be concluded that there are problems with the ESADE. Nonetheless, it should also be underlined that the current situation is a result of evaluating the results of the PCA and the PAF together. The PCA results alone may be misleading or insufficient for the scales that are newly developed and applied on a single sample.

In addition to the EFA, the AISP results were also used to determine the factor structure of the ESADE. Changing *c* values were considered for the AISP and the cut-off point was determined as .41. As stated in the literature, the choice of the *c* value depends on the researcher, and the higher the *c* value is, the stronger MSs are obtained (Emons et al., 2012; Sijtsma & Molenaar, 2002; Wismeijer et al., 2008). Accordingly, four MSs were determined, and few items were included in these scales. The most important reason for this is that many items in the ESADE could not be scaled according to the MHM.

According to the results obtained from the AISP for the ESADE, it is necessary to seriously work on it. Considering the steps in Figure 1, it is also necessary to return to test construction and item refinement stages and to perform the pilot survey, test tryout-1 and test tryout-2 stages again. In addition, as can be seen in Table 1, the ESADE was applied to a homogeneous group. It should be applied to heterogeneous and larger groups in the new application.

When the PCA, the PAF, and the AISP results for the ESADE were compared, four factors were formed in all three methods and there was only one common factor (including items 14, 15, and 16). Moreover, the PAF and the AISP results were found to be more similar. There are studies in the literature reaching similar results. For example, the factor structure of the Self-Concealment Scale was determined with the PCA and the MSA in Wismeijer et al.'s (2008) studies. The findings obtained showed that the results obtained from the PCA and the MSA methods were different as in this study. Gudergan et al. (2004), compared the factor analysis, the Rasch analysis, and the Mokken analysis results in scale development and stated that different structures emerged as a result of these three methods. They also stated that studies evaluating only factor analysis results are questionable. Chen, Watson, and Hilton (2016) determined the factor structure of the Mentors' Behavior scale with the MSA and the PAF. They stated that the factor structures determined according to the MSA and the PAF results were similar to each other, however, some items taken into account in the PAF were excluded in the MSA like in this study.

Additionally, the reliability of the scores obtained from the LOT-R and the ESADE was also estimated in this research. The reliability values obtained from the LOT-R were seen to be parallel to the literature (Gustems-Carnicer, Calderón, & Santacana, 2017; Scheier et al., 1994). The reliability values obtained from the ESADE were also found to be low. The effect of the number of the items and the fact that the scale has problems regarding the construct validity can be shown among the reasons for this situation.

When the results obtained from the LOT-R and the ESADE are evaluated together, it can be concluded that the PAF and the AISP results should be similar in the scales whose construct validity is provided. It can be expected that there should not be serious differences between the PAF and the AISP results in the scales with the construct validity.

In scale development studies, results obtained by a single statistical method, or a single study group-sample are shared. The scale should be re-applied in a different group and taken into the EFA process. In line with the findings of this research, it can be suggested that the PAF method and the direct oblimin rotation should be preferred in scale development studies, as stated in the literature (Field, 2009; Hair et al., 2019; Tabachnick & Fidell, 2014). In the context of this research, one of the most important suggestions is that the results obtained from a single sample and only factor analysis will not be sufficient to prove the construct validity of the scale. Also, the AISP under the MSA should take into consideration. It should also be noted that the PCA may give misleading results regarding the construct validity due to its mathematical infrastructure. It may also be recommended to examine the PAF or other factor extraction methods in scale development studies.

One of the most important issues in the AISP is the *c* cut-off point. The cut-off points for the LOT-R and the ESADE are high in this study. Therefore, it was provided more reliable results. However, when higher values are selected, fewer items will be selected for the scale as a matter of mathematical operation. Researchers should be careful at this point.

It would be useful to examine the AISP results together with the classical factor analysis methods in determining the construct validity of the scales, especially in the scale development studies. More attention should be paid to the construct validity of the scales when the AISP results that are different from the ones of the factor analysis methods are obtained.

The AISP is a method held under the MSA. Mokken Models can be used to scale short tests applied to small samples and high *H* values indicate that individuals are ordered reliably according to their total scores. Therefore, it is important to scale data according to the MHM in scale development studies. It is recommended that researchers take into account the AISP results with appropriate factor analysis methods in their scale development studies.

Acknowledgments

In this paper we make use of data of the LISS (Longitudinal Internet Studies for the Social Sciences) panel administered by CentERdata (Tilburg University, The Netherlands). The author sincerely thanks the LISS panel for sharing data sets. The author also thanks the participants who voluntarily participated in the scale development study.

References

Aljubaily, H. Y. (2010). *Measuring university students' perceptions of characteristics of ideal university instructor in Saudi Arabia and the United States: An application of nonparametric item response theory study* (Doctoral Dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 3434898).

