LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN
INSTITUT FÜR STATISTIK

Stella Bollmann, Moritz Heene, Helmut Küchenhoff, Markus Bühner

What can the Real World do for simulation studies? A comparison of exploratory methods

Technical Report Number 181, 2015
Department of Statistics
University of Munich

http://www.stat.uni-muenchen.de

Abstract

For simulation studies on the exploratory factor analysis (EFA), usually rather simple population models are used without model errors. In the present study, real data characteristics are used for Monte Carlo simulation studies. Real large data sets are examined and the results of EFA on them are taken as the population models. First we apply a resampling technique on these data sets with sub samples of different sizes. Then, a Monte Carlo study is conducted based on the parameters of the population model and with some variations of them. Two data sets are analyzed as an illustration. Results suggest that outcomes of simulation studies are always highly influenced by particular specification of the model and its violations. Once small residual correlations appeared in the data for example, the ranking of our methods changed completely. The analysis of real data set characteristics is therefore important to understand the performance of different methods.

Within the framework of psychological survey data, latent structure detection is an important aim. Items have to be assembled into groups in a way so that items within one group are as similar as possible to enable sufficiently accurate diagnoses. Generally, exploratory factor analysis (EFA) represents a widely accepted standard for exploration of test structures and detection of items that have a similar content. The first step in this structure detection process is to determine the appropriate number of true factors to retain (*dimensionality assessment*). The second step includes identifying which variable belongs to which factor (*assignment of variables*), what is usually done by assigning each variable to the factor on which it has the highest loading.

For the evaluation of the performance of EFA methods in both steps, previous studies have used real data and artificial data. Artificial data from Monte Carlo studies usually are generated by specifying the underlying population model correlation matrix that serves as covariance matrix of the distribution from which random data are drawn (Bacon, 2001; Beauducel, 2001; Crawford et al., 2010; D. A. Jackson, 1993; Ruscio & Roche, 2012; Smith, 1996; Velicer, Eaton, & Fava, 2000; Velicer & Fava, 1998; Zwick & Velicer, 1986). When using real data though, the true underlying model is unknown and cannot be tested. One possible way is to compare the results of sub samples to the corresponding result of the full-sample, the so-called resampling that has been used for evaluation of EFA in order to analyze sample sizes effects (Arrindell & Ende, 1985; Barrett & Kline, 1981; MacCallum, Widaman, Preacher, & Hong, 2001).

EFA simulation studies are based on more or less artificial conditions and most of them are built on simple population factor models. Typically, only few of the possible model parameters are integrated into the simulations in the first place while the others are set to zero. Model violations are almost never included. Note that the simulation of model violations is in fact much more common in Monte Carlo simulation studies on confirmatory factor analysis (CFA). Here, model error may be specified in the model in the form of residual correlations

(e.g. Hu & Bentler, 1999). But almost all simulation studies of EFA are based on the ideal case of exact fit of the common factor model in the population. This is a huge limitation, given that performance of EFA may change substantially in presence of model error (MacCallum et al., 2001). Generally speaking, studies that are based on simulated data lack a connection to real world data conditions. For example, in studies that investigated the influence of different sizes of loadings or communalities, cross-loadings are typically set to zero (Gerbing & Hamilton, 1996; Mundfrom, Shaw, & Ke, 2005; Velicer & Fava, 1998). If cross-loadings are integrated (Bacon, 2001; Sass & Schmitt, 2010; Sass, 2010; Zwick & Velicer, 1986), mostly either a two-factor model is investigated or only one or two high cross-loadings per variable are specified while the others remain restricted to zero. Real life psychometric data however often do consist of multiple factors and small but multiple cross-loadings of all variables on all factors (e.g. Church & Burke, P. J., 1994; Haynes, Miles, & Clements, 2000). In studies that examined factor inter-correlations on the other hand (Bacon, 2001; Crawford et al., 2010; Sass & Schmitt, 2010; Sass, 2010), either two-factor models were investigated or all factor inter-correlations were set to be equal. Data from the social sciences, however, often exhibit a complex model structure with many cross-loadings and factor inter-correlations of different sizes. And in addition, real data always exhibit more or less high model violations, e.g. residual correlations, because even complex models are never exactly true even in the population (Cudeck & Henly, 1991; MacCallum & Tucker, 1991). Model violations are a serious issue, which has to be considered since they might affect recovery of the population correlation matrix substantially (MacCallum et al., 2001). Moreover, it could have been shown that the disregard of existing model violations can lead to biased estimations of model parameters (Kaplan, 1988; Yuan, Marshall, & Bentler, 2003). Therefore, conclusions from Monte Carlo simulations that do not apply realistic factor models along with realistic model violations cannot be generalized to real data. This is in line with several studies showing that traditional Monte Carlo studies and rules of thumb commonly

lack validity (Fan & Sivo, 2005; D. L. Jackson, 2007; MacCallum & Tucker, 1991; MacCallum et al., 2001). They showed that simple rules of thumb regarding for example the minimum sample size for EFA are misleading. Thus, to draw valid conclusions from simulation studies, characteristics of the model and its violations have to be considered. These characteristics of the model and its violations in the shape they occur in a particular real data set shall in the following be called *data set characteristics*.

In studies where real data, exhibiting model violations, are analyzed, data set characteristics are not known and not manipulated. The advantage of the use of real data is that particular real data conditions can be described. These results can though hardly be transferred to other data sets; in so far as it is unknown which kind of data set might display the same behavior and which not. For a sufficient examination of EFA thus, we see it as reasonable approach to explore not only the performance of different methods on real data sets but also the data set characteristics of the respective data set. Not only all parameters of the model that best fits the data shall be taken into consideration but also their related occurred model violations. These parameters can be modified in a controlled way in a subsequent Monte Carlo study in order to detect their particular influence. In the present study, we explore different data sets with specific characteristics and aim to identify those characteristics that make the difference between the performance of EFA in simulation studies and real world data. We take parameter estimates from real data sets and integrate them into simulation studies to compare their impact. Only after taking these data characteristics into account, the influence of sample size on the performance of different methods will be investigated.

