Factor Retention Revised: Analyzing Current Practice and Developing New Methods

David Goretzko

München, 2020

Factor Retention Revised: Analyzing Current Practice and Developing New Methods

Inauguraldissertation

zur Erlangung des Doktorgrades der Philosophie

an der Ludwig-Maximilians-Universität München

vorgelegt von

David Goretzko
aus Hameln

München, 2020

Erstgutachter: Prof. Dr. Markus Bühner

Zweitgutachter: Prof. Dr. Moritz Heene
Datum der mündlichen Prüfung: 20.05.2020

1 Abstract

The present thesis consists of three studies covering different methodological decisions that have to be made when conducting exploratory factor analyses. Study 1 is a review on the current practice in psychological research and new developments concerning these methodological decisions, while both Study 2 and Study 3 focus on the issue of factor retention which is the determination of the number of factors. In Study 2, a new method - combining extensive data simulation with modern machine learning modelling - is proposed. This new approach was able to outperform common factor retention criteria in a large-scale simulation study where the number of factors, the number of manifest variables, the sample size, the loading magnitudes of primary and cross-loadings, the inter-factor correlations and the variables per factor were varied. As Study 2 focuses on the accuracy of different factor retention criteria trying to approximate the data generating process, Study 3 rather deals with the reproducibility of factor solutions. Bootstrapping is proposed as a way to assess the robustness of factor solutions against sampling error which then can be used as a proxy for replicability. Demonstrating this connection between robustness and replicability, Study 3 also shows that the new approach suggested in Study 2 has higher replication rates than other criteria and therefore seems to perform well not only for simulated data, but also on real data sets.

Since the current research practice often lacks informed decisions, especially with regard to the factor retention process (Study 1), the new factor retention criterion (Study 2) needs to be refined (different models need to be created for different data conditions) and then made available to a broad research community. Providing practitioners with an easy-to-use, yet accurate retention criterion may improve the application of EFA. Until then, bootstrapping or another way of evaluating the robustness of common factor retention criteria can be used as a confidence measure (and as a proxy for replicability) to choose the best retention criterion (and the best factor solution respectively) for a given context.

2 Zusammenfassung

Die vorliegende Arbeit ist aus drei Manuskripten (im Folgenden Studie 1, Studie 2 und Studie 3 genannt) aufgebaut, die verschiedene Aspekte der Exploratorischen Faktorenanalyse (EFA) beleuchten. Bei der EFA handelt es sich um eine statistische Methode zur Untersuchung latenter Variablen, die als ursächlich für die Zusammenhangsstrukturen mehrerer manifester Variablen angenommen werden. In der den Manuskripten vorangestellten Einleitung wird das (multidimensionale) tau-kongenerische Messmodell der klassischen Testtheorie - im Kontext der EFA auch *common factor model* genannt - eingeführt und die Herausforderungen und Fallstricke der EFA-Durchführung vorgestellt. Die zentralen Aspekte bei der EFA sind dabei das Studiendesign (womit in erster Linie die Stichprobenumfangsplanung gemeint ist), die Wahl einer Schätzmethode (*Extraction Method*), die Wahl einer Rotationsmethode (*Rotation Method*) und die Bestimmung der Faktorenanzahl (*Factor Retention*).

In Studie 1 wird die aktuelle Anwendung der EFA im Rahmen eines umfangreichen Reviews untersucht und die neuesten methodologischen Entwicklungen und Erkenntnisse im Hinblick auf die vier genannten Hauptaspekte der EFA diskutiert. Die Mehrzahl (50.3%) der untersuchten EFAs basierten auf Stichproben mit mehr als 400 Beobachtungen, was gemäß verschiedener Simulationsstudien (z.B. MacCallum, Widaman, Zhang, & Hong, 1999; Mundfrom, Shaw, & Ke, 2005) als Mindeststichprobe empfohlen werden kann (da man die Höhe von Kommunalitäten und den Grad der *Overdetermination*, welcher der Anzahl der Items, die jedem Faktor zugeordnet werden können, entspricht, nicht unbedingt vor der Datenerhebung absehen kann). Dies könnte für eine verbesserte Praxis im Vergleich zu den Befunden von Fabrigar, Wegener, MacCallum, und Strahan (1999) zwanzig Jahre zuvor sprechen, als noch an Regeln zur Stichprobenplanung festgehalten wurde, die die benötigte Größe der Stichprobe in Abhängigkeit von der Variablenanzahl abschätzen. Im Hinblick auf die verwendeten Rotationsmethoden zeigt sich, dass vermehrt auf oblique Rotationsmethoden gesetzt wird, was bereits bei Fabrigar et al. (1999) empfohlen wurde, jedoch dass beinahe nie unterschiedliche Rotationstechniken verglichen werden. Studie 1 diskutiert deshalb, welche Rotationsmethoden welche Annahmen treffen und wann sie entsprechend Anwendung finden sollten und fordert analog zu den Ausführungen von Browne (2001), dass verschiedene Rotationstechniken (falls möglich auch auf Teildatensätzen) getestet werden sollten. Außerdem werden sogenannte regularisierte Faktorenanalysen vorgestellt, welche den zusätzlichen Rotationsschritt in der EFA überflüssig machen könnten.

Hinsichtlich der Faktorenextraktion werden Simulationsstudien vorgestellt, die die verschiedenen Schätzmethoden vergleichen (z.B. Barendse, Oort, & Timmerman, 2015; De Winter & Dodou, 2012). Während die Hauptachsenanalyse (*Principal Axis Factoring*) am häufigsten in der aktuellen Forschung angewendet wird, argumentiert Studie 1, dass Maximum-Likelihood EFA für multivariat normalverteilte Daten und *weighted-least-squares*-Ansätze für ordinale Daten (speziell wenn die Anzahl der Antwortkategorien kleiner als fünf ist) - aufgrund der vorhandenen Fit-Indizes und der besseren Vergleichbarkeit mit konfirmatorischen Faktorenanalysen zur Validierung der Faktorstruktur - der Hauptachsenanalyse vorzuziehen sind. Bei der Bestimmung der Faktorenanzahl verlassen sich viele Anwender der EFA immer noch auf Methoden, die sich in zahlreichen Untersuchungen als nicht reliabel herausgestellt haben (z.B. im Fall des Kaiser-Kriteriums, Zwick & Velicer, 1986), aber in Statistikprogrammen wie *SPSS* (IBM Corp., 2019) die Standardeinstellung sind. Da die Parallelanalyse (Horn, 1965), die unter anderem aufgrund der Robustheit gegenüber unterschiedlichen Verteilungen der Daten (Dinno, 2009) als bisheriger "Goldstandard" gilt, in vielen Datenbedingungen keine akkurate Bestimmung der Faktorenanzahl erlaubt und moderne Methoden nur in manchen dieser Bedingungen überlegen sind, empfiehlt Studie 1 mehrere Kriterien (wenn möglich auf Teildatensätzen) zu vergleichen.

Neben neuen Kriterien zur Bestimmung der Faktorenanzahl existieren auch Kombinationsregeln, die vorgeben, wie Anwender der EFA auf Basis mehrerer dieser Methoden zu einer finalen Einschätzung der Dimensionalität kommen sollen (z.B. Auerswald & Moshagen, 2019). Da diese Kombinationsregeln und generell der Vergleich mehrer Methoden aufwendig und für Anwender mit Unsicherheiten (z.B. Welcher Methode ist in der spezifischen Situation eher zu trauen?) verbunden sind, wird in Studie 2 ein neuer Ansatz zur Bestimmung der Faktorenanzahl - genannt *Factor Forest* - vorgestellt. Dieser Ansatz verbindet eine umfassende Datensimulation mit dem Training eines modernen Machine-Learning (ML) Modells, das auf Basis von Eigenschaften der empirschen Daten die korrekte Faktorenanzahl vorhersagen soll.

Dafür wurden in Studie 2 zunächst 500000 Datensätze simuliert, wobei die Anzahl der latenten Faktoren ($k \in \{1, 2, ..., 8\}$), die Anzahl der manifesten Variablen ($p \in \{4, ..., 80\}$), die Stichprobengröße ($N \in [200; 1000]$), die Ladungshöhen (Haupt- und Nebenladungen) und die Korrelationen zwischen den latenten Variablen variierten. Anschließend wurden 181 *features* - also Variablen, die zur Vorhersage der Faktorenanzahl verwendet werden sollten - für jeden der Datensätze berechnet. Bei diesen Variablen handelte es sich überwiegend um Größen, die die empirische Korrelationsmatrix beschreiben (Eigenwerte, Matrixnormen,

etc.), da die Zerlegung der Korrelationsmatrix in eine systematische Komponente (der Teil der Varianz der manifesten Variablen, die durch die latenten Variablen erklärt wird) und eine unsystematische Komponente (*unique variance*) zentraler Bestandteil der EFA ist. Die aus den simulierten Datensätzen extrahierten *features* bildeten die Basis (das Trainingsset) für die Anwendung der ML-Algorithmen. Der neue Ansatz (die trainierten ML-Modelle) wurde anschließend in einer großen Simulationsstudie mit vier herkömmlichen Methoden (Parallelanalyse, Kaiser-Kriterium, Comparison Data und Empirical Kaiser Criterion) in 3204 Datenbedingungen hinsichtlich der Genauigkeit verglichen (pro Bedingung wurden 500 Replikationen durchgeführt). Dabei erzielte ein trainiertes *xgboost*-Modell (für den *xgboost* Algorithmus, siehe Chen & Guestrin, 2016) mit durchschnittlich 92.9% die höchste Genauigkeit. In einem zweiten Schritt wurde das *xgboost*-Modell noch weiter verbessert, indem sowohl sechs Hyperparameter des Algorithmus getuned (optimiert) und die Lösung der Parallelanalyse, des Comparison Data Ansatzes und des Empirical Kaiser Criterion als zusätliche *features* in das Modell aufgenommen wurden. Dadurch erzielt das finale *xgboost*-Modell eine Genauigkeit (*out-of-sample*) von 99.3%. Da die Ergebnisse aus Studie 2 auf multivariat-normalverteilten Daten basieren und psychologische Forschung (und damit viele Faktorenanalysen) auf ordinalen Fragebogendaten fußt, wurde das Vorgehen aus Studie 2 auch für ordinale Daten mit vier bis sieben Antwortkategorien getestet. Der ordinale *Factor Forest* erzielte eine Genauigkeit (*out-of-sample*) von durchschnittlich 98.5%, weshalb davon auszugehen ist, dass der neue Ansatz bei entsprechendem Training auf passenden Daten (die Trainingsdaten müssen das Anwendungsfeld möglichst umfassend abdecken) eine sehr hohe Genauigkeit liefert und deshalb als eigenständige Methode zur Bestimmung der Faktorenanzahl herangezogen werden kann. Die Ergebnisse dieser zusätzlichen Analyse werden im Kapitel “Additional Analyses: Ordinal Factor Forest” dieser Dissertation berichtet.

Der *Factor Forest* ermöglicht nicht nur eine genaue Vorhersage der Dimensionalität, er liefert zusätzlich Schätzwerte für die Wahrscheinlichkeit verschiedener Faktoranzahlen. Dies kann als Maß für die Sicherheit gesehen werden, dass die vorhergesagte Faktorenanzahl der wahren Dimensionalität entspricht. Herkömmliche Verfahren liefern in der Regel nur Punktschätzer für die Anzahl der Faktoren, so dass der Anwender keinen Anhaltspunkt für die Stabilität der Faktorlösung hat - bzw. dafür, ob Stichprobenbesonderheiten das Ergebnis verzerren (*sampling error*). Bei entsprechender Datenlage wird die Replikation der Faktorenlösung unter Umständen zum Problem. Obwohl im Zuge der Replikationskrise viele methodische Praktiken in der Psychologie auf dem Prüfstand stehen, wird die Bestimmung der Faktorenanzahl selten vor dem Hintergrund der Replizierbarkeit gesehen (Osborne &

Fitzpatrick, 2012). Für eine erfolgreiche Replikation, bzw. Kreuz-Validierung mittels konfirmatorischer Faktorenanalyse, erscheint jedoch die korrekte Bestimmung der Faktorenanzahl unabdingbar. Entsprechend untersucht Studie 3 den Zusammenhang zwischen der Robustheit der Methoden zur Bestimmung der Faktorenanzahl über Bootstrap-Stichproben hinweg und ihrer erfolgreichen Replikation. Dafür wurde mit vier verschiedenen Methoden die Faktorenanzahl für 19 Datensätze mit Persönlichkeitsmaßen (das *10 Item Big Five Inventory* von Rammstedt, Kemper, Klein, Beierlein und Kovaleva (2017), welches für eine *within-person* Replikation genutzt wurde und das *Big Five Structure Inventory* von Arendasy (2009), das den *between-person* Replikationskontext abbildet) bestimmt und die Robustheit dieser Lösung über 100 Bootstrap-Stichproben abgeschätzt. Es zeigte sich anschließend ein positiver Zusammenhang zwischen der Robustheit der Faktorenlösung und ihrer Replizierbarkeit. Während der *Factor Forest* und das Empirical Kaiser Criterion relativ robuste Lösungen über die Bootstrap-Stichproben hinweg lieferten und folglich höhere Repliaktionsraten aufwiesen, zeigten die Parallelanalyse und Comparison Data geringere Robustheit und schlechtere Replizierbarkeit.

Contents

List of Figures

List of Tables

3 General Introduction

Exploratory factor analysis (EFA) is a statistical method commonly used in psychological research to evaluate the intercorrelations among a set of observed variables and to discover underlying latent structures. The basic idea is that latent (unobservable) variables - namely psychological constructs like intelligence or personality traits - are represented by manifest variables which is known as the "common cause relation" (Reichenbach, 1956 as cited in Haig, 2005). Conversely, this means that using EFA one can find latent variables that can explain observations of a given set of manifest items. Spearman (1904) was the first to formulate the basic concept of factor analyses, followed by several researchers mainly from the field of intelligence research improving the methodology (Bartholomew, 1995). In this process, the influence of Thurstone was particularly important as he coined the terms "communality" and "uniqueness" (1940) and advocated the aim of simple structure solutions (1947). It was also Thurstone (1947) who formulated the multidimensional factor model - also known as the common factor model which reflects the (multidimensional) congeneric measurement model that is known from classical test theory (Jöreskog, 1971).

Within the process of questionnaire development (or test construction), EFA plays a very important role. Usually, conducting several EFAs is the starting point when indicators for a specific psychological construct are evaluated and subfacets of these constructs are explored. This procedure is often entangled with the development and the refinement of theories. In personality psychology, the most prominent example for the impact of EFA on construct definition and theory development is the history of the big five trait taxonomy as described by John and Srivastava (1999). However, the relevance of EFA is not limited to personality psychology as there are application context in all psychological disciplines - i.e. intelligence research (e.g. Cohen, 1957), organizational psychology (e.g. Smith, Organ, & Near, 1983), developmental psychology (e.g. Ronald, Happé, Hughes, & Plomin, 2005), clinical psychology (e.g. Comrey, 1957) or social psychology (e.g. Marsh, Barnes, & Hocevar, 1985). Thus, EFA is arguably one of the most influential statistical methods in psychological research. However, reviews considering the application of EFA (e.g. Fabrigar et al., 1999) have shown that the actual research practice often lacks informed methodological decisions - especially with regard to selecting an extraction method, a rotation method and a factor retention criterion. As determining the number of factors is probably "the most important decision a researcher will make" (Zwick & Velicer, 1986) when conducting an EFA, this thesis pays particular attention to the factor retention process presenting new approaches that cover different of its aspects. The main focus of this work is on an innovative method that promises to be accurate, user-friendly and replicable when estimating the number of factors. As this new factor retention criterion is tailored to specific application contexts, this thesis first reviews both the current use of EFA and methodological developments in this field (with regard to all major methodological decisions in EFA).

3.1 Manuscripts of this Thesis

The following manuscripts contain the three studies this thesis is based upon:

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

2. Goretzko, D., & Bühner, M. (under review). One model to rule them all? Using machine learning algorithms to determine the number of factors in exploratory factor analysis.[1]

3. Goretzko, D., & Bühner, M. (under review). Two factors, or rather four? Robustness of factor solutions in exploratory factor analysis.

Hereafter, Study 1 refers to the first manuscript, Study 2 to the second and Study 3 to the third respectively. Study 1 is a review on the current practice and methodological developments within EFA research focusing on the major decisions a researcher has to make - which extraction method to use, which rotation method to apply and how many factors to retain. In Study 2, a new approach for determining the number of factors in EFA is proposed and evaluated, whereas Study 3 focuses the relation of robustness and replicability in factor retention.

All manuscripts were written by the author of this thesis. Markus Bühner acted as the supervising author of all three papers, while Trang T. H. Pham collected data for the review on the current use of the EFA in Study 1. The idea and conception of the studies, especially the development of the new factor retention criterion in Study 2 was generated solely by the author of this thesis. As all manuscripts were created in consultation with the

[1] The article was published with some minor changes after the submission of this thesis: Goretzko, D., & Bühner, M. (2020). One model to rule them all? Using machine learning algorithms to determine the number of factors in exploratory factor analysis.*Psychological Methods.* doi:10.1037/met0000262

co-authors, the pronoun we is used in the summary of Study 1, Study 2 and Study 3 (as it has been done in the respective manuscripts).