Anastasia, A. (1988). *Psychological testing* (6th ed.). New York: Macmillan Publishing.

Bagnasco, A., Watson, R., Zanini, M., Rosa, F., Rocco, G., & Sasso, L. (2015). Preliminary testing using Mokken scaling of an Italian translation of the Edinburgh Feeding Evaluation in Dementia (EdFED-I) scale. *Applied Nursing Research*, *28*(4), 391-396. doi: 10.1016/j.apnr.2015.02.003

Bech, P., Carrozzino, D., Austin, S. F., Møller, S. B., & Vassend, O. (2016). Measuring euthymia within the Neuroticism Scale from the NEO Personality Inventory: a Mokken analysis of the Norwegian general population study for scalability. *Journal of Affective Disorders*, *193*, 99-102. doi: 10.1016/j.jad.2015.12.039

Chen, Y., Watson, R., & Hilton, A. (2016). An exploration of the structure of mentors' behavior in nursing education using exploratory factor analysis and Mokken scale analysis. *Nurse Education Today*, *40*, 161-167. doi: 10.1016/j.nedt.2016.03.001

Chou, Y. H., Lee, C. P., Liu, C. Y., & Hung, C. I. (2017). Construct validity of the depression and somatic symptoms scale: evaluation by Mokken scale analysis. *Neuropsychiatric Disease and Treatment, 13*, 205-211. doi: 10.2147/NDT.S11882

Cohen, R. J., & Swerdlik, M. (2009). *Psychological assessment: an introduction to tests and measurements* (7th ed.). McGraw-Hill Primis.

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology*, *78*(1), 98-104. doi: 10.1037/0021-9010.78.1.98

Crișan, D. R., Tendeiro, J. N., & Meijer, R. R. (2021). The Crit coefficient in Mokken scale analysis: a simulation study and an application in quality-of-life research. *Quality of Life Research*, 1-11. doi: 10.1007/s11136-021-02924-z

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Cengage Learning: USA.

De Ayala, R. J. (2009). *The theory and practice of item response theory*. Guilford Publications: New York.

de Cock, E. S., Emons, W. H., Nefs, G., Pop, V. J., & Pouwer, F. (2011). Dimensionality and scale properties of the Edinburgh Depression Scale (EDS) in patients with type 2 diabetes mellitus: the DiaDDzoB study. *BMC Psychiatry*, *11*(1), 141. doi: 10.1186/1471-244X-11-141

DeVellis, R. F. (2017). *Scale development: theory and applications* (4th ed.). Sage publications: Los Angeles.

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

Emons, W. H., Sijtsma, K., & Pedersen, S. S. (2012). Dimensionality of the Hospital Anxiety and Depression Scale (HADS) in cardiac patients: comparison of Mokken scale analysis and factor analysis. *Assessment*, *19*(3), 337-353. doi: 10.1177/1073191110384951

Field, A. (2009). *Discovering statistics using IBM SPSS statistics*. (4th ed.). Sage Publication: London.

George, D., & Mallery, P. (2016). *IBM SPSS statistics 23 step by step: A simple guide and reference*. Routledge: NY and OX.

Gudergan, S., Mathies, C., Kyngdon, A., & Kozicki, S. (2004, December). *Negotiation style measurement scale development and testing*. Paper presented at the Australian and New Zealand Marketing Academy Conference, Wellington. Abstract retrieved from https://opus.lib.uts.edu.au/handle/10453/3133

Gustems-Carnicer, J., Calderón, C., & Santacana, M. F. (2017). Psychometric properties of the Life Orientation Test (LOT-R) and its relationship with psychological well-being and academic progress in college students. *Revista Latinoamericana de Psicología, 49*(1), 19-27. doi: 10.1016/j.rlp.2016.05.001

Hair, J. F., Black, C. W., Babin, B. J., & Anderson, R. E. (2019). *Multivariate data analysis* (8th ed.). Annabel Ainscow: United Kingdom.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: principles and application.* Boston: Kluwer Academic Publishers Group.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. USA: Sage Publications.

Howard, M. C. (2016). A review of exploratory factor analysis decisions and overview of current practices: what we are doing and how can we improve? *International Journal of Human-Computer Interaction, 32*(1), 51-62. doi: 10.1080/10447318.2015.1087664

Kline, P. (1994). *An easy guide to factor analysis*. Routledge: USA.

Lawshe, C. H. (1975). A quantitative approach to content validity. *Personnel Psychology*, *28*(4), 563-575.