In a first step, we use real data sets with several thousand cases for the definition of a (finite) population. The entire data set is analyzed and the results of the assessed methods are taken as the population model. Next, samples are drawn with replacement from this data set to examine how well the detected structure can be reproduced in each of these sub-samples. This

analysis will be referred to as *Real World simulation*. The main goal of this process is to get a deeper insight into the performance of EFA under realistic data conditions in comparison to *traditional simulation* conditions. In the second step, the results of the *Real World simulation* set are compared to results from the commonly used simulation study design, using the same hypothetical factor model. In this way, we will test the hypothesis that most methods show different levels of performance in *traditional simulation* studies than they do in the *Real World*. The idea is to use sample based estimates of model parameters from the real data as a basis for the simulation of the artificial data in the simulation study. In doing so, we search for such data set characteristics that are necessary to conduct *traditional simulation* studies in order to resemble *Real World* conditions as good as possible. Another advantage of the subsequent Monte Carlo simulation is that although it is compared to the same model as in the *Real World simulation*, the simulated model is free of the effects of sample size. If thus the *Real World simulation* shows that a method is sensitive for sample size and therefore leads to different results in the population data set than in the sub samples, this can be verified in the subsequent Monte Carlo study. We use two different psychometric data sets and examine five different EFA methods for assessment of dimensionality and one EFA method for the *assignment of variables*.

Dimensionality assessment

EFA is based on the idea that latent variables are the cause of correlations between test items. Thus, the aim of EFA is the reproduction of the correlation matrix:

$$R = \Lambda\Phi\Lambda' + U^2$$

whereby R is the p x p correlation matrix of the p observed variables, U^2 is the diagonal p x p matrix of the unique variances, Λ is a p x k matrix of the factor loadings on the k < p factors, and Φ stands for the k x k matrix of the factor inter-correlations.

Different methods have been suggested to determine the number of factors in EFA. According to previous research, *Parallel Analysis* (PA, Horn, 1965) seems to be the most accurate procedure to determine the right number of factors to retain (Fabrigar, Wegener, MacCallum, & Strahan, 1999; Humphreys & Montanelli Jr., 1975; Lance, Butts, & Michels, 2006; Patil, Singh, Mishra, & Todd Donavan, 2008). PA is based on a simulated data set, containing random numbers for the same number of variables with the same sample size. This variables thus do not share a common factor. A *Principal Component Analysis* (PCA) or a *Principal Axis Factor Analysis* (PAF) for each simulated data set is conducted, and obtained Eigenvalues are recorded. Consequently, one obtains a sample distribution of eigenvalues. Finally, those factors with Eigenvalues exceeding the mean of the corresponding simulated factors, or its 95th-percentile, are retained. Given that in previous research the 95th-percentile criterion was found to be preferable over the mean criterion and its results mitigate the classic PA's slight tendency to recommend too many factors (Crawford et al., 2010; Glorfeld, 1995; Hayton, Allen, & Scarpello, 2004), in this study we decided to use the 95th-percentile criterion.

According to Zwick & Velicer (1982, 1986), the *Minimum Average Partial rule* (MAP rule, Velicer, 1976) is the second most accurate method for *dimensionality assessment*. It is based on a preliminary PCA in which only the first principal component is retained. This component is then partialed out of the correlations between variables. The off-diagonals of the resulting partial correlation matrix are used to compute the average squared coefficient. In the next step, the first two principal components are partialed out of the original correlation matrix and the average squared coefficient is computed again. These steps are repeated with one additional component per step until the average squared coefficient no longer decreases. At this point, when the average squared coefficient is minimal, the variance in the partial correlation matrix is interpreted as to no longer represent systematic variance. The number of

components that are retained at that point determines the number of components that has to be extracted.

The number of factors to be extracted can also be estimated using fit indices based on the maximum likelihood (ML) approach, such as the *Akaike information criterion* (AIC; Akaike, 1974) and the *Bayesian information criterion* (BIC; Schwarz, 1978). For both of these indexes, first the value of the likelihood function L for the estimated model is maximized. This value is then transformed in the following manners:

$$AIC = -2\ln(L) + 2p$$

$$BIC = -2\ln(L) + \ln(n) \times p.$$

with p being the number of parameters to be estimated and n the sample size. Lower values indicate better fit and, therefore, the model with the lowest AIC- or BIC-value is the optimal one.

There has been a large debate on the question of how sample size influences the *dimensionality assessment* in factor analysis (FA). The results are far from being clear and different researchers have come to different conclusions. When interpreting their results though, two aspects have to be considered. First, with increasing sample size the estimation error decreases and therefore parameter estimations become more stable across samples. Second, some methods select more complex models (i.e. more factors) the higher the sample size gets. The first influence has been examined in several previous simulation studies (Beauducel, 2001; Browne, 1968; Crawford et al., 2010; Guadagnoli & Velicer, 1988; Velicer & Fava, 1998). However, the more other parameters are included into the simulation, e.g. communalities and number of variables, the more the effect of more accurate estimation of number of factors with increasing sample size seems to diminish (Mundfrom et al., 2005; Pennell, 1968; Velicer & Fava, 1998). Interestingly, one study that analyzed real data could not find any relation between factor stability and sample size at all (Arrindell & Ende, 1985). The second effect of a systematically increasing number of factors with increasing sample

size has been examined by McDonald (1989). He found a strong dependency of model complexity and sample size when using AIC. Cudeck and Henley (1991) in their study though stated that this dependency is not necessarily undesired saying that "It may be better to use a simple model in a small sample rather than one that perhaps is more realistically complex but that cannot be accurately estimated." (Cudeck & Henly, 1991). It is important to consider that the target model of AIC changes with sample size while BIC assumes that there is a true model, independent of n (Burnham & Anderson, 2004; Kuha, 2004). AIC values parsimony less than BIC, therefore it can be expected that AIC favors more complex models than BIC, especially when sample size is growing large (Vrieze, 2012). Other simulation studies on the performance of AIC (Homburg, 1991; Lubke & Neale, 2006) did not find sample size to have such a big impact on model complexity. They rather found AIC to detect the right umber of factors more often with increasing sample size. These heterogeneous results suggest that when analyzing the influence of sample size on dimensionality assessment, one has to consider the characteristics of the data set, because for different data set characteristics different sample size effects are observed.

Assignment of variables

The first step necessary for assignment of variables to factors with EFA is the estimation of a loading matrix. Therefore, it is important that the specific method used yields one loading matrix that best reflects the underlying population loading pattern. Most of the existing studies compared the height of all loadings between the population loading matrix and the reproduced loading matrix (Arrindell & Ende, 1985; Guadagnoli & Velicer, 1988; Mundfrom et al., 2005; Sass & Schmitt, 2010; Sass, 2010; Schmitt & Sass, 2011; Velicer & Fava, 1998). To our knowledge there is only one study to date by Gerbing & Hamilton (1996) that evaluated not only the loading matrix, but also the *assignment of variables* to factors by assigning each indicator to the factor for which it showed the highest loading. They did not, however, include cross-loadings in their simulations. To assess the correct assignment

of facets to factors, we computed one EFA with PAF per data set. Variables were assigned to the factor for which they had the highest loading.