After a short introduction to the common factor model and its implications for the estimation of factor loadings and factor scores as well as an introduction to the issue of factor retention, the three manuscripts are summarized and discussed.

3.2 The Common Factor Model

The common factor model assumes linear relations between each manifest variable and the k underlying latent variables. When all p manifest variables (and the latent variables) are mean-centered, the common factor model for each manifest variable x_i (with $i \in \{1, ..., p\}$) can be written as:

$$x_i = \lambda_{i1}\xi_1 + \lambda_{i2}\xi_2 + ... + \lambda_{ik}\xi_k + \epsilon_i$$

where ξ_j is the j-th latent variable (or factor, where $j \in \{1, ..., k\}$) and ϵ_i[2] is an error term consisting of measurement error and item-specific “uniqueness”. The common factor model for all p manifest variables combined can be written as:

$$\mathbf{x} = \mathbf{\Lambda}\xi + \epsilon$$

with $\mathbf{x}$ being a vector containing all p manifest variables, ξ containing the k latent factors, ϵ containing the error terms and $\mathbf{\Lambda}$ being a $p \times k$ matrix containing the model parameters λ_{ij} called factor loadings. From this, one can derive that the covariance matrix of the manifest variables $\mathbf{\Sigma} = \mathbb{E}(xx^\top)$ can be written as:

$$\mathbb{E}(xx^\top) = \mathbb{E}((\mathbf{\Lambda}\xi + \epsilon)(\mathbf{\Lambda}\xi + \epsilon)^\top) = \mathbb{E}(\mathbf{\Lambda}\xi\xi^\top\mathbf{\Lambda}^\top) + \mathbb{E}(\mathbf{\Lambda}\xi\epsilon^\top) + \mathbb{E}(\epsilon\xi^\top\mathbf{\Lambda}^\top) + \mathbb{E}(\epsilon\epsilon^\top)$$

setting $\mathbb{E}(\xi\xi^\top) = \mathbf{\Phi}$ and $\mathbb{E}(\epsilon\epsilon^\top) = \mathbf{\Psi}^2$, this expression becomes:

$$\mathbf{\Sigma} = \mathbf{\Lambda}\mathbf{\Phi}\mathbf{\Lambda}^\top + \mathbf{\Psi}^2$$

[2] Note: ϵ_i is uncorrelated with ξ_j for all j and uncorrelated with ϵ_l for $i \neq l$.

since $\mathbb{E}(\mathbf{\Lambda}\xi\epsilon^\top) = \mathbb{E}(\epsilon\xi^\top\mathbf{\Lambda}^\top) = 0$ due to the independence of ξ and ϵ. $\mathbf{\Phi}$ is a $k \times k$ matrix containing the inter-factor correlations and $\mathbf{\Psi}^2$ is a $p \times p$ diagonal matrix[3] containing the unique variances of the manifest variables.

Figure 1 displays the common factor model for three latent variables and nine observed ones. It shows that correlations among factors are allowed, whereas correlations among error terms are not.

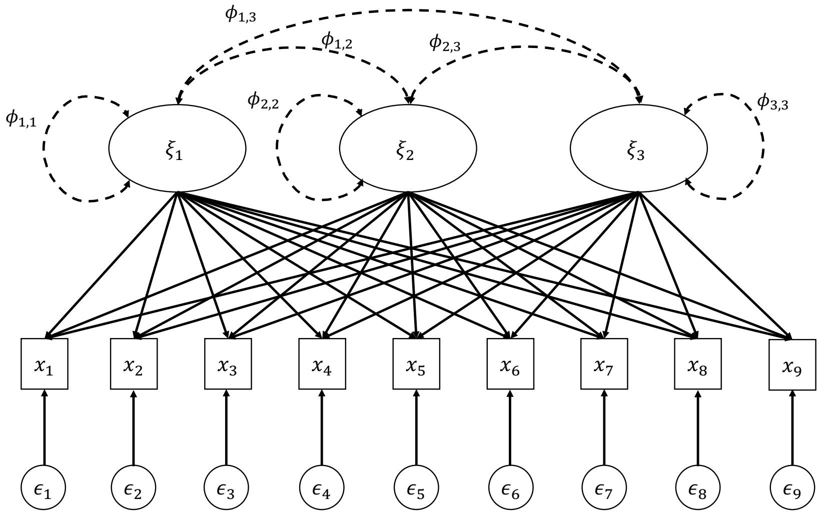

Figure 1. The common factor model with three factors and nine variables. The arrows connecting the latent variables with the manifest variables represent the respective factor loadings λ_{ij}. $\phi_{a,b}$ indicates a possible correlation between the latent factors a and b.

Introducing the error term ϵ_i and the unique variance $\mathbf{\Psi}^2$ respectively, the common factor model can be distinguished from principal component analysis (PCA) which is rather a tool of dimensionality reduction than an EFA in a narrow sense. Without the underlying measurement model, PCA does not account for measurement error and the resulting components (no latent variables per se) are positively biased compared with common factors[4], yet principal component scores are determinate unlike factor scores in the common factor model (Widaman, 2007). This so-called factor indeterminacy emerges from the fact that the common factor model contains more latent than manifest variables, so independent of

[3] $\mathbf{\Psi}^2$ is a diagonal matrix since all ϵ_i are independent of each other in the common factor model (i.e. no correlated errors are allowed in this measurement model).

[4] This bias decreases with higher communalities and smaller unique variances.

the particular sample size, infinite possible solutions are plausible for the factor scores ξ and the error terms (Steiger, 1979).

Not only determining the factor scores becomes problematic for the common factor model, but also estimating both factor loadings (λ_{ij}) and unique variances simultaneously can be challenging. A variety of extraction methods have been developed to tackle this problem with principal axis factoring (PAF, e.g. Holzinger, 1946) and Maximum-Likelihood estimation (ML, e.g. Jöreskog, 1967) being the most popular (see study 1). The objective or discrepancy functions (the functions that are minimized during the respective estimation process) reveal the differences between these two extraction methods:

$$F_{PAF} = \frac{1}{2}\,\mathrm{tr}\,[(\mathbf{S} - \mathbf{\Sigma})^2] = \sum_i \sum_j (s_{ij} - \sigma_{ij})^2$$

$$F_{ML} = log|\mathbf{\Sigma}| + \mathrm{tr}\,(\mathbf{S}\mathbf{\Sigma}^{-1}) - log|\mathbf{S}| - p \approx \sum_i \sum_j [\frac{(s_{ij} - \sigma_{ij})^2}{u_i^2 u_j^2}]$$

with $\mathbf{S}$ being the empirical correlation matrix, $\mathbf{\Sigma}$ being the model-implied correlation matrix and u_i^2 being the sample unique variance of the manifest variable i. Accordingly, for the ML approach the residuals between the empirical and model-implied correlation matrix is weighted by the sample unique variances. Hence, ML weighs down "weak" variables with low communalities (high uniqueness) as described by De Winter and Dodou (2012) or MacCallum, Browne, and Cai (2007).

Both objective functions are solved iteratively[5], yet the proceedings for PAF and ML are different. For PAF, initial communalities are estimated that replace the diagonal of the correlation matrix $\mathbf{S}$ which becomes a reduced correlation matrix $\mathbf{S}^*$ that can be decomposed via $\mathbf{S}^* = \hat{\mathbf{\Lambda}}\hat{\mathbf{\Lambda}}^\top$ with $\hat{\mathbf{\Lambda}}$ containing the estimated factor loadings. Based on these loading estimates, the communalities are re-estimated and the procedure continues with the next iteration until the communality estimates stabilize (for further readings, see De Winter & Dodou, 2012; Jöreskog, 2007). For the ML approach, however, no initial estimates of the communalities are necessary - both the unique variances and factor loadings are estimated directly in an iterative procedure. Jöreskog (1967) developed a computational feasible way to estimate the factor loadings $\mathbf{\Lambda}$ given current estimates for $\mathbf{\Psi}^2$ and vice versa.

[5] Therefore, so called Heywood cases can occur, where estimates of unique variances can become negative.

Since, both approaches are not useful for (very) small sample sizes and $n < p$ scenarios[6], new extraction methods emerged (Hirose & Yamamoto, 2014; Jung & Takane, 2008) that are based on regularized objective functions. These new methodological developments are discussed amongst others (inter alia the differences of common rotation methods that are used to obtain an interpretable pattern matrix) in the first study of this thesis.

3.3 Factor Retention

Before it is possible to estimate the model parameters (factor loadings and unique variances), researchers conducting an EFA have to determine the number of factors that should be retained. Often no theoretical assumptions can be made and the dimensionality (number of underlying factors) has to be estimated based on the empirical data set. Hence, several so-called factor retention criteria have been developed. Since Study 2 and Study 3 focus on the issue of factor retention, the following paragraph summarizes the most common criteria (see Study 1 for their proportions in the current research) and relevant new methods. The main element of the majority of these factor retention criteria are the eigenvalues of the correlation matrix of the manifest variables.

3.3.1 Eigenvalues. The symmetric $p \times p$ correlation matrix $\mathbf{S}$ (or the reduced correlation matrix based on the factor model $\mathbf{S}^*$) is characterized by p eigenvalues denoted η (the number of eigenvalues is equal to the rank of the respective matrix[7]) and p eigenvectors denoted x via the following transformation:

$$\mathbf{S}x = \eta x \; , \; x \neq 0$$

which holds for all p pairs of x and η.

Accordingly, the information about the correlations among the p manifest variables is transformed to p combinations of a scalar η and a vector x. The sum of the eigenvalues (of the correlation matrix) equals the number of manifest variables ($\sum_i^p \eta_i = p$) and the ratio $\frac{\eta_i}{\sum_i^p \eta_i} = \frac{\eta_i}{p}$ indicates the share of item variance that can be explained by the i-th linear combination of the p manifest variables. Hence, the higher the first eigenvalues become (and the higher this ratio gets) the less of these linear combinations (and therefore less latent factors) are needed to explain the variation of the observed variables. This is the

[6] $n < p$ scenarios are data conditions with less observations than manifest variables.

[7] Correlation matrices are always symmetric and positive semi-definite and therefore of full rank p.

gist of most factor retention criteria - the empirical eigenvalue distribution (or eigenvalue pattern) is used to determine the number of underlying factors.

3.3.2 Kaiser-Guttman Rule. The Kaiser-Guttman rule (KG; Kaiser, 1960) is a heuristic rule that suggests to extract all factors with eigenvalues greater than one. The idea is that an eigenvalue greater than one indicates that the respective factor explains more variance than a single manifest variable does[8] which is a reasonable argument on population level, but is flawed for empirical data due to sampling error as described by Braeken and Van Assen (2017). The weak performance (especially the tendency to retain too many factors [overfactoring]) of KG on sample level has been reported in several studies (e.g. Fabrigar et al., 1999; Velicer, Eaton, & Fava, 2000; Zwick & Velicer, 1986), yet it is still the most frequently used criterion in empirical research (see Study 1) and the default in statistical programs like *SPSS* (IBM Corp., 2019).

3.3.3 Scree Test. Another very popular method to determine the number of factors is the Scree test by Cattell (1966). This test is based on the visual inspection of a graphical representation of the empirical eigenvalue sequence. The researcher has to detect the "elbow" in a graph plotting the eigenvalues against the number of factors (see Figure 2 for an example of the so-called Scree plot). A heavy drop of the graph indicates that the factor before the drop substantially contributes, while all following factors have little further explanatory power. This idea seems to be reasonable and may be appealing to practitioners as it has high face validity, yet the interpretation can be rather difficult (as shown in Figure 2) and could lead to subjective decisions. Although there are some ideas to objectify the procedure like the Cattell-Nelson-Gorsuch approach (e.g. Nasser, Benson, & Wisenbaker, 2002) and other non-graphical methods (e.g. Raîche, Walls, Magis, Riopel, & Blais, 2013), it is not advisable to rely on these methods as they are inferior to state-of-art factor retention criteria as well (e.g. Ruscio & Roche, 2012).

3.3.4 Minimum Average Partial (MAP) Test. The MAP test by Velicer (1976) is designed for PCA, yet often applied to EFA settings as well (see Study 1). It is based on the following statistic describing the averaged squared partial correlation after x components are partialled out:

[8] In cases were all manifest variables are uncorrelated, the correlation matrix becomes the identity matrix and all eigenvalues would be equal to one.

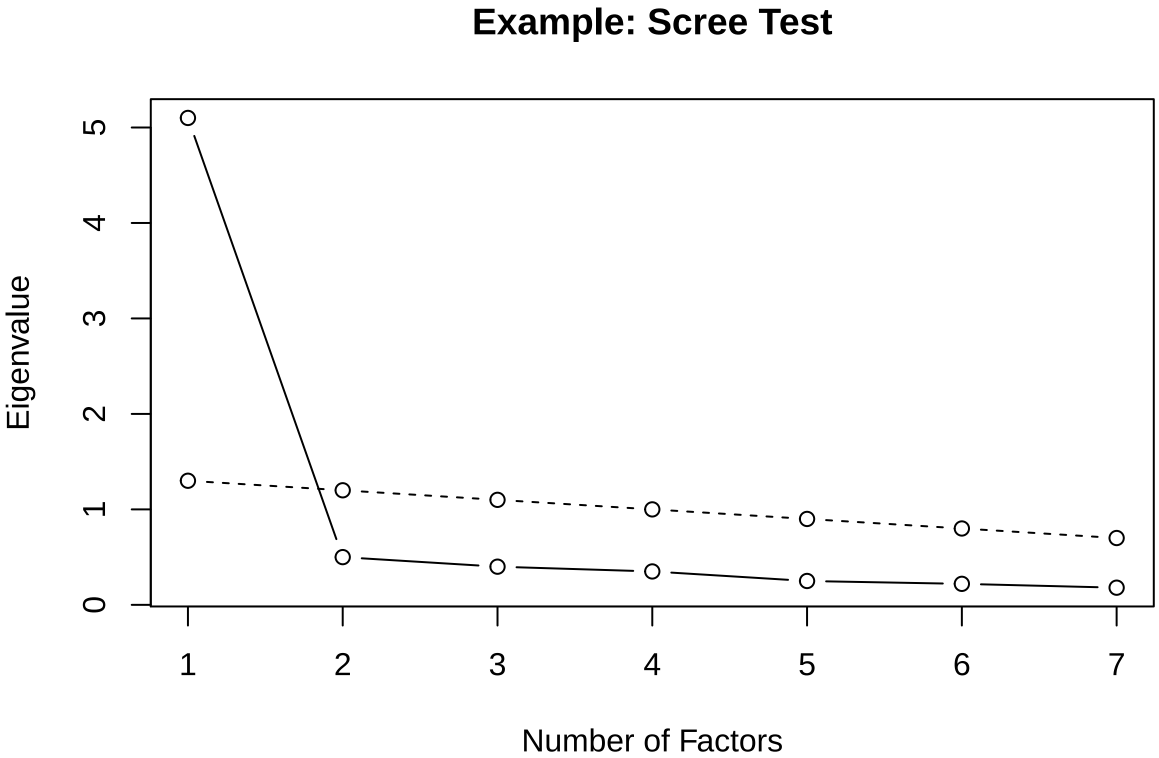

Figure 2. An exemplary Scree plot showing two cases - an unambiguous Scree plot (Scree1: solid line) where obviously one factor has to be extracted and an ambiguous Scree plot (Scree2: dashed line) where no "elbow" can be detected.

$$MAP_x = \sum_{i=1}^{p} \sum_{\substack{j=1 \\ j \neq i}}^{p} \frac{r_{xij}^2}{p(p-1)}$$

where x is the number of components that is currently assessed and r_{xij}^2 is the squared correlation among the i-th and j-th variable after x components are partialled out. The summary statistic MAP_x is calculated for $x = 0, ..., p-1$ and minimized to determine the number of components that are sufficient to explain the variation of all p variables. While the MAP test performs quite well when determining the number of components in PCA (e.g. Caron, 2019; Zwick & Velicer, 1986), it is not recommended for the common factor model (see Study 1).

3.3.5 Parallel Analysis. Parallel analysis developed by Horn (1965) compares the sequence of eigenvalues of the correlation matrix with a sequence of averaged eigenvalues from K random data sets with the same sample size as the empirical data set. While in this traditional approach the mean over K random data sets is used to compare each eigenvalue

Table 1

Traditional Parallel Analysis: Example for Simulated Data with Ten Variables based on Two Factors

Eigenvalues	PA_{mean}	$PA_{95\%}$
3.466	1.364	1.468
2.134	1.249	1.324
0.763	1.164	1.224
0.711	1.089	1.143
0.655	1.021	1.068
0.561	0.958	1.006
0.487	0.891	0.936
0.454	0.828	0.878
0.420	0.757	0.815
0.349	0.679	0.742

Note. 1000 random data sets were simulated for comparison. PA_{mean} are the mean eigenvalues based on random data and $PA_{95\%}$ are the 95 percentile eigenvalues based on random data that are used for comparison.

with, there are other implementations using the 95%-percentile instead. Both ideas are illustrated in Table 1. The empirical first eigenvalue is compared with the average of all K first eigenvalues of the simulated random data sets (here $K = 1000$) or the 95%-percentile of the respective K first eigenvalues. This is done for all p eigenvalues and factors are retained as long as the empirical eigenvalue is greater than the average or 95%-percentile of the comparison eigenvalues.