Lee, C. P., Chen, Y., Jiang, K. H., Chu, C. L., Chiu, Y. W., Chen, J. L., & Chen, C. Y. (2016). Development of a short version of the Aging Males' Symptoms scale: Mokken scaling analysis and Rasch analysis. *The Aging Male*, *19*(2), 117-123. doi: 10.3109/13685538.2016.1157861

Lee, C. P., Fu, T. S., Liu, C. Y., & Hung, C. I. (2017). Psychometric evaluation of the Oswestry Disability Index in patients with chronic low back pain: factor and Mokken analyses. *Health and Quality of Life Outcomes*, *15*(1), 1-7. doi: 10.1186/s12955-017-0768-8

Matsunaga, M. (2010). How to factor-analyze your data right: do's, don'ts, and how-to's. *International Journal of Psychological Research, 3*(1), 97-110. doi: 10.21500/20112084.854

Meijer, R. R., & Baneke, J. J. (2004). Analyzing psychopathology items: a case for nonparametric item response theory modeling. *Psychological Methods, 9*(3), 354-368. doi: 10.1037/1082-989X.9.3.354

Mokken, R. J. (1997). Nonparametric models for dichotomous responses. In W. J. van der Linden, & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 351-368). New York: Springer-Verlag.

Mooij, T. (2012). A Mokken scale to assess secondary pupils' experience of violence in terms of severity. *Journal of Psychoeducational Assessment*, *30*(5), 496-508. doi: 10.1177/0734282912439387

Scheier, M. F., Carver, C. S., & Bridges, M. W. (1994). Distinguishing optimism from neuroticism (and trait anxiety, self-mastery, and self-esteem): a reevaluation of the Life Orientation Test. *Journal of Personality and Social Psychology, 67*(6), 1063. doi: 10.1037/0022-3514.67.6.1063

Shenkin, S.D., Watson, R., Laidlaw, K., Starr, J. M., & Deary, I. J. (2014). The attitudes to ageing questionnaire: Mokken scaling analysis. *PLOS ONE 9*(9): e108766. doi: 10.1371/journal.pone.0099100

Sijtsma, K., & Molenaar, I. W. (2002). *Introduction to nonparametric item response theory*. Sage Publications.

Ski, C. F., Thompson, D. R., Hare, D. L., Stewart, A. G., & Watson, R. (2012). Cardiac Depression Scale: Mokken scaling in heart failure patients. *Health and Quality of Life Outcomes*, *10*(141). doi: 10.1186/1477-7525-10-141

Stewart, M. E., Allison, C., Baron-Cohen, S., & Watson, R. (2015). Investigating the structure of the autism-spectrum quotient using Mokken scaling. *Psychological Assessment*, *27*(2), 596-604. doi: 10.1037/pas0000058

Stout, W. (2001). Nonparametric item response theory: a maturing and applicable measurement modeling approach. *Applied Psychological Measurement*, *25*(3), 300-306. doi: 10.1177/01466210122032109

Tabachnick, B. G., & Fidell, L. S. (2014). *Using multivariate statistics* (6th ed.). Pearson: United States of America.

The jamovi project (2021). *jamovi* (Version 1.6) [Computer Software]. Retrieved from https://www.jamovi.org

Van der Ark, L. A. (2020). Package Mokken. https://cran.rproject.org/web/packages/mokken/mokken.pdf

VandenBos, G. R. (2015). *APA dictionary of psychology* (2nd ed.). Washington DC: American Psychiatric Pub. doi: 10.1037/14646-000

Vaughan, B., & Grace, S. (2018). A Mokken scale analysis of the peer physical examination questionnaire. *Chiropractic & Manual Therapies*, *26*(1), 6. doi: 10.1186/s12998-018-0176-0

Watson, R., van der Ark, L. A., Lin, L. C., Fieo, R., Deary, I. J., & Meijer, R. R. (2012). Item response theory: how Mokken scaling can be used in clinical practice. *Journal of Clinical Nursing*, *21*(19pt20), 2736-2746. doi: 10.1111/j.1365-2702.2011.03893.x

Wismeijer, A. A., Sijtsma, K., van Assen, M. A., & Vingerhoets, A. J. (2008). A comparative study of the dimensionality of the self-concealment scale using principal components analysis and Mokken scale analysis. *Journal of Personality Assessment*, *90*(4), 323-334. doi: 10.1080/00223890802107875

Yoon, S., Shaffer, J. A., & Bakken, S. (2015). Refining a self-assessment of informatics competency scale using Mokken scaling analysis. *Journal of Interprofessional Care, 29*(6), 579-586. doi: 10.3109/13561820.2015.1049340

Appendix

Example Items for the ESADE

Item 1: Lectures should be taught in accordance with the plan prepared beforehand.
Item 11: Students' opinions about the lecture should be taken into consideration.
Item 14: Student achievement should be evaluated with clear, understandable criteria.
Item 15: Criteria for evaluating achievement should be shared with students on the online platform.
Item 16: Students should be given feedback on the exam or homework (not just a score).
Item 17: Assessments should be shared online by using statistical information.