To summarize, this study aims to analyze the performance of EFA in *dimensionality assessment* and *assignment of variables* using *Real World simulation* and *traditional simulation* studies. We expect to find differences between the two simulations and hope to be able to establish useful guidelines for making simulation studies more comparable to Real World situations.

Method

We implemented a new approach, the *Real World simulation* for the exploration of the performance of EFA in the *dimensionality assessment* and *assignment of variables*. As population data sets we first used the norm data set of the German version of the revised NEO personality inventory (NEO-PI-R, Ostendorf & Angleitner, 2004) and then the second edition of the multidimensional intelligence structure test IST-2000-R by Amthauer, Brocke, Liepmann and Beauducel (Amthauer, Brocke, Liepmann, & Beauducel, 2007)[1]. Both are broadly used diagnostic tools in psychological practice. We specified a population model with each method on the population data and then tried to replicate the respective model in smaller samples drawn out of this population data. In a second step, these results were compared to the results from traditionally simulated conditions where estimators from the population data set were used to simulate data from a distribution.

Real World simulation.

The NEO-PI-R data set. The NEO-PI-R is a widely used personality inventory measuring personality in five major domains: neuroticism, extraversion, openness to

[1] We would like to express our thanks to Fritz Ostendorf and André Beauducel for providing the data sets for this study.

experience, agreeableness, and conscientiousness. Each domain scale is divided into six facets and eight items operationalize each facet. Thus, the questionnaire consists of 240 items. For this study, a self-report form was used in which participants provided self-reports on typical behaviors or reactions on a five point Likert scale, ranging from 0 = *strongly disagree* to 4= *strongly agree*. Validity and reliability for all questionnaires was shown by Ostendorf and Angleitner (2004).

As indicated in the manual of the NEO-PI-R, the correlation matrix of facets shows moderate to high inter-correlations within factors and low inter-correlations between factors (for more details see Ostendorf & Angleitner, 2004). The mean of factor inter-correlations is -.05 and the mean of absolute values of factor inter-correlations is .19 ranging from .02 to .52. Main factor loadings range from .37 to .88 (see Table 1). Cross-loadings of the data set are more or less normally distributed with a mean of .01, ranging from -.44 to .47. All in all the NEO-PI-R data set exhibits relatively high cross loadings and low factor inter-correlations.

The NEO-PI-R norm data set consists of 11,724 participants. The mean age of the sample is 29.92, ranging from 16 to 91 with 36% males and 64% females. Sum scores on facets were calculated for each subject.

The IST-2000-R data set. The basic module of the IST-2000-R measures intelligence in three major domains: verbal intelligence, numerical intelligence and spatial intelligence each of which is divided into three sub-tests. The test compromises a total of 180 questions, which can only be answered true or false.

Main factor loadings of the sub-test sum scores range from .47 to .83 (see Table 2). Validity and reliability for all questionnaires was shown by Amthauer et al. (2007). Cross-loadings range from -.15 to .20 with a mean of .01. The three factor inter-correlations have the values .66, .49 and .43. To sum it up, cross-loadings are lower than in the NEO-PI-R

Table 1
Loading Matrix of the Population Data Set NEO-PI-R

Facet	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5
N1	.88		.12		.10
N2	.76	-.44			
N3	.80			-.15	.12
N4	.65	.13		-.17	
N5	.37	-.18	-.23	.41	.12
N6	.77		-.16		
E1		.47	.10	.81	
E2		.15		.76	-.16
E3	-.27	-.37	.21	.37	
E4		-.22	.36	.50	
E5		-.23	-.17	.37	
E6	-.16	.14		.64	.18
O1		-.10	-.21		.62
O2	.15	.13			.76
O3	.26		.11	.21	.67
O4	-.24		-.16	.16	.38
O5	-.19	-.15	.12	-.25	.65
O6	-.16		-.18		.46
A1	-.23	.54		.38	
A2		.60			
A3		.73	.15	.44	
A4	-.14	.73			
A5	.21	.55			-.12
A6	.21	.52		.26	.25
C1	-.32		.55	.12	.14
C2	.12		.72		
C3		.22	.73		
C4	.10	-.19	.74		.18
C5	-.13		.76		
C6		.13	.48	-.40	

Note. N=Neuroticism; E=Extraversion; O=Openness to experiences; A=Agreeableness; C=Conscientiousness. Promax roation; ML estimation. Loadings below .10 are suppressed.

Table 2
Loading Matrix of the Population Data Set IST-2000-R

Sub Test	Factor 1	Factor 2	Factor 3
V1		.74	-.11
V2		.67	.18
V3	.18	.53	
N1	.47	.12	.18
N2	.83		
N3	.74		
S1			.64
S2			.49
S3	-.11		.51

Note. V=verbal intelligence; N=numerical intelligence; S=spacial intelligence. Promax roation; ML estimation. Loadings below .10 are suppressed.

data set and factor inter-correlations are much higher. The norm data set consists of 1,352 observations. The mean age is 19.09 ranging from 16 to 25 with 44% females and 56% males.

In both data sets the sub-facets (NEO-PI-R), and sub-tests (IST-2000-R) respectively, were the basis of our calculations. We regarded them as variables of our data set and assembled them into overlying factors. This was done in order to avoid the problem of categorical, and binary data that will be addressed in further research.

For determining the number of factors we used the MAP rule, PA-PAF, PA-PCA, AIC, and BIC. AIC and BIC were based on a PAF with ML estimation and Promax rotation. For the *assignment of variables,* also a PAF with ML estimation and Promax rotation was conducted. Van der Linden et al. (2012) showed in a study on factor inter-correlation of different personality inventories that factors are typically correlated. Their inter-correlations range from .52 to .67 (in absolute values) and the average inter-correlation is .60. For this reason, we chose oblique Promax rotation for our study. Facets were assigned to the factor for which they had the highest absolute loading.

Samples. Samples of sizes 100; 200; 300; 400; 500 and 1,000 were drawn randomly with replacement from the complete data set, with 1,000 replications each. Sampling with

replacement was chosen because otherwise a larger sample would have automatically meant a closer similarity to the population data set and therefore higher success rates (proportion of identical number of factors as in the population data set). By choosing sampling with replacement, we intended to create more comparable conditions between different sample sizes. For dimensionality assessment for each sample the MAP rule, PA-PAF and PA-PCA as well as AIC and BIC were computed. Subsequently, we determined the percentage of correct number of dimensions (success rate), i.e. identical to the number found in the population data and the mean of the suggested number of dimensions.