Other implementations of PA are based on eigenvalues from the reduced correlation matrix that takes the communalities into account and yet other implementations use permuted data instead of random data to preserve the skewness of the original data. These numerous types of PA vary in their performance as simulation studies show (e.g. Auerswald & Moshagen, 2019; Lim & Jahng, 2019), so practitioners should make educated decisions and report which implementation they use when selecting PA as their factor retention criterion.

3.3.6 New Approaches. Although PA has become the standard criterion that is often recommended due to its quite good performance across various conditions and its robustness against distributional assumptions (e.g. Fabrigar et al., 1999 and Study 1), there are several new approaches that are superior to PA in different conditions. The three most promising are discussed in Study 1 - the hull method (Lorenzo-Seva, Timmerman, & Kiers, 2011), the comparison data (CD) approach (Ruscio & Roche, 2012) and the empirical Kaiser criterion (EKC; Braeken & Van Assen, 2017), while both CD and EKC are used for comparison in Study 2 and Study 3.

4 Summary Study 1

In addition to planning the study design (sample size, degree of overdetermination [which is the number of variables per expected factor] and the choice of indicators), three major settings have to be chosen in EFA: the number of factors (or rather which factor retention criterion to use to determine this number), the extraction method and the rotation method. Study 1 combines a review on the current use of exploratory factor analysis regarding these decisions with a discussion of new methodological developments. It can be seen as a revision of the famous review by Fabrigar et al. (1999). For reviewing the current use of EFA, two journals focusing on psychological assessment (*Psychological Assessment* and *European Journal of Psychological Assessment*) were selected and every original article[9] from 2007 to 2017 reporting an EFA as a main statistical analysis was included in the review (304 reported EFAs).

4.1 Sample Size

More than half of the reported EFAs (50.3%) were conducted on samples with more than 400 observations (compared to 33.2% in the review of Fabrigar et al., 1999), whereas 16.4% were based on samples smaller than 200 observations (compared to 44.2% in the review of Fabrigar et al., 1999). This tendency for higher sample sizes can be seen as a sign of improving study designs in psychological research based on EFA. Since several simulation studies (e.g. Hogarty, Hines, Kromrey, Ferron, & Mumford, 2005; Mundfrom et al., 2005)

[9] 993 studies in *Psychological Assessment*, issues 19(1)-29(4) and 336 studies in the *European Journal of Psychological Assessment*, issues 23(1)-33(1) were examined.

showed the necessity for greater samples when communalities and overdetermination are low (which you cannot rule out entirely in advance), Study 1 advocates for samples with at least 400 observations. Higher sample sizes are also desirable as model parameters and factor scores are estimated with higher precision.

4.2 Extraction Methods

The majority of reviewed EFAs used PAF (51.3%) or ML estimation (16.4%), while for 22.4% of the analyses the extraction method was not reported. As no extraction method is always superior (e.g. see De Winter & Dodou, 2012 for a comparison of PAF and ML), we recommended to use a weighted least squares (WLS) approach for ordinal data with few categories (see Beauducel & Herzberg, 2006; Rhemtulla, Brosseau-Liard, & Savalei, 2012) and skewed data (Holgado–Tello, Chacón–Moscoso, Barbero–García, & Vila–Abad, 2010) and ML estimation when multivariate normality can be assumed. Both WLS and ML are implemented for confirmatory factor analyses (CFA) as well, so using these extraction methods and the respective fit indices allows for cross-validating results with CFA. PAF should be rather used in case where ML estimation produces Heywood cases, as it is less prone to such estimation problems (De Winter & Dodou, 2012). New estimation algorithms especially designed for small sample sizes (and $n < p$ scenarios) have emerged, since these conditions can be unfeasible for both PAF and ML estimation. We discussed these new approaches focusing on the proposed regularized exploratory factor analysis by Jung and Takane (2008) and the penalized EFA by Hirose and Yamamoto (2014) which is designed for wide data (many variables) and sparse loading patterns. For psychologists though, the latter seems to be more appealing as a way to replace the additional rotation step in EFA to get an interpretable solution. Using the penalized EFA instead of common EFA with subsequent rotation as discussed in Study 1 has recently been tested empirically by Scharf and Nestler (2019).

4.3 Rotation Methods

In addition to presenting the penalized EFA (Hirose & Yamamoto, 2014) as a new way to think about rotation, we focused on the common two step approach (extracting an initial factor solution and then rotating it to improve interpretability) and presented different rotation methods within the framework of the Crawford-Ferguson family (CF; Crawford & Ferguson, 1970). The general CF complexity function which is minimized with regard to constraints that are inherent to the respective rotation method covers several well-known rotation methods. We briefly discussed how these different criteria focusing either on row-complexity or column-complexity (see also Browne, 2001) provide simple structure patterns or benefit cross-loadings. In this context, additional weighting approaches were also introduced.

Besides reframing common rotation methods, we debated whether and when the rotation to a predefined target (Myers, Jin, Ahn, Celimli, & Zopluoglu, 2015) is appropriate and pointed out similarities of this approach to exploratory structure equation modelling (e.g. Marsh, Morin, Parker, & Kaur, 2014). Although, our review shows that current research practice has been improved as mainly oblique rotation methods were used (over 70% of the analyzed EFAs relied to oblique rotation) compared to the review of Fabrigar et al. (1999), where orthogonal *Varimax* rotation was applied in more than 50% of the cases, only two studies used different methods and compared the resulting patterns which is highly recommended by several authors (Browne, 2001; Fabrigar et al., 1999). Hence, in Study 1, we strongly advocated for comparing rotation methods with regard to the stability of factor patterns and the interpretability of the final solution.

4.4 Factor Retention Criteria

The most commonly used factor retention criterion was the Kaiser-Guttman rule (55.6%), followed by the Scree test (46.4%) and parallel analysis (42.1%). While parallel analysis (PA; Horn, 1965) has become a “gold-standard” for determining the number of factors, both Kaiser-Guttman (KG; Kaiser, 1960) and the Scree test (Cattell, 1966) are seen critical (e.g. Fabrigar et al., 1999). However, new promising alternatives have been developed that are superior to PA. We discussed the advantages and disadvantages of both the hull method (Lorenzo-Seva et al., 2011) and CD (Ruscio & Roche, 2012) in detail and introduced the modern version of KG - the EKC (Braeken & Van Assen, 2017). We recommended to compare the results of different retention criteria and to consider theoretical perspectives with regard to content validity for test construction purposes.

5 Summary Study 2

Besides new stand-alone factor retention criteria like CD or EKC, combination rules for several different criteria like the one proposed by Auerswald and Moshagen (2019) have been developed over the last years. Even though these rules promise high accuracies when determining the number of factors, they seem to be rather complex and therefore not very user-friendly. Furthermore, combining different criteria may reduce the confidence in the final solution. Therefore, this thesis introduces a new criterion that achieves a high accuracy, while it promises to be easily applicable for practitioners and provides probability estimates for different factor solutions that can be understood as confidence measures.

5.1 General Idea: The Factor Forest

This new approach (working title: *Factor Forest*) is based on extensive data simulation that reflects realistic data conditions of the application context and modern machine learning models that are used to predict the number of factors. The basic idea of the *Factor Forest* is that a (complex) statistical model can be found that describes the relationship between the characteristics of the empirical data set and the true number of underlying latent variables. The eigenvalues of the correlation matrix and the number of manifest variables as well as the sample size are obvious choices for the predictors of such a model. Further predictors (or features in the context of machine learning models) are developed to describe the empirical correlation matrix as EFA is based on the decomposition of the inter-item correlations.

Since modelling the relationship between these (observable) features and the (unobservable) number of factors is not feasible without knowing the true number of factors, a data basis (consisting of numerous data sets) with known factorial structure is simulated in a first step. This data basis is created varying the true number of factors (k), the sample size (N), factor loading magnitudes (primary and cross-loadings; $\mathbf{\Lambda}$), the number of variables per factor, the inter-factor correlations ($\mathbf{\Phi}$) and the number of manifest variables (p). For each of the data sets for this data basis, several features are calculated (mostly features that describe the correlation matrix, e.g. different eigenvalues and matrix norms) that are later used as predictor variables in the statistical model. Afterwards, a machine learning model - the *xgboost* algorithm (Chen & Guestrin, 2016; Chen, He, Benesty, Khotilovich, & Tang, 2018) seems to be a good choice (see Study 2) - is trained on the simulated data (depending on the algorithm several hyperparameters can be tuned to improve the predictive performance). The trained model can be evaluated on new simulated test data. In Study 2, this new approach is presented in detail (see Figure 3) and its performance is compared with common criteria.

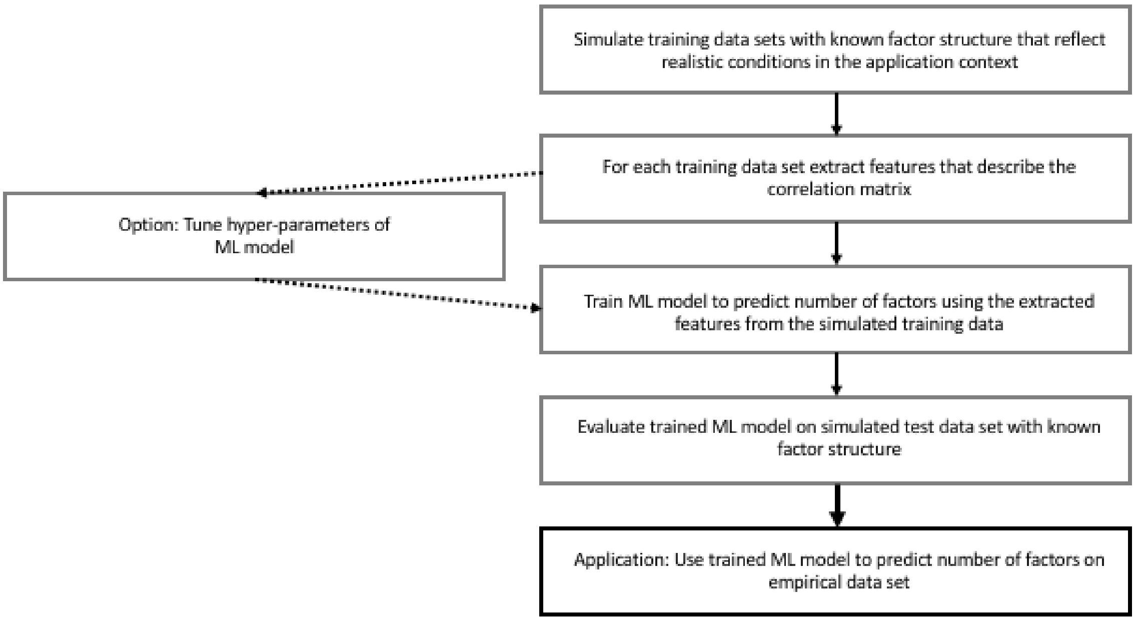
Simulate training data sets with known factor structure that reflect realistic conditions in the application context
For each training data set extract features that describe the correlation matrix
Option: Tune hyper-parameters of ML model
Train ML model to predict number of factors using the extracted features from the simulated training data
Evaluate trained ML model on simulated test data set with known factor structure
Application: Use trained ML model to predict number of factors on empirical data set

Figure 3. Visualization of the new factor retention approach (figure from Study 2).

5.2 Results

In Study 2, we first simulated 500000 data sets that served as the data basis for the machine learning model and extracted 181 different features that were presumed to have predictive power for the true number of factors. We then trained three different machine learning algorithms (mainly with default parameter settings) on this data and evaluated their performance compared to PA, EKC, CD and KG on new simulated data. For the evaluation, 3204 data conditions[10] were created varying the true number of factors (k), the sample sizes (N), factor loading magnitudes (primary and cross-loadings; $\mathbf{\Lambda}$), the number of variables per factor, the inter-factor correlations ($\mathbf{\Phi}$) and the number of manifest variables (p) as it was done for the simulation of the data basis.

The overall accuracy (out-of-sample) of the trained *xgboost* model was higher (92.9%) than the accuracy of all common retention criteria. When tuning six of the hyperparameters of the *xgboost* algorithm (and adding the other criteria as features) this performance could be further increased in a second step (99.3% out-of-sample accuracy). While all common criteria showed some kind of bias (either for one factor solutions or when the number of factors were higher), the *xgboost* model yielded unbiased estimates for all true values

[10] Using 500 replications per condition, in total this yielded 1512000 simulated data sets for the evaluation step.

of k (we evaluated $k \in \{1, 2, 4, 6\}$). As the provided machine learning model is a black box model, variable importance measures (e.g. standard permutation based importance measure) and tools for interpreting the final model were introduced in Study 2 as well. The first eigenvalues (not the primary eigenvalue though) were among the most important variables in the *xgboost* model, but the two most important features were two inequality measures that were applied to the empirical correlation matrices - the Gini coefficient (Gini, 1921) and the Kolm measure (Kolm, 1999). We also applied the so-called local interpretable model-agnostic explanations (LIME; Ribeiro, Singh, & Guestrin, 2016) that can help to understand how the tuned model comes up with a particular prediction, although it should be carefully interpreted as it relies on local linear approximations of the far more complex model (in our example: $r^2 = 0.235$ for the explaining model).

In addition to its higher accuracy compared to common retention criteria, the *xgboost* model can provide probability estimates for different factor solutions that can serve as confidence measures for practitioners. Since such uncertainty measures are not available with the common criteria - a new approach to assess the robustness of factor retention solutions is presented in Study 3.

5.3 Additional Analyses: Ordinal Factor Forest

In Study 2, the new approach is built on multivariate normal data (which is an assumption often made in the context of EFA, e.g. for ML estimation), yet psychological research data are often collected via questionnaires and therefore of ordinal nature. Accordingly, it was necessary to also create an ordinal implementation of the *Factor Forest*. The procedure was equal to the multivariate normal case (as described in Study 2) - first data was simulated (here: ordinal data[11] based on Gaussian copulas and binomial marginal distributions with varying numbers of item categories between four and seven and $\pi \in [0.2; 0.8]$) and then the tuned *xgboost* model was trained on the extracted features (using the same 184 features including the PA, CD and EKC solution as well as the number of categories as a special feature for the ordinal data). The performance of the new approach was just slightly worse for ordinal data (overall out-of-sample accuracy of 98.5%) than for normal data as reported in Study 2. The accuracy for different values of k varied between 97.9%

[11] 486563 data sets (initially 500000, but data sets with improper correlation matrices or errors in calculating features were excluded) were simulated and randomly assigned to the training or test set (70%/30% of the data sets).

($k = 3$) and 99.3% ($k = 1$), so the *Factor Forest* showed promising results as a stand-alone retention criterion for ordinal data as well.

6 Summary of Study 3

While the main aim of many factor retention criteria is to approximate the data generating process as close as possible (i.e. finding the "true" number of factors), replicability should not be ignored (Osborne & Fitzpatrick, 2012; Preacher, Zhang, Kim, & Mels, 2013). In practice, when there is only one empirical data set (and splitting this data set is not an option due to a small sample size), it is difficult to predict whether the assumed number of factors (based on one or several retention criteria) is robust against sampling error and reproducible later on. Study 3, therefore, evaluates an approach to assess the robustness of factor retention solutions and its usefulness as a proxy for possible replicability.

6.1 Results

Since replicability has become an issue in psychological research, researchers conducting an EFA should focus especially on the factor retention process as determining the number of factors may be the most far-reaching decision with regard to a successful replication of the factorial structure. Common retention criteria only provide estimates of the dimensionality, but no confidence measures (or measures of uncertainties like standard errors). When samples are small (and they often are as Study 1 shows) the accuracy of the criteria decreases (e.g. Study 2; Auerswald & Moshagen, 2019) as the retention process is prone to sampling error[12]. This hampers replicability and practitioners cannot decide how robust an estimate for the number of factors is.

Therefore, Study 3 investigated the robustness of different factor retention criteria on empirical data via bootstrapping and examined whether the robustness across bootstrap samples can be used as a proxy for reproducibility. The new *xgboost* model (the *Factor Forest*, see Study 2), PA, CD and EKC were compared with regard to their robustness and their reproducibility using 19 data sets consisting of personality measures (the *10 Item Big Five Inventory* by Rammstedt et al., 2017 and the *Big Five Structure Inventory* by Arendasy,

[12] This vulnerability to sampling error can be illustrated using KG: Given, for example, an eigenvalue of 1.05 on population level, values below one on sample level are very likely - especially when the sample size is small. So the Kaiser-Guttman criterion will retain less than the true number of factors just by chance very often.

2009). These data sets contain either different measurements (four time points) of the same participants or different cohorts within one study project. Accordingly, two different types of replication contexts are covered - within-subject and between-subject replication studies.

The results indicate that the *xgboost* model and EKC are more robust against sampling variations[13] than PA and CD and tend to reproduce the number of factors more often (an exact replication of the number of factors for two consecutive measurement periods) and more accurately (mean absolute difference of the suggested number of factors for two consecutive measurement periods). The *xgboost* model had the highest rate of replicability (61.5%), while using PA only 7.7% of the cases were exactly replicated. Although, EKC showed more robustness across the bootstrap samples than the *xgboost* model, its replication rate was only the second best (46.2%) and the mean absolute deviation of replication attempts was higher compared with the *xgboost* model (0.615 to 0.385). However, a positive relation between robustness across the bootstrap samples and reproducibility was found - as both robustness measures were positively associated with number of exact replications and negatively associated with the deviation from two consecutive dimensionality estimates. The respective results of our GLM analyses - even though not interpretable from a significance testing perspective - underline the relation between the robustness of factor solutions and their replicability.

7 Discussion

This thesis contains three studies that cover different aspects of EFA, but predominantly focuses on the issue of factor retention. Study 1 shows that current research often relies on factor retention criteria that are either not designed for the common factor model, rather subjective or not accurate on sample level. The latter is the case for the Kaiser-Guttman rule which is meaningful on population level[14], but prone to sampling error and therefore not useful as a (stand-alone) retention criterion for empirical data. Since all common (and new developed) factor retention criteria lack accuracy under some data conditions, reviews like Study 1 and large-scale simulation studies (e.g. Auerswald & Moshagen, 2019) urge practitioners to compare different criteria and to combine various estimates. This ap-

[13] The robustness was indicated by the standard deviation across 100 bootstrap samples and the percentage of bootstrap samples for which the criterion suggested the same number of factors as it did on the whole data set.

[14] Braeken and Van Assen (2017) describe why the eigenvalue $>$ 1 rule, that is associated with a positive Kuder-Richardson reliability, is a lower bound for the relevance of empirical eigenvalues.

proach is not new as Fabrigar et al. (1999) already suggested to compare different methods two decades ago. However, the results of Study 1 demonstrate that there is a discrepancy between methodological knowledge (presented in tutorial papers and how-to guidelines) and actual research practice. Although comparing criteria and using PA rather than, for example, KG or the Scree test is recommended, various studies were based on inappropriate methodological decisions considering the factor retention process. Hence, new approaches to this issue have to be both accurate for a great range of data conditions and easy to use, so that they will be applied by the majority of researchers conducting EFAs.

The new approach presented in Study 2 - the *Factor Forest* - promises to tackle both of these requirements. The tuned *xgboost* model showed almost perfect accuracy for multivariate normal data (Study 2) and is also applicable to ordinal data as the additional analyses reported above demonstrated. So far, the new approach combining extensive data simulation and the application of complex machine learning algorithms requires a lot of computational resources and is not yet easy to use. The final trained model can be used easily though and therefore seems to be a promising alternative for the actual research practice. Providing such a model that reflects all necessary data conditions would help practitioners with a highly accurate, objective and task-related method to determine the correct number of factors for their analyses. Accordingly, the *Factor Forest* covering different trained models for different types of data should be made accessible for EFA users in the future.

As discussed in Study 2, the supervised[15] learning approach (for a comparison of supervised and unsupervised learning, see James, Witten, Hastie, & Tibshirani, 2013) requires the simulation of a data basis that includes all important data conditions. So the trained model based on data sets that are simulated for $k \in \{1, 2, ..., 8\}$ (see Study 2) is able to suggest only one- to eight-factor solutions by design. Accordingly, the *Factor Forest* is only applicable to data sets that are somewhat similar to those it has been trained on. Study 2 shows that the new approach performs well in data conditions that are close to those in the training set, but not included. Nevertheless, for completely new conditions (e.g. panel data with considerably more manifest variables) new training data has to be simulated and

[15] Supervised learning - in contrast to unsupervised learning - requires a criterion for which true values are known. Here, a simulated data basis with known factor structure (known true number of factors) is necessary to create the machine learning (or statistical) model that can predict the number of factors based on data set characteristics. EFA itself (or the related PCA) can be assigned to unsupervised learning like clustering (e.g. Hansen & Larsen, 1996) where no target variable with known values is needed.

a new model has to be trained.

In Study 2, machine learning algorithms had to be chosen for the *Factor Forest* implementation. The *xgboost* algorithm (Chen & Guestrin, 2016) - especially when tuning six of its hyperparameters - showed the best results and nearly perfect accuracy, yet implementing other algorithms in further versions would be possible as well. Using other tree-based methods (e.g. random forest implementations like the *ranger* by Wright & Ziegler, 2017 as evaluated in Study 2 or the *cforest* by Hothorn & Zeileis, 2015 which is an implementation of a conditional random forest that promises unbiased partitioning) or kernel-based methods like support vector machines (Cortes & Vapnik, 1995), while relying on the same features as predictors will probably not be superior to the tuned *xgboost* model. However, selecting other/further features could help to improve the accuracy as the addition of the common retention criteria EKC, PA and CD as features in Study 2 demonstrated. The current implementation of the *Factor Forest* could be extended by features specially designed for new data conditions (like it was done for ordinal data where the number of categories served as an additional feature).

Another idea would be to use the correlation matrix itself as the input instead of creating features that describe the correlation matrix mathematically and use those features as the input for the model. This could be done applying neural networks (e.g. Cheng & Titterington, 1994) since they can handle large numbers of input variables (in this case a correlation matrix of 80 variables has $80^2 = 6400$ entries and consists of $\frac{80 \times 79}{2} = 3160$ unique bivariate correlations). The numerical value of each correlation could be directly used as the input for one input node and no feature engineering would be necessary (contrary to an implementation where a graphical representation of the correlation matrix is used as the model input and the neural network is used for image processing, see e.g. Egmont-Petersen, Ridder, & Handels, 2002). The ordering of the variables would be a severe problem for this approach, though, since the structure of a correlation matrix is completely arbitrary and the position of an item X is only determined by its position in the data set. To solve this problem, the variables could be clustered in a first step to create a correlation matrix that shows concentrations of high and low correlations or the assignment of bivariate correlations to the input nodes could be randomized and repeated several times to cover the various possible structures that can occur for the same data (this would lead to a much bigger training set).

All features extracted for Study 2, on the contrary, are independent of the positioning of the manifest variables. Therefore, this procedure seems to be meaningful and worth pursuing in this context. Nevertheless, further research could evaluate whether the use of neural networks (using the same or an extended feature set) might be beneficial - especially with regard to an integration of different data types (ordinal data, multivariate normal data, count data, etc.) into one model.

As mentioned above, further features might improve the accuracy of the new approach for the data conditions that were assessed in Study 2, for ordinal data and for other data types that have not been evaluated yet. Although the features that describe the eigenvalue patterns seem to be predictive for all types of data, additional features could be useful for specific data conditions. In contexts where the variable distributions are highly skewed and the assumption of multivariate normality has to be questioned, features describing the joint item distribution or the marginals could be helpful. As EFA is often conducted on questionnaire data and this kind of data is prone to missing values due to item non-response (e.g. Shoemaker, Eichholz, & Skewes, 2002), missingness should also be considered during the factor retention process (for more details on the impact of missing data on factor retention, see, for example, Goretzko, Heumann, & Bühner, 2019). Thus, the proportion of missing values for each variable or the type of missing data method used to handle missingness (the default is often pairwise-deletion when calculating correlation matrices) could provide valuable information for the prediction task.

Depending on the context it would also be possible to change the loss function[16] (the performance measure) which is used for the training (and evaluation) of the machine learning model. In Study 2, predicting the number of factors is framed as a classification task for which several performance measures have been developed. Ferri, Hernández-Orallo, and Modroiu (2009) compare the behavior of 18 measures for classification tasks and show which measures provide similar results and which measures differ strongly. The authors also discuss which measures are independent of prior class distributions (e.g. when the classes are highly imbalanced in the training data a classifier predicting the majority class achieves a high accuracy, yet poor values for recall or precision which correspond to sensitivity and $1-$specificity in the psychological research context) which is not a problem in Study 2, but could be of interest when data conditions are not evenly distributed in the training set. Since

[16] The loss function describes how well the model fits the data and is therefore minimized when training the model. In the classical regression context, the quadratic loss function (see least-squares approach) or the mean squared error (*MSE*) is typically used as the loss function, for example. The loss functions weigh the different aspects of model misfit differently (e.g. outlier sensitivity of quadratic loss vs. absolute loss), so changing the loss function can benefit another candidate model (or leads to different model parameters in the context of model training).

the *Factor Forest* is currently optimized with regard to the overall accuracy (that means that the loss function weighs every false prediction equally - no matter if the prediction is $\hat{k} = 3$ or $\hat{k} = 5$ when the actual number of factors is $k = 2$), it could be meaningful to adjust the loss function to take the ordinality[17] of the criterion into account. One possibility would be to define the loss as a symmetric Toeplitz matrix (e.g. Gray, 2006) with a zero diagonal, for example:

$$\mathbf{L} = \begin{bmatrix} 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 \\ 1 & 0 & 1 & 2 & 3 & 4 & 5 & 6 \\ 2 & 1 & 0 & 1 & 2 & 3 & 4 & 5 \\ 3 & 2 & 1 & 0 & 1 & 2 & 3 & 4 \\ 4 & 3 & 2 & 1 & 0 & 1 & 2 & 3 \\ 5 & 4 & 3 & 2 & 1 & 0 & 1 & 2 \\ 6 & 5 & 4 & 3 & 2 & 1 & 0 & 1 \\ 7 & 6 & 5 & 4 & 3 & 2 & 1 & 0 \end{bmatrix}$$

which can be read like a confusion matrix with the rows indicating the true values and the columns indicating the predictions while all elements of **L** quantify the respective loss.

This loss matrix punishes higher deviations from the true number of factors more strongly than the loss function behind the accuracy does, so further research should evaluate whether the nearly perfect overall accuracy of the model can also be reached when considering the ordinality of the criterion. This seems to be an important research question, because in practice the general model accuracy is not relevant when the prediction for a particular empirical data set heavily diverges from the true number of factors. In fact, the smaller this difference is the more useful the respective research results are.

There are several authors (e.g. Fabrigar et al., 1999; Thurstone, 1947) that regard

[17] Predicting the number of factors was implemented as a classification task since the number of factors has to be an integer. However, it would also be possible to implement it as a regression task instead. In this case, the ordinality of the criterion would be considered by a quadratic loss function associated with the *MSE*, for example. In doing so, estimates like 3.48 can occur, for which common rounding strategies yield three suggested factors. As underfactoring is considered worse compared with overfactoring (e.g. Fabrigar et al., 1999) this might be not the best strategy for this application. In addition, implementing the *Factor Forest* as a regression task could result in implausible solutions like a negative number of factors (e.g. when using regularized linear models as the statistical model, no problem for regression trees or our xgboost implementation though), so relying on the classification framework seems to be more promising.

overfactoring as less severe compared to underfactoring (as pseudo relations between factors and manifest variables that would load on other factors that were not included in the model distort the results). Accordingly, the loss function could be adjusted by increasing the costs of underfactoring or decreasing the costs of overfactoring. This might help to avoid underfactoring, but as Study 2 showed that the *Factor Forest* provided unbiased results for all values of k, the new model with the adjusted loss function would probably suggest too many factors in many cases.

While the possibility to change the loss function or the machine learning algorithm, to tune more hyperparameters, to extract further features and to extend the data basis makes the *Factor Forest* approach very flexible and versatile, the new method has another advantage over common retention criteria - the model does not only provide point estimates for the number of factors, but also probability estimates that can be used as an implicit certainty measure. When for example, the trained model suggests two factors with an probability estimate of 44% (while a three factor solution has a probability of 38%), the user may double check the results calculating the EKC or conducting PA. However, when the estimated probability of the two factor solution is 99.97%, the user can have more confidence in that result.

Study 3 introduces bootstrapping to the factor retention process as an easy-to-use tool to evaluate the robustness of the criteria used to determine the number of factors. Furthermore, it illustrated the positive relation between robustness across bootstrap samples and replicability as well as the comparably good performance of the *Factor Forest* in terms of reproducible results. Thus, the new approach serves both goals of factor retention - the accurate approximation of the data generating process and the replicability in different samples (see, Preacher et al., 2013). The latter is particularly important considering debates about the replication crisis in psychology (e.g. Shrout & Rodgers, 2018).

As described in Study 3, debates about the replication crisis mainly focus on hypothesis testing, under-powered studies and phenomena like p-hacking (e.g. Simmons, Nelson, & Simonsohn, 2011), but completely leave out EFA. Although EFA is a purely exploratory analysis and results should always be confirmed by CFA (Fabrigar et al., 1999), results of EFAs - especially results of the dimensionality assessment - strongly shape further research (inter alia which candidate models are tested in CFA). This can be observed in the disagreements about the factorial structure of different psychological constructs. While for the most prominent example, the *BIG-5* (see Study 3), the research community predominantly[18] agrees on five factors, there are several on-going debates in other research areas, for example, debates about the dimensionality of the *Rosenberg Self-Esteem Scale* that was intended as a unidimensional measure, yet appears to rather have a two-factor structure (Huang & Dong, 2012; Marsh, Scalas, & Nagengast, 2010). In some cases, methodological artifacts like factors created by special item wording (e.g. the *Sense of Community Index*, Peterson, Speer, & Hughey, 2006) can explain these debates, but often the lack of reproducibility (and robustness against sampling error) of the factor retention criteria might be an explanation as Study 3 shows. Therefore, determining the number of factors should not be done by solely focusing on finding the "true" number of factors, but also considering the stability of solutions and their replicability.

With regard to replicability, the frequent use of PAF as extraction method (see Study 1) has to be seen critically since this extraction method does not allow for exact cross-validation with CFA. As also discussed in Study 1, researchers should be more transparent when reporting EFA results to facilitate replication attempts. Accordingly, the current research practice in EFA can be improved in terms of replicability in several ways - there should be guidelines for transparent and comprehensive reporting of decisions (extraction method, rotation method, factor retention criterion, etc.) and results of EFAs like the ones proposed by the *Journal of the Society for Social Work and Research* (Cabrera-Nguyen, 2010). For factor extraction, PAF should be avoided in favor of ML-EFA or WLS approaches, rotation methods should be compared and selected in consideration of theoretical assumptions (see Study 1) and factor retention criteria should be supplemented by a procedure to assess their stability like bootstrapping (Study 3) or evaluated on subsamples (Study 1). The new approach - the *Factor Forest* - seems to be a promising factor retention criterion considering both its accuracy (Study 2) and its reproducibility (Study 3) and may bridge the gap between methodological research and the actual practice due to its easy application (when a trained model is provided).

7.1 Limitations

In Study 1, the current use of EFA is evaluated focusing on articles published in two journals over a time span of approximately ten years. When discussing the current practice both the journal selection and the time period have to be taken into account. The selected journals focus on psychological assessment and questionnaire development,

[18] However, there are opposing findings and diverging opinions for the *BIG-5* as well (e.g. Saggino, 2000; Thalmayer, Saucier, & Eigenhuis, 2011).

which arguably attracts rather methodological papers with probably more elaborate decision making concerning EFA than journals from other (applied) research areas. Gaskin and Happell (2014), for example, evaluated the use of EFA in nursing journals in the year 2012 and reported considerably higher rates of EFAs based on orthogonal *Varimax* rotation and less usage of parallel analysis. As Study 1 considered a comparably broad time period, it would also be interesting to take a closer look at the more recent practice analyzing possible changes in response to the replication crisis (e.g. Shrout & Rodgers, 2018) that raised awareness for the importance of profound research practices.

Although the new approach to factor retention proposed in Study 2 seems to be promising, some limitations have to be mentioned. As discussed above, the *Factor Forest* needs to be trained on simulated data that reflect all data conditions within the application context. Even though Study 2 showed that minor deviations from the data conditions included in the training set did not deteriorate its performance, the *Factor Forest* might not be usable for new data that deviate strongly from the training data. Therefore, different application contexts have to be covered by separate trained models. Since the simulation of the respective training data and the model training (plus the tuning of the hyperparameters) is computationally expensive, it is not practical that every researcher has to go through the complete procedure all by him or herself when conducting an EFA. Therefore, trained models should be provided clearly stating the data conditions they were trained on, so that practitioners can select a model that suits their application context best. However, the simulated data cannot cover all potential data conditions due to economical reasons. So when working with rare data conditions that are not fully covered by the trained model, it might be necessary to evaluate the performance of the model on exemplary simulated data first. However, for most of the cases in which EFA is used, it should be possible to create a standard model with the new approach that estimates the number of factors very accurately.

As Study 3 demonstrates the link between the robustness of factor retention criteria and their replicability, it has to be stated that replicability has no value in itself. Replicating a wrong factor structure is obviously not desirable. Therefore, the robustness of the factorial structure should not be used as a stand-alone criterion when choosing a method to determine the number of factors. Nonetheless, in combination with Study 2 or other simulation studies comparing different factor retention criteria with regard to their accuracy, the robustness assessed for an empirical data set can provide valuable information on which criterion might be more reliable in a specific context. Although the relation between robustness and replicability of factor retention criteria seems to be logical and the results of Study 3 in fact gave evidence for this link, the number of observations (data sets that were investigated) for the statistical modelling was quite small, so that statistical inference should be avoided due to the lack of power. Thus, Study 3 has to be regarded as solely descriptive and further replication studies should be used to increase the number of cases for the analyses to confirm the positive relation between robustness and replicability in factor retention.