Thus, we had 6 x 5 different conditions in each data set for determining the number of factors as we compared the five different methods with six different sample sizes each.

For the report of similarities of factor solutions in *assignment of variables* we set the number of factors to the theoretically assumed number: Five factors for the NEO-PI-R and three factors for the IST-2000-R. Similarity was then determined using the Rand Index (Rand, 1971), calculated by counting the number of correctly classified pairs of elements. Thus, the Rand Index is defined by:

$$R(C, C') = \frac{2(n_{11} + n_{00})}{n(n-1)}$$

C is the actual cluster solution in the sample, C′ is the cluster solution in the population data set, n_{11} is the number of pairs that are in the same cluster under C and C′, and n_{00} is the number of pairs that are in different clusters under C and C′.

Traditional simulation Study.

We specified the model from the NEO-PI-R with five factors and six variables each. To make the *traditional simulation* comparable to the *Real World model*, estimators of main factor loading and uniquenesses from the EFA of the norm data set were used for the simulations. The design contained three fixed factors: sample size, cross-loadings and factor

inter-correlations. In the first simulation condition we did not manipulate cross-loadings and factor inter-correlations, but took their estimators from the population data set (*population model*). Building on that, we manipulated them in four different ways, which were devised to best reflect the common practice in Monte Carlo simulations. Second, we set all cross-loadings and third, all factor inter-correlations to zero (*zero cross-loadings / zero factor correlations*). Fourth, we set one cross-loading per variable to a high value (*one cross-loading*) while taking factor inter-correlations from the estimators of the real data set. Fifth, we set inter-correlations to random values in the range given by the population data set (*varying factor inter-correlations*) and obtained cross-loadings from the real data set. We determined the values of cross-loadings by summing up the squares of all cross-loadings of each variable of the real data and then taking the square root. In doing so, we made sure to retain the communalities of all variables. Setting cross-loadings and/or factor inter-correlations to zero is an unrealistic but in simulation studies often used procedure. First, in all simulation conditions correlations of residual variances terms were also set to zero. Resulting loading matrices can be seen in Tables 1 and 2. Later on, we also included residual correlations into the simulation by adding the residual correlations we had found in the EFA of the population data set to the correlation matrix calculated for the *population model*. Residual correlations in the NEO-PI-R data set ranged between -.10 and .15 and in the IST-2000-R data set between -.04 and .06. Additionally, we applied different sample sizes (100; 200; 300; 400; 500 and 1,000). Multivariate normally distributed data are drawn from a distribution with given population covariance matrix that is calculated from loadings, factor inter-correlations and uniquenesses.

For *dimensionality assessment*, the MAP rule, PA-PAF, PA-PCA and AIC and BIC were used, and the variations described above were tested. This resulted in a fully crossed 5 x 6 x 5 design. Each of the resulting conditions was replicated 1,000 times to obtain reliable results from the simulations. For each method, the success rates and means of indicated

numbers of factors were reported. For the *assignment of variables*, a fully crossed 6 x 5 design was tested (six different sample sizes and five simulation models). Means and Rand Indexes in EFA then were calculated. All calculations were programmed in the open source software R 0.94.110 using the package 'psych' (Revelle, 2012).

Results

Real world simulation

Dimensionality assessment. In the population data set of the NEO-PI-R, the MAP rule and PA-PCA resulted in five factors while PA-PAF suggested six factors, which served as reference for the sub samples. Therefore, in the sub samples we expected five factors for the MAP rule, five factors for PA-PCA and six factors for PA-PAF. AIC reached its lowest value for the maximum number of factors possible, which was 29. BIC suggested 22 factors.[2]

The success rates for the sub samples can be obtained from Table 3. For smaller samples, PA-PAF indicated five factors to retain more often than PA-PCA, although in the population data set, PA-PAF had suggested six factors. For this reason, PA-PAF achieved comparatively low success rates in the sub-samples, as the reference was six factors. For samples with n > 250, in PA-PAF, the frequency of six factors increased and the frequency of five factors decreased. This finding suggests that PA-PAF, at least for the Real World data set, is sensitive for sample size. AIC and BIC both did not suggest the same number of factors as in the population data set in any of the Real World samples. This was caused by the effect that AIC, as expected, was highly sensitive for sample size (see Table 4). And therefore in the large population data set, AIC reached its minimum with the maximum number of possible

[2] Please note that from 15 factors on it was not possible to estimate communalities for PAF so 1s were used instead.

Table 3
Success Rates for dimensionality assessment in Real World Simulation for Different Sample Sizes for both Data Sets

n	MAP	PA-PAF	PA-PCA	AIC	BIC
			NEO-PI-R		
	5 Factors	6 Factors	5 Factors	29 Factors	22 Factors
100	.619	.023	.582	.000	.000
200	.822	.025	.912	.000	.000
300	.922	.025	.988	.000	.000
400	.966	.030	.997	.000	.000
500	.977	.038	1.00	.000	.000
1,000	.999	.100	1.00	.000	.000
			IST-2000-R		
	1 Factor	4 Factors	2 Factors	4 Factors	4 Factors
100	.994	.180	.577	.000	.000
200	1.00	.249	.789	.038	.002
300	1.00	.282	.926	.125	.012
400	1.00	.316	.973	.250	.029
500	1.00	.347	.991	.432	.073
1,000	1.00	.638	1.00	.947	.497

Notes. MAP=MAP rule; PA-PAF= Parallel Analysis with Principal Axis Factoring; PA-PCA= Parallel Analysis with Principal Component Analysis; AIC= Akaike information criterion; BIC=Bayesian information criterion; n=sample size. All success rates are based on 1,000 replications.

Table 4
Mean Number of indicated Factors for different sample sizes in Real World Samples, AIC and BIC - NEO-PI-R

	n					
	100	200	300	400	500	1,000
AIC	5.40	6.72	7.80	8.54	9.12	11.29
BIC	4.19	4.98	5.23	5.69	6.32	8.38

Note. AIC= Akaike information criterion; BIC=Bayesian information criterion. n= sample size. Facets are assigned to factors.

factors (29). Contrary to our expectations, BIC also seemed to be susceptible to sample size in the *Real World simulation*, increasing up to eight suggested factors for 1,000 observations (see Table 4).