7.2 Conclusion

Even though the research practice concerning EFA seems to have improved since the review of Fabrigar et al. (1999), Study 1 showed that some methodological decisions and their reporting can still be enhanced. Especially, when it comes to determining the number of factors - arguably the most important and far-reaching decision during EFA, many researchers rely on outdated or even invalid methods. The new approach presented in Study 2 promises to be more accurate and robust (Study 3) than common criteria. Therefore, further research should focus on improving the *Factor Forest* and to provide an implementation for practitioners. Different data conditions call for different models (one for multivariate normal data, one for ordinal data, one for panel data, etc.). However, Study 2 and the additional implementation for ordinal data suggest that the new approach can be applied in varying contexts. The replication crisis and on-going debates in several areas of psychological research demonstrate the importance of a robust and reproducible factor retention process. Accordingly, Study 3 introduced the idea of bootstrapping to assess the robustness of factor retention solutions and found a relation between the robustness (across bootstrap samples) and actual replication. Thus, a robustness check like this should always be used to decide whether an assumed factor solution is stable and will be replicable or whether it is likely that sampling error has deteriorated the factor retention process.

The new factor retention criterion proposed in this thesis also demonstrates that modern machine learning techniques can be used to improve classical statistical procedures. In particular, when heuristic rules are used to select parameters of statistical methods (here: the number of factors for EFA), because no analytic solutions are feasible (or corrupted by sampling error), the idea of training complex machine learning models on simulated data can lead to more accurate heuristics or decision rules. Consequently, this new approach could potentially be adopted in other areas of classical statistics where heuristic rules are used - for example in sample size planning, outlier detection or as a model test in structural equation modeling.

8 References

Arendasy, M. (2009). *BFSI: Big-Five Struktur-Inventar (Test & Manual).* Mödling: SCHUHFRIED GmbH.

Auerswald, M., & Moshagen, M. (2019). How to determine the number of factors to retain in exploratory factor analysis: A comparison of extraction methods under realistic conditions. *Psychological Methods*, *24*(4), 468–491. doi:10.1037/met0000200

Barendse, M. T., Oort, F. J., & Timmerman, M. E. (2015). Using exploratory factor analysis to determine the dimensionality of discrete responses. *Structural Equation Modeling: A Multidisciplinary Journal*, *22*(1), 87–101. doi:10.1080/10705511.2014.934850

Bartholomew, D. J. (1995). Spearman and the origin and development of factor analysis. *British Journal of Mathematical and Statistical Psychology*, *48*(2), 211–220. doi:10.1111/j.2044-8317.1995.tb01060.x

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

Braeken, J., & Van Assen, M. A. (2017). An empirical Kaiser criterion. *Psychological Methods*, *22*(3), 450–466. doi:10.1037/met0000074

Browne, M. W. (2001). An overview of analytic rotation in exploratory factor analysis. *Multivariate Behavioral Research*, *36*(1), 111–150. doi:10.1207/S15327906MBR3601_05

Cabrera-Nguyen, P. (2010). Author guidelines for reporting scale development and validation results in the Journal of the Society for Social Work and Research. *Journal of the Society for Social Work and Research*, *1*(2), 99–103. Retrieved from http://www.jstor.org/stable/10.5243/jsswr.2010.8

Caron, P.-O. (2019). Minimum average partial correlation and parallel analysis: The influence of oblique structures. *Communications in Statistics-Simulation and Computation*, *48*(7), 2110–2117. doi:10.1080/03610918.2018.1433843

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

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In

Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785–794). New York, NY, USA: ACM. doi:10.1145/2939672.2939785

Chen, T., He, T., Benesty, M., Khotilovich, V., & Tang, Y. (2018). Xgboost: Extreme gradient boosting. R package version 0.6. 4.1.

Cheng, B., & Titterington, D. M. (1994). Neural networks: A review from a statistical perspective. *Statistical Science*, *9*(1), 2–30. Retrieved from http://www.jstor.org/stable/2246275

Cohen, J. (1957). A factor-analytically based rationale for the Wechsler Adult Intelligence Scale. *Journal of Consulting Psychology*, *21*(6), 451–457. doi:10.1037/h0044203

Comrey, A. L. (1957). A factor analysis of items on the MMPI Depression Scale. *Educational and Psychological Measurement*, *17*(4), 578–585. doi:10.1177/001316445701700412

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273–297. doi:10.1007/BF00994018

Crawford, C. B., & Ferguson, G. A. (1970). A general rotation criterion and its use in orthogonal rotation. *Psychometrika*, *35*(3), 321–332. doi:10.1007/BF02310792

De Winter, J. C., & Dodou, D. (2012). Factor recovery by principal axis factoring and maximum likelihood factor analysis as a function of factor pattern and sample size. *Journal of Applied Statistics*, *39*(4), 695–710. doi:10.1080/02664763.2011.610445

Dinno, A. (2009). Exploring the sensitivity of Horn's parallel analysis to the distributional form of random data. *Multivariate Behavioral Research*, *44*(3), 362–388. doi:10.1080/00273170902938969

Egmont-Petersen, M., Ridder, D. de, & Handels, H. (2002). Image processing with neural networks - a review. *Pattern Recognition*, *35*(10), 2279–2301. doi:https://doi.org/10.1016/S0031-3203(01)00178-9

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

Ferri, C., Hernández-Orallo, J., & Modroiu, R. (2009). An experimental comparison of performance measures for classification. *Pattern Recognition Letters*, *30*(1), 27–38. doi:https://doi.org/10.1016/j.patrec.2008.08.010

Gaskin, C. J., & Happell, B. (2014). On exploratory factor analysis: A review of recent evidence, an assessment of current practice, and recommendations for future use. *International Journal of Nursing Studies*, *51*(3), 511–521. doi:10.1016/j.ijnurstu.2013.10.005

Gini, C. (1921). Measurement of inequality of incomes. *The Economic Journal*, *31*(121), 124–126.

Goretzko, D., Heumann, C., & Bühner, M. (2019). Investigating parallel analysis in the context of missing data: A simulation study comparing six missing data methods. *Educational and Psychological Measurement*. doi:10.1177/0013164419893413

Gray, R. M. (2006). Toeplitz and circulant matrices: A review. *Foundations and Trends in Communications and Information Theory*, *2*(3), 155–239. doi:10.1561/0100000006

Haig, B. D. (2005). Exploratory factor analysis, theory generation, and scientific method. *Multivariate Behavioral Research*, *40*(3), 303–329. doi:10.1207/s15327906mbr4003_2

Hansen, L. K., & Larsen, J. (1996). Unsupervised learning and generalization. In *Proceedings of international conference on neural networks (icnn'96)* (Vol. 1, pp. 25–30). doi:10.1109/ICNN.1996.548861

Hirose, K., & Yamamoto, M. (2014). Estimation of an oblique structure via penalized likelihood factor analysis. *Computational Statistics & Data Analysis*, *79*, 120–132. doi:10.1016/j.csda.2014.05.011

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

Holgado–Tello, F. P., Chacón–Moscoso, S., Barbero–García, I., & Vila–Abad, E. (2010). Polychoric versus pearson correlations in exploratory and confirmatory factor analysis of ordinal variables. *Quality & Quantity*, *44*(1), 153–166. doi:10.1007/s11135-008-9190-y

Holzinger, K. J. (1946). A comparison of the principal-axis and centroid factor. *Journal of*

Educational Psychology, *37*(8), 449–472. doi:10.1037/h0056539

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, *30*(2), 179–185. doi:10.1007/BF02289447

Hothorn, T., & Zeileis, A. (2015). partykit: A modular toolkit for recursive partytioning in R. *Journal of Machine Learning Research*, *16*, 3905–3909. Retrieved from http://jmlr.org/papers/v16/hothorn15a.html

Huang, C., & Dong, N. (2012). Factor structures of the Rosenberg Self-Esteem Scale. *European Journal of Psychological Assessment*, *28*(2), 132–138. doi:10.1027/1015-5759/a000101

IBM Corp. (2019). *IBM SPSS Statistics for Windows*. Armonk, NY: IBM Corp. Retrieved from www.ibm.com/SPSS/Statistics

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning*. New York, NY: Springer.

John, O. P., & Srivastava, S. (1999). The big five trait taxonomy: History, measurement, and theoretical perspectives. In L. A. Pervin & O. P. John (Eds.), *Handbook of personality: Theory and research*. New York: Guilford Press.

Jöreskog, K. G. (1967). Some contributions to maximum likelihood factor analysis. *Psychometrika*, *32*(4), 443–482. doi:10.1007/BF02289658

Jöreskog, K. G. (1971). Statistical analysis of sets of congeneric tests. *Psychometrika*, *36*(2), 109–133. doi:10.1007/BF02291393

Jöreskog, K. G. (2007). Factor analysis and its extensions. In R. Cudeck & R. C. MacCallum (Eds.), *Factor analysis at 100* (pp. 47–77). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Jung, S., & Takane, Y. (2008). Regularized common factor analysis. In *New trends in psychometrics* (pp. 141–149). Tokyo: Universal Academy Press.

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

Kolm, S.-C. (1999). The rational foundations of income inequality measurement. In *Handbook of income inequality measurement* (pp. 19–100). Dordrecht, NL: Springer.

Lim, S., & Jahng, S. (2019). Determining the number of factors using parallel analysis and its recent variants. *Psychological Methods*, *24*(4), 452–467. doi:10.1037/met0000230

Lorenzo-Seva, U., Timmerman, M. E., & Kiers, H. A. L. (2011). The hull method for selecting the number of common factors. *Multivariate Behavioral Research*, *46*(2), 340–364. doi:10.1080/00273171.2011.564527

MacCallum, R. C., Browne, M. W., & Cai, L. (2007). Factor analysis models as approximations. In R. Cudeck & R. C. MacCallum (Eds.), *Factor analysis at 100* (pp. 153–175). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

MacCallum, R. C., Widaman, K. F., Zhang, S., & Hong, S. (1999). Sample size in factor analysis. *Psychological Methods*, *4*(1), 84–89. doi:10.1037/1082-989X.4.1.84

Marsh, H. W., Barnes, J., & Hocevar, D. (1985). Self-other agreement on multidimensional self-concept ratings: Factor analysis and multitrait-multimethod analysis. *Journal of Personality and Social Psychology*, *49*(5), 1360–1377. doi:10.1037/0022-3514.49.5.1360

Marsh, H. W., Morin, A. J., Parker, P. D., & Kaur, G. (2014). Exploratory structural equation modeling: An integration of the best features of exploratory and confirmatory factor analysis. *Annual Review of Clinical Psychology*, *10*(1), 85–110. doi:10.1146/annurev-clinpsy-032813-153700

Marsh, H. W., Scalas, L. F., & Nagengast, B. (2010). Longitudinal tests of competing factor structures for the Rosenberg Self-Esteem Scale: Traits, ephemeral artifacts, and stable response styles. *Psychological Assessment*, *22*(2), 366–381. doi:10.1037/a0019225

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. doi:10.1207/s15327574ijt0502_4

Myers, N. D., Jin, Y., Ahn, S., Celimli, S., & Zopluoglu, C. (2015). Rotation to a partially specified target matrix in exploratory factor analysis in practice. *Behavior Research Methods*, *47*(2), 494–505. doi:10.3758/s13428-014-0486-7

Nasser, F., Benson, J., & Wisenbaker, J. (2002). The performance of regression-based variations of the visual scree for determining the number of common factors. *Educational and Psychological Measurement*, *62*(3), 397–419. doi:10.1177/00164402062003001

Osborne, J. W., & Fitzpatrick, D. C. (2012). Replication analysis in exploratory factor analysis: What it is and why it makes your analysis better. *Practical Assessment, Research & Evaluation*, *17*(14/15), 1–8.

Peterson, N. A., Speer, P. W., & Hughey, J. (2006). Measuring sense of community: A methodological interpretation of the factor structure debate. *Journal of Community Psychology*, *34*(4), 453–469. doi:10.1002/jcop.20109

Preacher, K. J., Zhang, G., Kim, C., & Mels, G. (2013). Choosing the optimal number of factors in exploratory factor analysis: A model selection perspective. *Multivariate Behavioral Research*, *48*(1), 28–56. doi:10.1080/00273171.2012.710386

Raîche, G., Walls, T. A., Magis, D., Riopel, M., & Blais, J.-G. (2013). Non-graphical solutions for Cattell's scree test. *Methodology*, *9*, 23–29. doi:10.1027/1614-2241/a000051

Rammstedt, B., Kemper, C. J., Klein, M. C., Beierlein, C., & Kovaleva, A. (2017). A short scale for assessing the big five dimensions of personality: 10 item Big Five Inventory (BFI-10). *Methods, Data, Analyses*, *7*(2), 233–249. doi:10.12758/mda.2013.013

Reichenbach, H. (1956). The direction of time. Los Angeles, CA: University of California Press.

Rhemtulla, M., Brosseau-Liard, P. É., & Savalei, V. (2012). When can categorical variables be treated as continuous? A comparison of robust continuous and categorical sem estimation methods under suboptimal conditions. *Psychological Methods*, *17*(3), 354–373. doi:10.1002/mpr.352

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Model-agnostic interpretability of machine learning. *arXiv Preprint arXiv:1606.05386.*

Ronald, A., Happé, F., Hughes, C., & Plomin, R. (2005). Nice and nasty theory of mind in preschool children: Nature and nurture. *Social Development*, *14*(4), 664–684. doi:10.1111/j.1467-9507.2005.00323.x

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), 282–292. doi:10.1037/a0025697

Saggino, A. (2000). The big three or the big five? A replication study. *Personality and Individual Differences*, *28*(5), 879–886. doi:10.1016/S0191-8869(99)00146-4

Scharf, F., & Nestler, S. (2019). Should regularization replace simple structure rotation in exploratory factor analysis? *Structural Equation Modeling: A Multidisciplinary Journal*, 1–15. doi:10.1080/10705511.2018.1558060

Shoemaker, P. J., Eichholz, M., & Skewes, E. A. (2002). Item nonresponse: Distinguishing between don't know and refuse. *International Journal of Public Opinion Research*, *14*(2), 193–201. doi:10.1093/ijpor/14.2.193

Shrout, P. E., & Rodgers, J. L. (2018). Psychology, science, and knowledge construction: Broadening perspectives from the replication crisis. *Annual Review of Psychology*, *69*(1), 487–510. doi:10.1146/annurev-psych-122216-011845

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*(11), 1359–1366. doi:10.1177/0956797611417632

Smith, C. A., Organ, D. W., & Near, J. P. (1983). Organizational citizenship behavior: Its nature and antecedents. *Journal of Applied Psychology*, *68*(4), 653–663. doi:10.1037/0021-9010.68.4.653

Spearman, C. (1904). "General intelligence" objectively determined and measured. *American Journal of Psychology*, *15*(2), 201–292. Retrieved from https://www.jstor.org/stable/1412107

Steiger, J. H. (1979). Factor indeterminacy in the 1930's and the 1970's some interesting parallels. *Psychometrika*, *44*(2), 157–167. doi:10.1007/BF02293967

Thalmayer, A. G., Saucier, G., & Eigenhuis, A. (2011). Comparative validity of brief to medium-length big five and big six personality questionnaires. *Psychological Assessment*, *23*(4), 995–1009. doi:10.1037/a0024165

Thurstone, L. L. (1940). Current issues in factor analysis. *Psychological Bulletin*, *37*(4), 189–236.

Thurstone, L. L. (1947). *Multiple factor analysis*. Chicago, IL: University of Chicago Press.

Velicer, W. F. (1976). Determining the number of components from the matrix of partial correlations. *Psychometrika*, *41*(3), 321–327. doi:10.1007/BF02293557

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 R. Goffin & E. Helmes (Eds.), *Problems and solutions in human assessment* (pp. 41–71). Boston, MA: Springer.

Widaman, K. F. (2007). Common factors versus components: Principals and principles, errors and misconceptions. In R. Cudeck & R. C. MacCallum (Eds.), *Factor analysis at 100* (pp. 177–204). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Wright, M. N., & Ziegler, A. (2017). ranger: A fast implementation of random forests for high dimensional data in C++ and R. *Journal of Statistical Software*, *77*(1), 1–17. doi:10.18637/jss.v077.i01

Zwick, W. R., & Velicer, W. F. (1986). Comparison of five rules for determining the number of components to retain. *Psychological Bulletin*, *99*(3), 432–442. doi:10.1037/0033-2909.99.3.432

9 Study 1

The article entitled “Exploratory factor analysis: Current use, methodological developments and recommendations for good practice.” published in *Current Psychology* (Goretzko, Pham, & Bühner, 2019) is referred to as Study 1 throughout this thesis. It is presented hereinafter.

9.1 Abstract

Psychological research often relies on exploratory factor analysis (EFA). As the outcome of the analysis highly depends on the chosen settings, there is a strong need for guidelines on the decisions a researcher faces when conducting an EFA. Therefore, we want to examine the recent methodological developments as well as the current practice in psychological research. We reviewed ten years of studies containing EFAs and contrasted them with new methodological options. We focused on four major issues: an adequate sample size, the extraction method, the rotation method and the factor retention criterion determining the number of factors to extract. Finally, we present modified recommendations based on these reviewed empirical studies and practical considerations.

9.2 Introduction

Exploratory factor analysis (EFA) is a frequently used statistical method in psychology. There is hardly any other statistical method shaping the field of test construction as strongly as the EFA, simultaneously causing as many controversial debates about its correct application. Over the years, several publications dealt with recommendations on how to use EFA, trying to familiarize researchers with the most important decisions they have to make.

One of the most influential papers in this context, a meta-analytic review by Fabrigar, Wegener, MacCallum and Strahan (1999), investigated the use of EFA in 217 papers published from 1991 through 1995 in the *Journal of Personality and Social Psychology* (JPSP) and the *Journal of Applied Psychology* (JAP). The authors made recommendations for the practical application of EFA regarding an appropriate sample size, the number of items per factor, the extraction method, the factor retention criterion as well as the rotation method and the general applicability of the procedure. In the following, these recommendations will be discussed briefly and afterwards compared with the current use of EFA in psychological research. This is done by reviewing publications in two highly relevant journals for psychological assessment published over the last decade. The latest developments and empirical findings concerning methodological decisions during the EFA process are presented and merged with those of (Fabrigar et al., 1999) to obtain enriched recommendations for EFA. We focus on sample size considerations, the choices of rotation and extraction methods as well as the best way to determine the number of factors.

9.2.1 Theory and Purpose of EFA. EFA is used to explore correlative relations among manifest variables and to model these relations with one or more latent variables. In the common factor model a causal link between latent variable(s) and manifest indicators is assumed ("common cause relation") – an assumption that is comprehensively discussed with all its implications by Borsboom, Mellenbergh and Van Heerden (2003). Based on the common factor model, the covariance matrix of the manifest variables can be decomposed into a part of shared variance $\mathbf{\Lambda}\mathbf{\Lambda}^\top$ (impact of the latent variable(s) or the "common cause") and unique variance $\mathbf{\Psi}^2$:

$$\Sigma = \mathbf{\Lambda}\mathbf{\Lambda}^\top + \mathbf{\Psi}^2$$

(Jöreskog, 1967).

When factor loadings and unique variances are estimated, one faces the problem of rotation indeterminacy which means that the loading matrix can only be defined up to a rotation, because more latent variables have to be estimated than manifest variables are observed (that is why ML estimation, for example, uses an iterative estimation procedure, see Jöreskog, 1967). Steiger (1979) discusses the related issue of factor indeterminacy in detail including historical perspectives. Given a rotated solution, there are no unique solutions for factor scores – a problem that should be considered when interpreting EFA results (see also Steiger & Schönemann, 1978 for a simple numerical example). When EFA is used as a tool for defining psychological constructs and developing associated questionnaires, rotation indeterminacy might be the predominant issue, so we focus on the related methodological decisions, yet inviting the readers to keep in mind the problem of factor indeterminacy especially when considering the results (factor scores) for diagnostic purposes.

9.2.2 Recommendations of Fabrigar et al. (1999). Fabrigar et al. (1999) established basic guidelines for the general study design, the extraction methods, the rotation methods and the factor retention criteria. In the following, these recommendations are presented briefly.

9.2.2.1 Study design (number of items and sample size).

Besides others, there are two important issues the researcher has to consider when designing a study – the number of variables representing a latent construct and the sample size. Fabrigar et al. (1999) suggest that one should find at least four items with acceptable reliabilities (> .70) for each expected factor. Contrary to former opinions (e.g. Ford, MacCallum, & Tait, 1986; Gorsuch, 1983), the authors do not support the idea of a subjects-per-variable ratio as a guiding value for the sample size. In fact, they recommend sample sizes greater than 400 as desirable, as smaller samples might yield invalid results under unfavorable conditions (e.g. low communalities, MacCallum, Widaman, Zhang, & Hong, 1999).

9.2.2.2 Extraction methods.

When comparing different extraction methods, Fabrigar et al. (1999) conclude that Maximum Likelihood (ML) estimation might be the preferred approach due to the numerous fit indices available for this method. The authors propose three alternatives, when the assumption of multivariate normality is violated: transforming the data, correcting the fit indices or using a different method like principal axis factoring (PAF).

9.2.2.3 Factor retention criteria.

To determine the number of factors, Fabrigar et al. (1999) recommend to use different criteria and never just one method. They advise researchers to combine fit indices (when using ML EFA) like RMSEA, as proposed by Browne and Cudeck (1992), with common methods such as parallel analysis (PA, Horn, 1965). In the case of a sufficiently large sample, they encourage researchers to split the data set and compare the results of the factor retention criteria among the subsets.

9.2.2.4 Rotation methods.

When it comes to the various rotation methods provided by statistical software, Fabrigar et al. (1999) have a strong call for oblique procedures as these can lead to uncorrelated and correlated factors which usually occur in psychology, whereas orthogonal rotation methods force an uncorrelated factor solution. However, they do not give any further recommendations which specific oblique rotation should be favored in which condition.

9.2.2.5 General recommendations.

The paper of Fabrigar et al. (1999) has two key learnings. First, it can be seen as a strong call for a more thoughtful application of EFA. The authors emphasized that EFA and principal component analysis (PCA) are different methods, especially with regard to unique variances, and should not be exchanged unintentionally. Second, the paper draws attention to the fact that the presentation of the method and its results often does not allow for assessing the quality of the analysis. Therefore, Fabrigar et al. (1999) emphasize the importance of transparent and coherent presentations of the whole EFA procedure and criticize the rare documentation of important EFA settings or characteristics of the data. Especially, the lack of information on item communalities is noted by the authors.

9.3 Review of the Current Use of EFA

As 20 years have passed since the work of Fabrigar et al. (1999), we want to examine what has changed in the meantime and whether the discussed recommendations have been adopted by a broader community. Therefore, we sifted every original article in *Psychological Assessment* and in the *European Journal of Psychological Assessment* (EJPA) from 2007 to 2017. These journals were selected due to their special focus on test construction and the variety of studies using EFA. The database research yielded 993 studies in Psychological Assessment, issues 19(1)-29(4) and 336 studies in EJPA, issues 23(1)-33(1). For our analysis, we focused on articles reporting an EFA (e.g. studies on test construction) and excluded articles which did not report an EFA as a main analysis (e.g. studies only referring to EFA results in the footnotes). We analyzed a total of 304 EFAs, 44 from EJPA and 260 from Psychological Assessment (some papers with more than one EFA). To quantify the current EFA practice, we classified the respective sample sizes, the extraction methods, the rotation methods, the factor retention criteria, the number of variables per factor as well as the average communalities in each EFA. Articles directly referring to Fabrigar et al. (1999) were also considered separately, as we wanted to examine whether these articles showed a higher compliance to the presented recommendations.

9.3.1 Study Design (Number of Items and Sample Size). Table 2 shows the sample sizes reported for each of the 304 EFAs. About half of the analyses (50.3%) were based on samples larger than 400, while only eight cases (2.6%) had samples smaller than 100. In 1.6% of all cases the sample size was not presented at all.

The ratios of variables per factor are shown in Table 3, reporting the general ratio for each EFA as well as the minimum number of items associated with a factor in each analysis. In 10.5% of the EFAs the general ratio was not provided. In nearly one third of the analyses

Table 2

Study 1: Sample Sizes in EFAs in Current Psychological Research

Sample Size	N	%	$\%_{Fabrigar}$
< 100	8	2.6	17.5
100 - 200	42	13.8	26.7
200 - 300	44	14.5	15.7
300 - 400	52	17.1	6.9
> 400	153	50.3	33.2
Not Reported	5	1.6	0.0

(31.6%), the smallest number of items of a factor was not listed as well. On the other hand, more than half of the considered EFAs (52.6%) had a general item to factor ratio of five or higher with 22.0% reporting a ratio of 10 or greater. At least three variables associated with the smallest factor were reported for 57.9% of the EFAs, with 11.5% having at least six variables associated with each factor.

We were also interested in the means of communalities for each EFA as those can be seen as indicators for the quality of the measurement (when developing scales and seeking for unidimensional constructs that are represented by several manifest indicators) or rather as measures for the soundness of the extracted factors (Table 4). The vast majority of studies neither specified the communalities nor gave enough information to calculate them (pattern matrix and correlation among factors). Thus, 87.5% of all EFAs were published neglecting the communalities. When item communalities were reported, they mostly fell between .40 and .70 (10.5% of all EFAs).

Table 3

Study 1: Item to Factor Ratio (A) and Minimum Variables per Factor (B)

	N	%
A		
$<$ 3:1	10	3.3
3:1 - 5:1	102	33.6
5:1 - 7:1*	47	15.5
$>$ 7:1	113	37.2
Not Reported	32	10.5
B		
$<$ 3	32	10.5
3 - 5	141	46.4
$>$ 5	35	11.5
Not Reported	96	31.6

Note. *The item to factor ratio of exactly five to one is included here

Table 4

Study 1: Average Communalities in EFAs in Current Psychological Research

Average Communalities $\bar{h}^2$	N	%
$<$ .40	4	1.3
.40 - .49	11	3.6
.50 - .59	12	3.9
.60 - .69	9	3.0
$>$ .69	2	0.7
Not Reported	266	87.5

Note. Communalities are averaged over each EFA, so the range of communalities is ignored.

9.3.2 Extraction Method. The present usage of extraction methods is shown in Table 5. It should be noted that PCAs are excluded from this review. Therefore, only EFAs in the narrow sense, those allowing for unique variances, are included. With 51.3%, the majority of EFAs was based on PAF, followed by ML estimation (16.4%). Least-Squares approaches made up less than ten percent of the used extraction methods. In more than 22% of the cases the extraction method was not reported at all.

9.3.3 Rotation Method. Table 6 shows the rotation methods used in the analyzed EFAs. As two of the EFAs were conducted using two different rotation methods for comparison, a total 306 cases are reported. 71.4% of the reported EFAs were implemented with oblique rotation methods, while 20.4% did not report the rotation method. Most researchers chose Promax (32.2%) or Oblimin (14.5%) for oblique rotation. Varimax (8.9%) was the only orthogonal rotation method found in our sample.

9.3.4 Factor Retention Criterion. More than half of the time researchers did not rely solely on one factor retention criterion to determine the number of factors, but used multiple criteria instead (note that because of the usage of multiple criteria the percentages in Table 7 do not add up to one). The most common method was the Kaiser-Guttman criterion (often referred to as Eigenvalue $>$ 1 rule) used in 55.6% of the cases, followed by the Scree test (46.4%), PA (42.1%) and theoretical considerations or interpretability of the solution (35.5%). In some cases, only one of these four methods were used as a single criterion. When reporting just one stand-alone criterion, Kaiser-Guttman was the

Table 5

Study 1: Extraction Methods in EFAs in Current Psychological Research

Extraction Method	N	%
Principal Axis Factoring	156	51.3
Maximum Likelihood	50	16.4
Unweighted Least Squares	11	3.6
Weighted Least Squares*	16	5.3
MinRes	3	1.0
Not Reported	68	22.4

Note. *WLSMV: Weighted Least Squares Means and Variance adjusted

Table 6

Study 1: Rotation Methods in EFAs in Current Psychological Research

Rotation Method	N	%	$\%_{Fabrigar}$
Varimax	27	8.9	53.0
Promax	98	32.2	1.8
Oblimin	44	14.5	0.0
Geomin	23	7.6	0.0
Equimax	5	1.6	0.0
GeoMax	2	0.7	0.0
Varimax (oblique)	11	3.6	0.0
Other oblique rotations	34	11.2	13.8
Not Reported	62	20.4	15.2

most common (10.5%), followed by Scree test (9.5%) and PA (8.2%). In total, we found 16 different methods applied as retention criterion in our review (Table 7).

9.3.5 Studies with References to Fabrigar et al. (1999). The analyses from articles directly citing Fabrigar et al. (1999) produced quite different results. PAF was prevalently used as the extraction method (88%) while ML estimation was used only once. Every applied rotation method was oblique with Promax being reported the most frequently (82%). 80% of the articles reported multiple criteria and PA was the predominant retention criterion with 88%, while Kaiser-Guttman (82%), Scree test (74%) and MAP test (66%) were used at least roughly two-thirds of the time as well. In general, these articles showed a higher tendency to report our variables of interest. At least 40% of them provided both the information about communalities as well as complete information about the other relevant variables.

Table 7

Study 1: Factor Retention Criteria in Current Psychological Research

Factor Retention Criterion	N	%
Kaiser-Guttman*	169	55.6
Scree test*	141	46.4
Parallel Analysis*	128	42.1
Theory/Interpretability*	108	35.5
AIC	1	0.3
BIC	8	2.6
Chi-Square-Test	1	0.3
Comparison Data	1	0.3
Eigenvalue > .70	1	0.3
Variance accounted for	24	7.9
At least 3 Variables per factor	16	5.3
MAP test	59	19.4
RMSEA	3	1.0
SRMR	1	0.3
Standard Error Scree	16	5.3
Very Simple Structure	1	0.3
Not Reported	13	4.3

Note. Percentages do not add up to 1, because multiple criteria were used among the majority of EFAs. * Criteria, that were used standalone to determine the number of factors in at least one study.

9.4 Methodological Developments

As there are several new methodological developments in the field of EFA, we want to present an updated review of the methodological questions arising when conducting EFA. The discussed recommendations of Fabrigar et al. (1999) serve as the basis of our overview, which is why the following sections focus primarily on concepts and empiricism published after the year 1999.

9.4.1 Study Design (Number of Items and Sample Size). When it comes to EFA, an adequate sample size is a heavily discussed issue. As Fabrigar et al. (1999) point out, recommendations concerning subject to item ratios ($\frac{N}{p}$) are out of date. In fact, MacCallum et al. (1999) showed in a simulation study that these ratios are not useful, and furthermore, that the communalities of the analyzed variables and the number of items per factor should be considered when searching for an appropriate sample size. Rouquette and Falissard (2011) evaluated the requirements for sample sizes in EFA in the context of psychiatric scales. They found that the subject to item ratio rules did not work appropriately and concluded that it is not necessarily true that shorter scales need smaller samples than larger scales or vice versa. Therefore, they recommended a rule of thumb of 300 subjects or more when using EFA in this specific context.

Other studies followed the findings of MacCallum et al. (1999). Hogarty, Hines, Kromrey, Ferron and Mumford (2005) reported a strong influence of item communalities on the accuracy of EFA solutions. Especially when overdetermination was strong (e.g. three factors represented by 20 variables) and communalities were high (h^2 between .60 and .80), sample factor loadings and population factor loadings corresponded vastly. Quite similar results were obtained in simulations by Mundfrom, Shaw and Ke (2005): the higher the item communalities were and the stronger overdetermination was, the smaller the sample could have been to find accurate factor solutions. Thus, even samples smaller than 100 observations could be appropriate when communalities are sufficiently high and factors are represented by a great number of items.

Contrary to EFA, there are some methods to determine sample size for CFA which go beyond common rules of thumb (Schmitt, 2011). One of them is a method based on Monte Carlo simulations evaluating the minimum sample size for a particular model and a desired power for the Likelihood ratio test (Muthen & Muthen, 2002). This process determining the sample size analogously to sample size planning for other analyses (e.g. ANOVA) seems to be a practicable solution for CFA, but will not fit in the context of EFA as necessary assumptions about the factor structure and the size of loadings cannot be made in advance (otherwise CFA would be the preferred analysis method).

As there is often little or no evidence in advance about the concrete size of the item communalities, one has to come back to rough rules of thumb. We therefore recommend to (highly) overdetermine the expected factors and stick with an item to factor ratio of at least 4, better 5, so that samples of approximately 400 subjects will promise trustworthy results (see, Mundfrom et al., 2005). Hogarty et al. (2005) likewise recommended overdetermination to limit the need of excessive sample sizes due to potentially low item communalities. Increasing the item to factor ratio can be harmful though, when the content validity is not considered. Artificial duplication of items can lead to violations of local independence. The item to factor ratio should therefore be increased carefully.

The number of observations which allows for stable estimations of correlations (as EFA is based on the correlative relations among variables) might be another reference value for a desirable sample size. Schönbrodt and Perugini (2013) demonstrate at which sample sizes Pearson correlations stabilize depending on different levels of confidence and definitions of stability. As secondary loadings in EFA are often based on rather small correlations more than 300 observations seem to be necessary to achieve reasonably stable correlations in this context. Therefore, this rough assessment is an additional indicator that the rule of thumb of Rouquette and Falissard (2011) with sample sizes greater than 300 might be a good lower bound when planning the sample for an EFA. We agree with Fabrigar et al. (1999) that samples containing at least 400 observations should be aimed for to avoid estimation problems. Even though there are some methods especially designed for small samples (e.g. Jung & Takane, 2008; see Extraction Methods), using those should be exceptional and reserved for cases in which strong ethical or resource-related objections can be made. In general, researchers should collect greater samples so that factor loadings and factor scores are estimated more precisely – especially when tests are designed for clinical diagnostics.