In the IST-2000-R population data set, none of the used methods suggested the theoretically assumed three factors. The MAP rule indicated one and PA-PCA resulted in two factors while the other methods all found four factors. Consequently, these are the respective numbers we expected them to recover for the sub samples as well. As can be obtained from table 3, the MAP rule almost perfectly reproduced the one factor it had suggested in the population data set. Also PA-PAF performed satisfactory for at least medium sample sizes in finding two factors. In contrast, the other methods failed to recover the four factors they had previously suggested because they all resulted in smaller numbers of factors in most of the samples. Taking a closer look at the mean number of indicated factors (Table 5), in this data set none of the methods was overly sensitive for sample size while all methods (except for the MAP rule) showed at least a slight tendency towards sensitivity for sample size. Maybe this is because this data set reveals less residual correlations. It might also be caused by the fact that the total number of possible factors (i.e. the number of variables) was much lower.

Assignment of variables. Factor loadings on the five factors for the entire data set of the NEO-PI-R can be obtained from Table 1. When assigning facets to the factor with the highest absolute loading, EFA revealed a grouping of facets according to theoretical assumptions, except for facet N5 *impulsiveness*, which was assigned to the facets of the factor Extraversion instead of Neuroticism. Therefore, in our sub-samples, we also expected EFA to assign N5 to the factor Extraversion for a perfect fit (Rand Index = 1.0). In the sub-samples the Rand Index increased with increasing sample size, from .891 (100 obs.) to .967 (1,000 obs.).

When conducting an EFA for the entire data set of the IST-2000-R with three factors, all sub-tests showed the highest loadings on the factor they were theoretically assigned to. This was thus, the assignment pattern, we also expected in the sub samples. Here, the Rand Index increased from .832 for 100 observations to .996 for 1,000 observations. Although for 100 observations the Rand Index was slightly lower than in the NEO-PI-R data set, it was higher in the IST-2000-R for all other sample sizes.

Table 5
Mean Number of indicated Factors for different Sample sizes in Real World Samples IST-2000-R

	n					
	100	200	300	400	500	1,000
MAP	1.01	1.00	1.00	1.00	1.00	1.00
PA-PAF	3.02	3.25	3.28	3.32	3.35	3.64
PA-PCA	1.64	1.86	1.97	1.99	2.00	2.00
AIC	2.22	2.93	3.12	3.25	3.45	3.95
BIC	1.30	2.14	2.72	2.97	3.07	3.50

Note. MAP=MAP rule; PA-PAF= Parallel Analysis with Principal Axis Factoring; PA-PCA= Parallel Analysis with Principal Component Analysis; AIC= Akaike information criterion; BIC=Bayesian information criterion. n= sample size.

Traditional simulation study

Dimensionality assessment. Comparing the *zero cross-loadings* condition with the *population model* condition of the NEO-PI-R, all of the used EFA methods performed almost equally well in both conditions (see Table 6). The MAP rule and AIC seemed to achieve the best success rates. BIC achieved the poorest results in both conditions, but still converged to five factors in all *Traditional simulation* conditions in contrast to the results of the *Real World simulation*. In the condition where only one cross-loading per variable was set to a higher value, performance of the MAP rule and both PA methods decreased to success rates below .90. AIC and BIC yielded almost the same results as in the *zero cross-loadings* condition as well as the *population model* condition. When setting all factor inter-correlations to zero, all methods showed their best success rates. This suggests that the performance of these methods is more affected by the population factor inter-correlations than by cross-loadings, although factor inter-correlations were much lower than in other personality inventories (Van der Linden et al., 2012). When choosing random values in the range of the population data set, results were almost equal to the *population model*. Hence, this seems to be a reasonable way of simulating factor inter-correlations.

PA-PAF and AIC only in the Real World condition were sensitive for sample size and not in any of the simulation conditions where their success rates increased with increasing

Table 6

Success Rates and mean Number of Factors across all Sample Sizes for all Traditional Simulation Conditions - NEO-PI-R

	MAP	PA-PAF	PA-PCA	AIC	BIC
		Zero cross-loadings			
Success rate	.98	.95	.91	.98	.80
Mean	4.99	4.95	4.91	4.98	4.77
		One cross-loading			
Success rate	.97	.88	.76	.96	.77
Mean	4.97	4.88	4.76	4.97	4.77
		Zero factor inter-correlations			
Success rate	.99	1.00	1.00	.94	.68
Mean	4.99	5.00	5.00	5.05	4.51
		Varying factor inter-correlations			
Success rate	.99	.93	.88	. 98	.77
Mean	4.99	4.98	4.95	5.00	4.91
		Population model			
Success rate	.99	.96	.92	.99	.83
Mean	4.99	4.96	4.92	4.99	4.83
		Population model + residual correlations			
Success rate	.91	.93	.92	.14	.44
Mean	5.03	4.98	4.92	7.88	8.25
		Real World			
Ratio 5 Factors	.88	.91	.91	.13	.45
Mean	5.06	4.99	4.92	7.83	5.65

Note. MAP=MAP rule; PA-PCA= Parallel Analysis for Principal Component Analysis; PA-PAF= Parallel Analysis for Principal Axis Factoring; AIC= Akaike information criterion; BIC=Bayesian information criterion. Facets are assigned to factors. All success rates are based on 1,000 replications. In the *zero cross-loadings* and *one cross-loading* conditions the factor inter-correlations were set to *population model* values; In *the zero factor inter-correlations* and *varying factor inter-correlations* conditions cross-loadings were set to population model values. Sample sizes=100; 200; 300; 400; 500; 1,000.

sample size. Given the fact that the only parameter that was not included into the model for computing the population correlation matrix in the *traditional simulation* study was the residual correlations, one can assume that the non-zero residual correlations, present only in the *Real World* condition, caused this sample size effect. To prove this assumption, we performed an additional simulation in which we included the residual correlations from the

Real World data to the *population model*. Results of this simulation can be obtained from Table 6. As expected, success rates of all methods were almost identical to those from the *Real World* condition and therefore, for AIC and BIC much worse than when population parameters without residual correlations were simulated. AIC, in contrast to the PA methods and the MAP rule, does not seem to be able to indicate a useful number of factors in the presence of residual correlations despite its very accurate detection of five factors for every other simulation condition.