9.4.1.1 Current practice.

Against this background, it is encouraging to see that sample sizes in our review tend to be higher than in the study of Fabrigar et al. (1999) twenty years ago. This might be an effect of the differing journals we used for our review, but it could also indicate real improvements in current practice. As the sample size can be judged only when communalities and item-to-factor ratios are known, one has to be cautious with results of studies based on extremely small samples when these measures are not reported. Thus, we recommend to provide this information within every article. Sample sizes of more than 400 observations are essential when conducting EFAs and should not be smaller to improve estimation precision and to prevent estimation problems such as Heywood cases. The tendency of studies directly referring to Fabrigar et al. (1999) showing higher sample sizes than the average of the considered studies, can be seen as a confirmation that methodological education can help to improve psychological research.

9.4.2 Extraction Method. Another central decision when performing EFA is the choice of an appropriate extraction method. It has been stated repeatedly that PCA is not the same as EFA and therefore PCA is not an equivalent alternative when dealing with latent variables measured by manifest items (Costello & Osborne, 2005; Fabrigar et al., 1999; Gorsuch, 1990, 1997). A short introduction on the differences between EFA and PCA is presented by Suhr (2005). When item communalities are close to one both methods yield similar results while results can differ heavily when communalities decrease. The decision between EFA and PCA should be linked directly to the purposes of the analysis – when exploring latent constructs that are measured (measurement error!) via manifest indicators common EFA should be preferred.

Even when excluding PCA from the set of possible extraction methods, researchers are confronted with various different options: ML estimation, Minres introduced by Harman and Jones (1966), different least squares approaches, Minimum Rank Factor Analysis (MRFA) and PAF, just to name the most common ones. Jöreskog, Olsson and Yang-Wallentin (2013) point out that ML estimation can be described as an iteratively reweighted least squares approach (for more detail, see Browne, 1974). So, the framework of the weighted least squares family (WLS) covers ML, unweighted least squares (ULS) as well as generalized least squares (GLS) as special cases.

Despite these methodological similarities among them, the choice of an extraction method can have a severe impact on the concrete EFA solution and the literature lacks advice which exact extraction method should be used under which conditions (Costello & Osborne, 2005; Fabrigar et al., 1999). Numerous researchers (e.g. Conway & Huffcutt, 2003; Costello & Osborne, 2005) follow Fabrigar et al. (1999) preferring ML estimation when multivariate normality is given. Again, the main reason for this preference is the variety of fit indices one can use for model evaluation and comparison. In addition, ML estimation is implemented in all major statistical programs (e.g. SPSS, FACTOR, R, MPLUS).

However, using Likert type items multivariate normality might be questionable. When multivariate normality is violated, Costello and Osborne (2005) recommend PAF, while Yong and Pearce (2013) suggest to conduct PCA at first to reduce the dimensionality of the data and subsequently perform a "real" factor extraction using one of the methods above.

Accordingly, PAF is often used as an alternative extraction method. De Winter and Dodou (2012) compared PAF and ML estimation via simulations and showed that ML estimation was more likely to produce Heywood cases throughout all conditions, but outperformed PAF when loadings were unequal and underextraction was given. PAF, on the other hand, performed better when the factor structure was orthogonal and when overextraction was present. So, neither PAF nor ML estimation can be seen as preferable in general.

Barendse, Oort and Timmerman (2015) compared ML with WLS and robust WLS for different response scales (continuous, dichotomous and polytomous) and found robust WLS with polychoric correlations to yield better results when discrete data was evaluated – findings comparable to those that have been made in the field of confirmatory factor analysis (CFA). Beauducel and Herzberg (2006), for example, compared ML estimation to weighted least squares means and variance adjusted (WLSMV) estimation for CFA by simulating data sets based on variables with different response scale formats. They found that WLSMV performed better for variables with two or three categories which are situations where normality assumptions might be questionable anyway. Comparable results were reported by Rhemtulla, Brosseau-Liard and Savalei (2012), who showed that ML estimation can be used when variables have five or more categories yielding results of equal quality as WLSMV. Both simulation studies revealed a slight greediness of WLSMV estimation for greater sample sizes. These findings might not be applicable directly to EFA, but they can give some evidence which conditions might be suitable for either ML estimation or WLS approaches.

All of these estimation algorithms require a minimum sample size (another reason for rather big samples, see section Sample Size) and do not provide reliable results with small samples. Therefore, a regularized EFA for small sample sizes has been proposed (Jung & Takane, 2008). Contrary to common estimation methods (e.g. ML), it does not estimate the unique variances for each item and the factor loadings iteratively, but rather estimates a single regularization parameter[19] λ to avoid improper solutions. The regularization pa-

[19] Regularization means that an additional term is added to an objective function to solve an otherwise not solvable problem. Here, instead of estimating several unique variances which can be infeasible when the sample size is too small, a so-called regularization parameter is selected that adjusts the initial estimates of

rameter λ shrinks the initial estimates of the unique variances, while the factor loadings are estimated as usual with common ML, ULS or GLS estimations. Initially, the unique variances are either assumed to be constant across all variables, proportional to the anti-image[20] covariance (see, e.g. Kaiser, 1976) or proportional to the Ihara-Kano estimates (see, Ihara & Kano, 1986). Jung and Lee (2011) showed in a simulation study that this procedure (with ML estimation and anti-image assumption) works better for small samples (less than 50 observations) than common ML estimation or PCA. Nevertheless, the assumptions for the unique variances are hardly ever met in psychological studies, so this procedure should be reserved for situations where common extraction procedures are not feasible for a given sample size.

9.4.2.1 Current practice.

In the majority of studies PAF was used - a tendency that was even stronger for those referring to Fabrigar et al. (1999). Yet, there are several advantages of ML and the Least-Squares approaches as mentioned above. EFA results should be cross-validated with CFA, so we recommend to use ML or LS approaches instead of PAF as these estimation methods are available for CFA as well and therefore provide comparable outcomes among the analyses. For normally distributed data, one should rely rather on ML estimation, whereas WLS estimation should be preferred for non-normal and ordinal data (especially when Likert type items with less than five categories are used). Extracting via PAF should rather be restricted to cases where the other extraction methods suffer from non-convergence or improper solutions. Depending on the particular data, more than one method can be tried, though and results can be examined for matching patterns as suggested by Widaman (2012).

9.4.3 Factor Retention Criterion. Determining the number of factors is a very decisive issue in the EFA process because of its influential power within the exploratory analysis. While in many articles authors write about the true number of factors and the problem to find this exact number, Preacher, Zhang, Kim and Mels (2013) argue that there is no true factor model and researchers rather have to approximate the data generating

the unique variances.

[20] The anti-image can be depicted as the negative of the image of a matrix. The image covariance matrix contains the variation of each variable that can be explained by the other variables (partial covariance coefficients), the respective anti-image consists of the negatives which can be described as the unique variance components. For more details, have a look at Kaiser (1976) or EFA textbooks since the anti-image correlation matrix is a commonly used tool to evaluate whether an EFA is applicable to the data (see also Measuring Sampling Adequacy (MSA) as described in Kaiser, 1970).

process. The authors describe an error framework which covers two different directions in the factor retention issue.

Preacher et al. (2013) explain that one has to choose the aim of the EFA - approximating the “true” factor structure (approximation goal) or finding the most replicable solution (replicability goal) which is a decision analogue to the bias-variance trade-off. They conclude that different factor retention criteria are best for these different goals. In simulation studies the authors focused on fit indices based on ML estimation and found the RMSEA (to be more precise: its confidence interval’s lower bound) to perform best for the approximation goal while AIC and BIC were far less accurate especially in scenarios with great sample sizes. Contrary, for the replicability goal and in cases of small samples BIC performed best.

Often the approximation goal has priority in EFA research. There is a broad range of evidence that in this case, PA produces the best results when comparing the most common criteria (see, Fabrigar et al., 1999; Peres-Neto, Jackson, & Somers, 2005; Zwick & Velicer, 1986). The generally good performance of PA might be based on its robustness against varying distributional assumptions (Dinno, 2009). Timmerman and Lorenzo-Seva (2011) evaluated different extraction methods within the PA and recommended to use MRFA instead of PCA or PAF for ordered polytomous items which are usually used in psychological questionnaires.

PA has become some kind of gold standard for factor retention criteria, but promising alternatives have been proposed recently. Lorenzo-Seva, Timmerman and Kiers (2011) developed the so-called hull method. This method is based on four major steps. First the researcher chooses a range of possible numbers of factors, then an arbitrary fit index is evaluated for each number of factors (CFI performed best in simulations). Afterwards the degrees of freedom of this set of factor solutions is assessed and finally the values of the chosen fit index are plotted against the respective degrees of freedom. The higher boundary[21] of the convex hull of the plotted data points shows an elbow which defines the number of factors to retain. The authors showed a superiority of their method to PA and the minimum average partial test (MAP) in simulations and for a real data set. This reported superiority of the hull method was based on cases with an extremely high item to factor ratio ($\frac{p}{k} = 20$). In cases of smaller ratios PA yielded equivalent or even better results.

Another method is the comparison data (CD) approach (Ruscio & Roche, 2012). CD can be framed as an extended PA which reproduces the observed correlation matrix instead

[21] Sometimes (in the context of two dimensional convex hulls) also referred to as the upper hull.

of using random data. The researcher specifies the upper bound for the possible number of factors. Then data for populations with one, two, etc. factors (up to this predefined upper bound) are simulated each reproducing the given empirical covariance structure as closely as possible. Samples (the authors suggest 500) of the same size as the empirical data are drawn from each population and the respective eigenvalues of the item correlation matrix are compared to the observed eigenvalues via the Root-Mean-Square-Error (RMSE)[22]. One gets as many RMSE values as samples drawn from each population. These values of each factor solution are then compared to those of the next factor solution by a nonparametric Mann-Whitney U test (the one factor solution against the two factor solution, the two factor solution against the three factor solution and so on). The iterative procedure stops when no significant improvement is indicated (Ruscio & Roche, 2012).

In simulation studies the authors showed that an α-level of .30 may be adequate (note that an α-error means possible overextraction, while a β-error means underextraction) and that CD outperformed PA and other minor retention criteria under several conditions[23].

As this method (and similar approaches using simulated data) can be computationally demanding, Braeken and Van Assen (2017) proposed the Empirical Kaiser Criterion (EKC) which makes use of the statistical properties of eigenvalues and does not require any simulations[24]. It is based on the so-called Marčenko-Pastur distribution, which asymptotically describes the distribution of sample eigenvalues under the null model (no underlying factor structure) and is therefore closely related to the results of PA, and the idea of the Kaiser criterion that only eigenvalues greater than one should be taken into account. The theoretically expected eigenvalues (that are used for the comparison) are corrected for the magnitude of all previous eigenvalues - so when, for example, the first empirical eigenvalue already accounts for 60% of the item variance, the expectations for the following eigenvalues are adjusted downwards. The authors were able to show superiority over PA for oblique

[22] The RMSE is defined as the root of the MSE which is the averaged squared distance between parameters and its estimates. In this case, the differences between the given eigenvalues and the eigenvalues obtained of the simulated data sets of the specific k-factor population are computed: $RMSE_t = \sqrt{\frac{\sum_{i=1}^{p}(\eta^*_{t,i}-\eta_i)^2}{p}}$ where the empirical eigenvalues are denoted η and the comparison eigenvalues of t-th comparison data set ($t = 1, ..., T$ where the default for T is 500) are denoted η^*_t and p is the number of eigenvalues (variables in the empirical data set).

[23] They varied the number of factors (one to five), the number of response categories (two to 20), used correlated and uncorrelated solutions and sample sizes between 200 and 1000.

[24] It only requires the number of items and the sample size, so it can be applied without knowing much about the structure of the data – for example when evaluating published results.

structures and found comparable results to CD and other simulation based approaches. However, this evaluation was based on simple structure assumptions, so little is known so far about the performance of EKC when cross-loadings are present.

9.4.3.1 Current practice.

In current research, more than 50% of the EFAs are based on multiple factor retention criteria, whereas Fabrigar et al. (1999) reported just about 20% of studies to do so. In articles referring to their article, the percentage rises to 80%. That speaks in favor of the current research practice, although the frequent use of invalid methods such as Kaiser-Guttman rule or Scree test (also as a single criterion) has to be criticized. There are even tutorial papers for EFA recommending these methods Maroof (2012) or completely ignoring more appropriate tools (Beavers et al., 2013). In addition to avoid these criteria, it should become scientific standard not to rely on the MAP-test as a (stand-alone) factor retention criterion in common factor analysis as it is created for PCA and therefore associated with different assumptions.

As there is enough evidence demonstrating problems with some of the commonly used criteria, we want to encourage researchers to use the whole spectrum of methods determining the number of factors and whenever feasible to split the sample and evaluate the subsamples separately. A practical solution could be using PA and CD in combination with a descriptive measure like the explained variance or theoretical considerations. Nevertheless, this decision still remains the most difficult to make within EFA. Thus, it is inevitable to be aware of its consequences and to report every consideration concerning this issue.

9.4.4 Rotation Method. After the primary extraction, researchers almost always decide to rotate the factor solution to obtain results that are easier to interpret. It has become common understanding in literature on EFA methods that oblique rotation is preferable (e.g. Baglin, 2014; Conway & Huffcutt, 2003; Costello & Osborne, 2005; Fabrigar et al., 1999), but it is also stated that it is not clear which oblique rotation has to be used. Browne (2001) gives a detailed overview of the different rotation methods and highly recommends a multi-method approach. He argues that using various complexity functions[25] might be an appropriate way to handle a situation in which no solution is undoubtedly superior. One could use a method from the Crawford-Ferguson (CF) family, plus Infomax

[25] The so-called complexity function is the objective function which is minimized with regard to specific constraints to achieve a particular rotation of the initial loading matrix. We recommend the article of Browne (2001) explaining the link between constraints and rotation criteria in greater detail.

rotation and Geomin rotation, for example. The CF family (Crawford & Ferguson, 1970) covers several well-known rotation methods by formulating the complexity function as a function of row complexity (items) and column complexity (factors):

$$f(\Lambda) = (1-\kappa)\sum_{i=1}^{p}\sum_{j=1}^{k}\sum_{l\neq j}^{k}\lambda_{ij}^2\lambda_{il}^2 + \kappa\sum_{j=1}^{k}\sum_{i=1}^{p}\sum_{m\neq i}^{p}\lambda_{ij}^2\lambda_{mj}^2$$

with p indicating the number of variables, k indicating the number of factors and κ being an arbitrary constant weighting the row-wise (first part of the sum) or column-wise complexity. Some values of κ lead to commonly known criteria. $\kappa = 0$, for example, corresponds to the Quartimin-criterion and $\kappa = \frac{1}{p}$ to the Varimax rotation (Browne, 2001). Browne (2001) explains that in cases of almost perfect cluster patterns most complexity functions work perfectly fine, but when complexity in the factor patterns increases, one has to weigh up the stability of the solution against its accuracy.

When complex structures are expected (higher amount and amplitude of cross-loadings), rotation methods like CF-Equamax or CF-Facparsim should be used. When fewer or smaller cross-loadings are expected, common techniques like Geomin or CF-Quartimin might be more appropriate (Sass & Schmitt, 2010; Schmitt & Sass, 2011).

Browne (2001) states that a standardization like the Cureton-Mulaik (CM) weighting (for more details, see Cureton & Mulaik, 1975) can improve the solution. Nonetheless, if there are only a few complex variables among many perfectly discriminative ones (only loadings on one factor), weighting procedures might focus too much on these variables. The advantages of CM weighting were empirically shown by Lorenzo-Seva (2000) comparing weighted Oblimin with Direct Oblimin, Promaj, Promin and weighted Promax.

Another interesting, yet different rotation method is the rotation to target procedure, where the factor matrix is rotated in a way that a partially specified target matrix (some coefficients of the factor pattern matrix are defined in advance) is replicated as closely as possible (Myers, Jin, Ahn, Celimli, & Zopluoglu, 2015). It seems to be an appropriate rotation method when additional information is available as it has some similarities to exploratory structure equation modelling (ESEM; for further readings, see Marsh, Morin, Parker, & Kaur, 2014). Therefore, it might be the right rotation method when theoretical or empirical information is given and when many factor cross loadings tend to be zero, because this seems to be the most reasonable specification a researcher can make in advance. Browne (2001) suggests to apply this procedure iteratively, updating the target matrix in every step.

As the choice of the best rotation method appears to be arbitrary to some degree, we want to present a totally different approach: the penalized factor analysis (Hirose & Yamamoto, 2014). Instead of conducting EFA in the classical two steps – extracting k factors and afterwards rotating the solution to increase interpretability – this new method obtains sparsity in the pattern matrix through penalizing the likelihood. The penalty[26] is analogue to the complexity function discussed before, but it is now integrated into the estimation process, so that cross-loadings get shrunk towards zero in the first place.

First simulations revealed some promising results for wide data with many variables and sparse loading matrices (Hirose & Yamamoto, 2015) as well as for a real data set (Hirose & Yamamoto, 2014). The latter was analyzed with both the new approach with a MC+ penalty and the common ML estimation with Promax rotation. The penalized factor analysis produced quite similar yet sparser and well interpretable results. Nevertheless, the penalized factor analysis still has to be evaluated under a broader range of conditions to investigate whether it will be an appropriate tool for psychological research questions.

In general, it is appropriate to use different rotation methods and to choose the one with the most reasonable solution as all rotated solutions are mathematically equal[27] (in case of the two-step process[28], not the penalized ML estimation). For replication purposes, it is necessary to report the chosen rotation procedure.