PA-PCA revealed big problems with the comparatively high factor inter-correlations with the IST-2000-R data set (see Table 7). When setting them to zero, PA-PCA jumped back to almost 100% success rate. PA-PCA seems to perform bad only in presence of high factor inter-correlations. The MAP rule however does not find the right number of factors even when factor inter-correlations are set to zero. Given the specific loading and factor inter-correlation pattern in the IST-2000-R data set, the MAP rule seems to detect one factor no matter what specific simulation is used. Across all simulation conditions PA-PAF showed the best success rates compared to the other methods across simulation conditions. Even though PA-PAF showed a slight dependency on sample size in the NEO-PI-R data set and was therefore unable to achieve a good performance in the *Real World simulation*, this method seems to be the most consistent one across all simulation conditions and data sets. It is the only method that did not show the same dramatic drop as all the other methods in presence of factor inter-correlations. Both AIC and BIC performed reasonably well in almost all simulation conditions, although they did not overall reach the same performance as in the NEO-PI-R data set. The decrease from the *population model* without residual correlations to the *population model + residual correlations* though was by far not as bad as in the NEO-PI-R data set given the residual correlations where much lower. For the same reason, performance was also better in the *Real World simulation*. Across all conditions, BIC had the lowest success rates of all criteria in both data sets.

Table 7

Success Rates and mean Number of Factors across all Sample Sizes for all Traditional Simulation Conditions - IST-2000-R

	MAP	PA-PAF	PA-PCA	AIC	BIC
		Zero cross-loadings			
Success rate	.00	.93	.09	.81	.56
Mean	1.00	3.00	1.94	2.78	2.31
		One cross-loading			
Success rate	.00	.88	.00	.76	.47
Mean	1.00	2.90	1.26	2.72	2.15
		Zero factor inter-correlations			
Success rate	.06	.99	1.00	.98	.90
Mean	1.47	3.01	3.00	2.98	2.90
		varying factor inter-correlations			
Success rate	.00	.76	.08	1.00	.99
Mean	1.00	2.99	1.96	2.81	2.41
		Population model			
Success rate	.00	.93	.04	.82	.03
Mean	1.00	2.99	1.93	2.80	2.44
		Population model + residual correlations			
Success rate	.00	.59	.04	.56	.68
Mean	1.00	3.36	1.95	3.16	2.50
		Real World			
Success rate	1.00	.27	.85	.30	.50
Mean	1.00	3.24	1.89	3.15	2.61

Note. MAP=MAP rule; PA-PCA= Parallel Analysis for Principal Component Analysis; PA-PAF= Parallel Analysis for Principal Axis Factoring; AIC= Akaike information criterion; BIC=Bayesian information criterion. Facets are assigned to factors. All success rates are based on 1,000 replications. In the *zero cross-loadings* and *one cross-loading* conditions the factor inter-correlations were set to *population model* values; In *the zero factor inter-correlations* and *varying factor inter-correlations* conditions cross-loadings were set to population model values. Sample sizes=100; 200; 300; 400; 500; 1,000.

Previous results, that AIC tends to overfactor with large sample sizes (Cudeck & Henly, 1991; McDonald, 1989; Vrieze, 2012) could be confirmed only in the *Real World* data of NEO-PI R and seems to be due to the moderate residual correlations (-.10 – .15). It does not occur with small residual correlations as in the IST-2000-R data set (-.04 – .06).

All of the investigated *dimensionality assessment* methods proved to be more sensitive for factor inter-correlations than to cross-loadings on the level present in these data sets, even in the NEO-PI-R data set where factor inter-correlations are comparatively low. This dependency resulted in a dramatic drop in performance for the IST-2000-R data set that almost disappeared completely when setting factor inter-correlations to zero. For the *dimensionality assessment* we found no substantial difference among the methods regarding presence or absence of population cross-loadings, whereas when factor inter-correlations were set to zero, almost all methods had a success rate of 1.00. These findings are congruent with other previous findings (Bacon, 2001; Beauducel, 2001).

Assignment of variables. For the *traditional simulation*, we used the factor solution with five factors. Results can be obtained from Table 8. In the NEO-PI-R data set, EFA showed its best performance in the *zero cross-loadings* condition with almost 100% correct, and its poorest performance in the *one cross-loading* condition, where its Rand Index was even lower than in the *Real World simulation*. In the IST-2000-R data set the Rand Index was best when all factor inter-correlations were set to zero and worst for *varying factor inter-correlations*, while the *one cross-loading* condition still showed a poor performance. The fact that in the IST-2000-R data factor inter-correlations also for *assignment of variables* mattered the most might be due to their relatively high level here. Our results concerning the *assignment of varialbes* confirmed the previous results of Gerbing and Hamilton (1997): in cases of low factor inter-correlations and no cross-loadings, almost perfect structure detection can be achieved. Only when applying cross-loadings, and especially when both cross-loadings and factor correlations are present, performance of EFA decreases substantially (Rand Indexes of .95 or less).

Table 8
Rand Indexes across all Sample Sizes for Real World Simulation and Traditional Simulation for the NEO-PI-R Data Set and the IST-2000-R Data Set

	Data set	
Simulation condition	NEO-PI-R	IST-2000-R
Zero cross-loadings	1.00	.98
One cross-loading	.93	.95
Zero factor inter-correlations	.96	1.00
Varying factor inter-correlations	.95	.93
Population model	.95	.96
Real world	.93	.94

Note. NEO-PI-R data set: Facets are assigned to factors. All Rand Indexes are based on 1,000 replications. In the *zero cross-loadings* and *one cross-loading* conditions the factor inter-correlations were set to *population model* values; In *the zero factor inter-correlations* and *varying factor inter-correlations* conditions cross-loadings were set to population model values. Sample sizes=100; 200; 300; 400; 500; 1,000.

Discussion

The present paper examined data set characteristics of real data sets and their impact on the performance of factor structure detection methods in psychometrics. As an example, two real large data sets were examined and the consistency of EFA methods in sub samples of different sizes was analyzed. This new *Real World simulation* addresses the issue of lacking validity of *traditional simulation* studies, i.e. Monte Carlo studies, by detecting crucial data set characteristics from real data sets and manipulating them systematically in a subsequent simulation study. It therefore enables researchers to better understand data sets in psychological settings. Constellations are simulated that represent the real world in psychological data and thus results of these simulations yield information that is relevant to what happens in practice. When applying unrealistic conditions to simulations, produced results often are misleading. We found that when using one high cross-loading per variable, a common procedure in *traditional simulation* studies (e.g. Zwick & Velicer, 1986), all

methods seem to perform worse than in the other simulation conditions in which the same total amount of cross-loadings per variable was applied. These results suggest that the manner in which cross-loadings are manipulated is at least as important as the level of cross-loadings. But the results reported here show that not only the performance of a particular method depends on characteristics of a model and its violations in the particular data set but also the ranking of different methods can change. For example, for the NEO PI-R data set, we demonstrated a dramatic decrease in performance of the MAP rule and AIC, formerly the best methods, when introducing the small residual correlations from the population data, whereas the PA methods were hardly affected at all.

One particular finding of this exemplary *Real world simulation* is that there is no sense in establishing general statements on an absolute minimum sample size or a minimum sample size depending only on the level of communality. The influence of sample size on structure recovery depends on multiple parameters. The results of this study show that one important parameter is the residual correlations in small sizes such as they are present in nearly every real data set. Once they are included in the specified correlation matrix, some methods become very sensitive for sample size and therefore their performance decreases with increasing sample size. This effect was strongest for AIC and lowest for both PA methods. Since AIC does not assume that there is a true model but only aims to minimize the error of prediction (Vrieze, 2012), its sensitivity for sample size in presence of residual correlations is not surprising. The larger the sample size becomes the more residual correlations can be explained by further minor factors and, therefore, the prediction can be optimized. Yet, one needs to bear in mind that residual correlations in the height present in the NEO-PI-R data set, indicate a non-diagonal matrix $\mathbf{U}^2$ and imply model violations that might indicate model misfits. Consequently, one fair conclusion is that in case of the NEO-PI R data set, the five-factor model is possibly not entirely correct, even though it might be the practically most useful one. So, given the existing residual correlations, there are many small additional factors

that need to be considered when aiming to explain the observed correlation matrix R (p x p). In such cases, the data structure in oblique EFA (and under the assumption that the minor factors themselves are not inter-correlated) is then

$$R = \Lambda\phi\Lambda' + \mathrm{MM}' + U^2$$

with M as a k x q matrix of k minor factor loadings of the p variables.

This might be the reason why also BIC is sensitive for sample size in the *Real World* data. Even BIC cannot be expected to converge if the true model is not among the proposed. Indicating more factors with increasing sample size is not necessarily undesired in data with residual correlations. Still, this decision depends upon the particular aim of the study. Possibly, all previously found relations between sample size and model recovery were also influenced by the particular model and the level of model violations in the data i.e. the data set characteristics. This is a possible explanation why in some studies researchers did find an effect of sample size and in some they did not depending on the specified models in the simulations (e.g. Browne, 1968b; Guadagnoli & Velicer, 1988) and the data set characteristics in real data (Arrindell & Ende, 1985; Barrett & Kline, 1981).

Please note that with the parameters we manipulated in the subsequent simulation study we were not able to explain all the variations of the performance of the methods in the real data sets. What is left to explain for example, is that the decrease in performance of the MAP rule in the IST data that was not reduced even after setting factor inter-correlations to zero. By additionally varying the number of variables per factor or the number of factors, this issue can probably be resolved. But this shows again how important the use of real data is to receive indications about important data set characteristics.

Practical implications

For the application of EFA in practice, it shall be noted that no definite statement can be made on which method is superior to all the others not even depending on sample size. The ranking of EFA methods strongly depends on data set characteristics and the aim of the study.

In particular, this study shows that EFA results are highly unstable against even small variations of specific parameters as for example residual correlations. For a comprising analysis of the underlying structure, different methods shall thus be used in combination. For example, factor inter-correlations appear to be the decisive issue for the election of the *dimensionality assessment* method to be used. PA-PAF seems to be a good choice in presence of factor inter-correlations. Further, practitioners are advised to examine residual correlations closely when using the MAP rule or PA, since they are not very sensitive for such model violations. Whenever AIC and BIC suggest more factors than PA methods and the MAP rule, correlated residual variance is likely and minor factors should be considered. Anyhow, the particular data set characteristics of the data set under examination shall be examined comprehensively. Maybe, in the end the best solution will be to run data specific simulations prior to any analysis for selection of the best method for the particular data set (Muthén & Muthén, 2002).

For the examination of the comparative performance of statistical methods in simulations, the results of this study suggest that characteristics of real data sets should be considered exhaustively before starting the simulation study. Based on single aspects as sample size or number of variables per factor, no final conclusions can be drawn as long as the impact of their interaction with other data set characteristics is neglected. The *Real World simulation* technique is one possible method to identify key data set characteristics for the performance of specific methods from real data sets. The association of data set characteristics with specific performance rankings may subsequently be used for *traditional simulation* studies. Here, these core characteristics are varied systematically and thus their causal relation can be tested. Future work should address the following issues: Collecting data set characteristics from real large data sets, that are crucial for the performance of the specific method tested and examination of the performance of these methods in systematic simulation studies where these realistic variations are tested.

Akaike, H. (1974). A new look at the statistical model identification. *Automatic Control, IEEE Transactions On*, *19*(6), 716–723.

Amthauer, R., Brocke, R., Liepmann, D., & Beauducel, A. (2007). *Intelligenz-Struktur-Test 2000 R - IST 2000-R* (2. erw.). Göttingen: Hogrefe.

Arrindell, W. A., & Ende, J. van der. (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.

Bacon, D. R. (2001). An evaluation of cluster analytic approaches to initial model specification. *Structural Equation Modeling*, *8*(3), 397–429.

Barrett, P. T., & Kline, P. (1981). The observation to variable ratio in factor analysis. *Personality Study & Group Behaviour*, *1*(1), 23–33.

Beauducel, A. (2001). Problems with parallel analysis in data sets with oblique simple structure. *Methods of Psychological Research Online*, *6*(2), 141–157.

Browne, M. W. (1968). A comparison of factor analytic techniques. *Psychometrika*, *33*(3), 267–334.

Burnham, K. P., & Anderson, D. R. (2004). Multimodel inference understanding AIC and BIC in model selection. *Sociological Methods & Research*, *33*(2), 261–304.

Church, A. T., & Burke, P. J. (1994). Exploratory and confirmatory tests of the Big Five and Tellegen's three- and four-dimensional models. *Journal of Personality and Social Psychology*, *66*(1), 93–114.

Crawford, A. V., Green, S. B., Levy, R., Lo, W. J., Scott, L., Svetina, D., & Thompson, M. S. (2010). Evaluation of parallel analysis methods for determining the number of factors. *Educational and Psychological Measurement*, *70*(6), 885–901.

Cudeck, R., & Henly, S. J. (1991). Model selection in covariance structures analysis and

the problem of sample size: A clarification. *Psychological Bulletin*, *109*(3), 512–519.

Fabrigar, L. R., Wegener, D. T., MacCallum, R., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods*, *4*(3), 272.

Fan, X., & Sivo, S. A. (2005). Sensitivity of fit indexes to misspecified structural or measurement model components: Rationale of two-index strategy revisited. *Structural Equation Modeling*, *12*(3), 343–367.

Gerbing, D. W., & Hamilton, J. G. (1996). Viability of exploratory factor analysis as a precursor to confirmatory factor analysis. *Structural Equation Modeling: A Multidisciplinary Journal*, *3*(1), 62–72.

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), 377–793.

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

Haynes, C. A., Miles, J. N. V., & Clements, K. (2000). A confirmatory factor analysis of two models of sensation seeking. *Personality and Individual Differences*, *29*(5), 823–839.

Hayton, J. C., Allen, D. G., & Scarpello, V. (2004). Factor retention decisions in exploratory factor analysis: A tutorial on parallel analysis. *Organizational Research Methods*, *7*(2), 191–205.

Homburg, C. (1991). Cross-validation and information criteria in causal modeling. *Journal of Marketing Research*, 137–144.

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, *30*(2), 179–185.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure

analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*(1), 1–55.

Humphreys, L. G., & Montanelli Jr., R. G. (1975). An Investigation of the Parallel Analysis Criterion for Determining the Number of Common Factors. *Multivariate Behavioral Research*, *10*(2), 193–205.

Jackson, D. A. (1993). Stopping rules in principal components analysis: a comparison of heuristical and statistical approaches. *Ecology*, 2204–2214.

Jackson, D. L. (2007). The effect of the number of observations per parameter in misspecified confirmatory factor analytic models. *Structural Equation Modeling*, *14*(1), 48–76.

Kaplan, D. (1988). The Impact of Specification Error on the Estimation, Testing, and Improvement of Strucutural Equation Models. *Multivariate Behavioral Research*, *23*(1), 69–86.

Kuha, J. (2004). AIC and BIC comparisons of assumptions and performance. *Sociological Methods & Research*, *33*(2), 188–229.

Lance, C. E., Butts, M. M., & Michels, L. C. (2006). The Sources of Four Commonly Reported Cutoff Criteria What Did They Really Say? *Organizational Research Methods*, *9*(2), 202–220.

Lubke, G., & Neale, M. C. (2006). Distinguishing between latent classes and continuous factors: Resolution by maximum likelihood? *Multivariate Behavioral Research*, *41*(4), 499–532.

MacCallum, R., & Tucker, L. R. (1991). Representing sources of error in the common-factor model: Implications for theory and practice. *Psychological Bulletin*, *109*(3), 502–511.

MacCallum, R., Widaman, K. F., Preacher, K. J., & Hong, S. (2001). Sample Size in Factor

Analysis: The Role of Model Error. *Multivariate Behavioral Research*, *36*(4), 611–637.

McDonald, R. P. (1989). An index of goodness-of-fit based on noncentrality. *Journal of Classification*, *6*(1), 97–103.

Mundfrom, D. J., Shaw, D. G., & Ke, T. L. (2005). Minimum sample size recommendations for conducting factor analyses. *International Journal of Testing*, *5*(2), 159–168.

Muthén, L. K., & Muthén, B. O. (2002). How to Use a Monte Carlo Study to Decide on Sample Size and Determine Power. *Structural Equation Modeling: A Multidisciplinary Journal*, *9*(4), 599–620.

Ostendorf, F., & Angleitner, A. (2004). *NEO-Persönlichkeitsinventar nach Costa und McCrae, Revidierte Fassung*. Göttingen: Hogrefe.

Patil, V. H., Singh, S. N., Mishra, S., & Todd Donavan, D. (2008). Efficient theory development and factor retention criteria: Abandon the "eigenvalue greater than one"criterion. *Journal of Business Research*, *61*(2), 162–170.

Pennell, R. (1968). The influence of communality and on the sampling distributions of factor loadings. *Psychometrika*, *33*(4), 423–439.

Rand, W. M. (1971). Objective Criteria for the Evaluation of Clustering Methods. *Journal of the American Statistical Association*, *66*(336), 846–850.

Revelle, W. (2012). psych: Procedures for Personality and Psychological Research. (Version 1.0-91).

Ruscio, J., & Roche, B. (2012). Determining the number of factors to retain in an exploratory factor analysis using comparison data of known factorial structure. *Psychological Assessment*, *24*(2), 282.

Sass, D. A. (2010). Factor loading estimation error and stability using exploratory factor analysis. *Educational and Psychological Measurement*, *70*(4), 557–577.

Sass, D. A., & Schmitt, T. A. (2010). A comparative investigation of rotation criteria

within exploratory factor analysis. *Multivariate Behavioral Research*, *45*(1), 73–103.

Schmitt, T. A., & Sass, D. A. (2011). Rotation criteria and hypothesis testing for exploratory factor analysis: Implications for factor pattern loadings and interfactor correlations. *Educational and Psychological Measurement*, *71*(1), 95–113.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461–464.

Smith, R. M. (1996). A comparison of methods for determining dimensionality in Rasch measurement. *Structural Equation Modeling: A Multidisciplinary Journal*, *3*(1), 25–40.

Van der Linden, D., Tsaousis, I., & Petrides, K. V. (2012). Overlap between General Factors of Personality in the Big Five, Giant Three, and trait emotional intelligence. *Personality and Individual Differences*.

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

Velicer, W. F., Eaton, C. A., & Fava, J. L. (2000). Construct explication through factor or component analysis: A review and evaluation of alternative procedures for determining the number of factors or components. In *Problems and solutions in human assessment* (pp. 41–71). Springer.

Velicer, W. F., & Fava, J. L. (1998). Effects of variable and subject sampling on factor pattern recovery. *Psychological Methods*, *3*(2), 231.

Vrieze, S. I. (2012). Model selection and psychological theory: a discussion of the differences between the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). *Psychological Methods*, *17*(2), 228.

Yuan, K., Marshall, L. L., & Bentler, P. M. (2003). Assessing the Effect of Model Misspecifications on Parameter Estimates in Structural Equation Models. *Sociological*

Methodology, *33*(1), 241–265.

Zwick, W. R., & Velicer, W. F. (1982). Factors influencing four rules for determining the number of components to retain. *Multivariate Behavioral Research*, *17*(2), 253–269.

Zwick, W. R., & Velicer, W. F. (1986). Comparison of five rules for determining the number of components to retain. *Psychological Bulletin*, *99*(3), 432.