9.4.4.1 Current practice.

For one out five cases in our review, the rotation method was not reported, so the current research practice clearly lacks transparency here. Only two studies used different rotation methods and compared different solutions – a procedure highly recommended by Browne (2001). However, a positive aspect is that more than 70% of all EFAs used oblique rotation methods. A number that increased to 98% for studies referring to Fabrigar et al. (1999), where 53% of the examined studies had used the orthogonal Varimax rotation.

[26] In common ML estimation an objective function (that is derived from the log-likelihood) is minimized. Here a so-called penalty term is added to this function. It penalizes a high number of parameters (in this case loadings, especially cross-loadings). The more parameters are estimated to be non-zero, the higher this term gets and it “becomes harder” to achieve a minimum, so in turn adding this penalty yielding more small (or even zero-) loadings (depending on the type of penalty). You can read about penalizing the likelihood in the EFA estimation process in more detail in Jin, Moustaki and Yang-Wallentin (2018).

[27] That is the known as the problem of rotation indeterminacy (see introduction section)

[28] The optimization process is done with respect to different constraints, but apart from that it is equivalent for all rotation methods. Therefore, theoretical considerations must be taken into account to make a reasonable decision (are cross-loadings consistent with theoretical assumptions, etc.).

Accordingly, psychological research seems to be on the right track regarding this issue.

When using different rotation methods on the same sample or on subsamples as suggested by Browne (2001), Fabrigar et al. (1999) or Preacher et al. (2013), it might be helpful to evaluate the similarity of different solutions. Lorenzo and Ferrando (1996, 1998), for example, developed the *FACOM/NFACOM* library which allows for comparison of different factor solutions based on different methodological decisions. A decision for a rotation can be made by weighing up the mathematical interpretability and the theoretical plausibility of the respective solution.

9.4.5 Further Recommendations. EFA is often applied to questionnaire items which are not normally distributed but rather skewed. Investigating this problem, Holgado–Tello, Chacón–Moscoso, Barbero–García and Vila–Abad (2010) found EFA based on polychoric correlations to reproduce the true factor model more accurately than EFA based on Pearson correlations. Baglin (2014) nicely illustrated this issue and also recommended polychoric correlations for these cases. Hence, item skewness should be evaluated before conducting EFA and in case of severely skewed variables polychoric correlations should be chosen when using ML estimation. When extracting via WLS approaches, polychoric or tetrachoric correlations are used instead of Pearson correlation anyway, so the type of correlation does not have to be selected in these cases. Other approaches worth considering for ordinal data are those that are based on response patterns (IRT models) instead of approximating the correlation matrix assuming underlying normal distributed latent variables (e.g. Jöreskog & Moustaki, 2001). As full-information item factor analysis (full information maximum likelihood, FIML) can be computationally challenging (IRT approach as well as when assuming an underlying continuous variable) and problematic with small samples in particular, Katsikatsou, Moustaki, Yang-Wallentin and Jöreskog (2012) proposed a pairwise likelihood (PML) estimation approach that closely matched the results of FIML and can be seen as a practical alternative.

When conducting EFA, researchers should specify their research goals precisely and select the best suited methods. We want to clarify that the presented methods and related recommendations are designed for the researcher's goal to approximate the data generating process as precisely as possible. Often interpretability and theoretical considerations can be equally important. For test construction purposes, in particular, content validity should be the top priority. Researchers therefore should report transparently which objectives they have, which methodological decisions they take and all outcomes they collect. This ensures that the quality of a solution can be evaluated and implications of particular studies can be weighted. The *Journal of the Society for Social Work and Research* has taken a leading role in demanding certain reporting guidelines for EFA (see Cabrera-Nguyen, 2010). Other journals should follow this example and call for transparency in reporting EFAs. Especially in the light of the current discussion about the replication crisis in psychology (e.g. Shrout & Rodgers, 2018), transparency with regard to data, research material and methodological decisions is essential (for further readings: OSF Guidelines for Transparency, Klein et al., 2018). Furthermore, we encourage researchers to consider various procedures in this context, instead of performing a standard practice based on default settings or personal routines.

9.5 Summary

As pointed out, EFA is a very complex analysis and it is therefore not easy to make general recommendations on how to conduct it properly. Each case should be evaluated individually, so this paper tries to sensitize researchers for careful decisions and transparent reporting. Nevertheless, we want to formulate some “default” settings which can be seen as a basis for further considerations. Samples for EFA should be greater than 400 participants to get reliable factor patterns and precisely estimated factor scores. One should use ML or WLS estimation as extraction method depending on the respective item distributions and the response format, because these methods allow for evaluations of model fit and cross-validation with CFAs. To determine the number of factors, we recommend combining PA and CD (or EKC) with a descriptive measure (e.g. explained variance) and theoretical considerations. Latter should be included for test construction purposes, but should be ignored when the data generating process of the specific data is approximated. In any case, multiple retention criteria should be applied and reported later on to provide the full picture. As different rotation methods yield mathematically indeterminate factor solutions, researchers should compare factor patterns between different methods and choose the solution that fits theoretical considerations best. Again, it is necessary to report the chosen method to enable other researchers to replicate the respective solution.

9.6 References

Baglin, J. (2014). Improving your exploratory factor analysis for ordinal data: A demonstration using FACTOR. *Practical Assessment, Research & Evaluation*, *19*(5). Retrieved from http://pareonline.net/getvn.asp?v=19&n=5

Barendse, M. T., Oort, F. J., & Timmerman, M. E. (2015). Using exploratory factor analysis to determine the dimensionality of discrete responses. *Structural Equation Modeling: A Multidisciplinary Journal*, *22*(1), 87–101. doi:10.1080/10705511.2014.934850

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

Beavers, A. S., Lounsbury, J. W., Richards, J. K., Huck, S. W., Skolits, G. J., & Esquivel, S. L. (2013). Practical considerations for using exploratory factor analysis in educational research. *Practical Assessment, Research & Evaluation*, *16*(6). Retrieved from https://scholarworks.umass.edu/pare/vol18/iss1/6

Borsboom, D., Mellenbergh, G. J., & Van Heerden, J. (2003). The theoretical status of latent variables. *Psychological Review*, *110*(2), 203–219. doi:10.1037/0033-295X.110.2.203

Braeken, J., & Van Assen, M. A. (2017). An empirical Kaiser criterion. *Psychological Methods*, *22*(3), 450–466. doi:10.1037/met0000074

Browne, M. W. (1974). Generalized least squares estimators in the analysis of covariance structures. *South African Statistical Journal*, *8*(1), 1–24. Retrieved from https://journals.co.za/content/sasj/8/1/AJA0038271X_175

Browne, M. W. (2001). An overview of analytic rotation in exploratory factor analysis. *Multivariate Behavioral Research*, *36*(1), 111–150. doi:10.1207/S15327906MBR3601_05

Browne, M. W., & Cudeck, R. (1992). Alternative ways of assessing model fit. *Sociological Methods & Research*, *21*(2), 230–258. doi:10.1177/0049124192021002005

Cabrera-Nguyen, P. (2010). Author guidelines for reporting scale development and validation results in the Journal of the Society for Social Work and Research. *Journal of the Society for Social Work and Research*, *1*(2), 99–103. Retrieved from http://www.jstor.org/stable/10.5243/jsswr.2010.8

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

Costello, A. B., & Osborne, J. W. (2005). Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Practical Assessment, Research & Evaluation*, *10*, 1–9. Retrieved from https://scholarworks.umass.edu/pare/vol10/iss1/7/

Crawford, C. B., & Ferguson, G. A. (1970). A general rotation criterion and its use in orthogonal rotation. *Psychometrika*, *35*(3), 321–332. doi:10.1007/BF02310792

Cureton, E. E., & Mulaik, S. A. (1975). The weighted varimax rotation and the promax rotation. *Psychometrika*, *40*(2), 183–195. doi:10.1007/BF02291565

De Winter, J. C., & Dodou, D. (2012). Factor recovery by principal axis factoring and maximum likelihood factor analysis as a function of factor pattern and sample size. *Journal of Applied Statistics*, *39*(4), 695–710. doi:10.1080/02664763.2011.610445

Dinno, A. (2009). Exploring the sensitivity of Horn's parallel analysis to the distributional form of random data. *Multivariate Behavioral Research*, *44*(3), 362–388. doi:10.1080/00273170902938969

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

Ford, J. K., MacCallum, R. C., & Tait, M. (1986). The application of exploratory factor analysis in applied psychology: A critical review and analysis. *Personnel Psychology*, *39*(2), 291–314. doi:10.1111/j.1744-6570.1986.tb00583.x

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

Gorsuch, R. L. (1983). *Factor analysis*. Hillsdale, NY: Lawrence Erlbaum Associates.

Gorsuch, R. L. (1990). Common factor analysis versus component analysis: Some well and little known facts. *Multivariate Behavioral Research*, *25*(1), 33–39.

doi:10.1207/s15327906mbr2501_3

Gorsuch, R. L. (1997). Exploratory factor analysis: Its role in item analysis. *Journal of Personality Assessment*, *68*(3), 532–560. doi:10.1207/s15327752jpa6803_5

Harman, H. H., & Jones, W. H. (1966). Factor analysis by minimizing residuals (minres). *Psychometrika*, *31*(3), 351–368. doi:10.1007/BF02289468

Hirose, K., & Yamamoto, M. (2014). Estimation of an oblique structure via penalized likelihood factor analysis. *Computational Statistics & Data Analysis*, *79*, 120–132. doi:10.1016/j.csda.2014.05.011

Hirose, K., & Yamamoto, M. (2015). Sparse estimation via nonconcave penalized likelihood in factor analysis model. *Statistics and Computing*, *25*(5), 863–875. doi:10.1007/s11222-014-9458-0

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

Holgado–Tello, F. P., Chacón–Moscoso, S., Barbero–García, I., & Vila–Abad, E. (2010). Polychoric versus pearson correlations in exploratory and confirmatory factor analysis of ordinal variables. *Quality & Quantity*, *44*(1), 153–166. doi:10.1007/s11135-008-9190-y

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, *30*(2), 179–185. doi:10.1007/BF02289447

Ihara, M., & Kano, Y. (1986). A new estimator of the uniqueness in factor analysis. *Psychometrika*, *51*(4), 563–566. doi:10.1007/BF02295595

Jin, S., Moustaki, I., & Yang-Wallentin, F. (2018). Approximated penalized maximum likelihood for exploratory factor analysis: An orthogonal case. *Psychometrika*, *83*(3), 628–649. doi:10.1007/s11336-018-9623-z

Jöreskog, K. G. (1967). Some contributions to maximum likelihood factor analysis. *Psychometrika*, *32*(4), 443–482. doi:10.1007/BF02289658

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

doi:10.1207/S15327906347-387

Jöreskog, K. G., Olsson, U. H., & Yang-Wallentin, F. (2013). *Multivariate analysis with lisrel.* Basel, Sui: Springer International Publishing.

Jung, S., & Lee, S. (2011). Exploratory factor analysis for small samples. *Behavior Research Methods*, *43*(3), 701–709. doi:10.3758/s13428-011-0077-9

Jung, S., & Takane, Y. (2008). Regularized common factor analysis. In *New trends in psychometrics* (pp. 141–149). Tokyo: Universal Academy Press.

Kaiser, H. F. (1970). A second generation little jiffy. *Psychometrika*, *35*(4), 401–415. doi:10.1007/BF02291817

Kaiser, H. F. (1976). Image and anti-image covariance matrices from a correlation matrix that may be singular. *Psychometrika*, *41*(3), 295–300. doi:10.1007/BF02293555

Katsikatsou, M., Moustaki, I., Yang-Wallentin, F., & Jöreskog, K. G. (2012). Pairwise likelihood estimation for factor analysis models with ordinal data. *Computational Statistics & Data Analysis*, *56*(12), 4243–4258. doi:10.1016/j.csda.2012.04.010

Klein, O., Hardwicket, T. E., Aust, F., Breuer, J., Danielsson, H., Mohr, A. H., ... Frank, M. C. (2018). A practical guide for transparency in psychological science. *COLLABRA-PSYCHOLOGY*, *4*(1). doi:10.1525/collabra.158

Lorenzo, U., & Ferrando, P. J. (1996). FACOM: A library for relating solutions obtained in exploratory factor analysis. *Behavior Research Methods, Instruments, & Computers*, *28*(4), 627–630. doi:10.3758/BF03200553

Lorenzo, U., & Ferrando, P. J. (1998). NFACOM: A new program for relating solutions in exploratory factor analysis. *Behavior Research Methods, Instruments, & Computers*, *30*(4), 724–725. doi:10.3758/BF03209493

Lorenzo-Seva, U. (2000). The weighted oblimin rotation. *Psychometrika*, *65*(3), 301–318. doi:10.1007/BF02296148

Lorenzo-Seva, U., Timmerman, M. E., & Kiers, H. A. L. (2011). The hull method for selecting the number of common factors. *Multivariate Behavioral Research*, *46*(2), 340–364. doi:10.1080/00273171.2011.564527

MacCallum, R. C., Widaman, K. F., Zhang, S., & Hong, S. (1999). Sample size in factor analysis. *Psychological Methods*, *4*(1), 84–89. doi:10.1037/1082-989X.4.1.84

Maroof, D. A. (2012). Exploratory factor analysis. In *Statistical methods in neuropsychology: Common procedures made comprehensible* (pp. 23–34). Boston, MA: Springer US. doi:10.1007/978-1-4614-3417-7_4

Marsh, H. W., Morin, A. J., Parker, P. D., & Kaur, G. (2014). Exploratory structural equation modeling: An integration of the best features of exploratory and confirmatory factor analysis. *Annual Review of Clinical Psychology*, *10*(1), 85–110. doi:10.1146/annurev-clinpsy-032813-153700

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. doi:10.1207/s15327574ijt0502_4

Muthen, L. K., & Muthen, 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. doi:10.1207/S15328007SEM0904_8

Myers, N. D., Jin, Y., Ahn, S., Celimli, S., & Zopluoglu, C. (2015). Rotation to a partially specified target matrix in exploratory factor analysis in practice. *Behavior Research Methods*, *47*(2), 494–505. doi:10.3758/s13428-014-0486-7

Peres-Neto, P. R., Jackson, D. A., & Somers, K. M. (2005). How many principal components? Stopping rules for determining the number of non-trivial axes revisited. *Computational Statistics & Data Analysis*, *49*(4), 974–997. doi:10.1016/j.csda.2004.06.015

Preacher, K. J., Zhang, G., Kim, C., & Mels, G. (2013). Choosing the optimal number of factors in exploratory factor analysis: A model selection perspective. *Multivariate Behavioral Research*, *48*(1), 28–56. doi:10.1080/00273171.2012.710386

Rhemtulla, M., Brosseau-Liard, P. É., & Savalei, V. (2012). When can categorical variables be treated as continuous? A comparison of robust continuous and categorical sem estimation methods under suboptimal conditions. *Psychological Methods*, *17*(3), 354–373. doi:10.1002/mpr.352

Rouquette, A., & Falissard, B. (2011). Sample size requirements for the internal validation of psychiatric scales. *International Journal of Methods in Psychiatric Research*, *20*(4), 235–249. doi:10.1002/mpr.352

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), 282–292. doi:10.1037/a0025697

Sass, D. A., & Schmitt, T. A. (2010). A comparative investigation of rotation criteria within exploratory factor analysis. *Multivariate Behavioral Research*, *45*(1), 73–103. doi:10.1080/00273170903504810

Schmitt, T. A. (2011). Current methodological considerations in exploratory and confirmatory factor analysis. *Journal of Psychoeducational Assessment*, *29*(4), 304–321. doi:10.1177/0734282911406653

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. doi:10.1177/0013164410387348

Schönbrodt, F. D., & Perugini, M. (2013). At what sample size do correlations stabilize? *Journal of Research in Personality*, *47*(5), 609–612. doi:10.1016/j.jrp.2013.05.009

Shrout, P. E., & Rodgers, J. L. (2018). Psychology, science, and knowledge construction: Broadening perspectives from the replication crisis. *Annual Review of Psychology*, *69*(1), 487–510. doi:10.1146/annurev-psych-122216-011845

Steiger, J. H. (1979). Factor indeterminacy in the 1930's and the 1970's some interesting parallels. *Psychometrika*, *44*(2), 157–167. doi:10.1007/BF02293967

Steiger, J. H., & Schönemann, P. H. (1978). A history of factor indeterminacy. In S. S. (Ed.), *Theory construction and data analysis* (pp. 136–178). San Francisco, CA: Jossey-Bass.

Suhr, D. D. (2005). Principle component analysis versus exploratory factor analysis. Philadelphia, PA.

Timmerman, M. E., & Lorenzo-Seva, U. (2011). Dimensionality assessment of ordered polytomous items with parallel analysis. *Psychological Methods*, *16*(2), 209–220. doi:10.1037/a0023353

Widaman, K. F. (2012). Exploratory factor analysis and confirmatory factor analysis. In H.

Cooper, P. M. Camic, D. L. Long, A. T. Panter, D. Rindskopf, & K. J. Sher (Eds.), *APA handbook of research methods in psychology, vol 3: Data analysis and research publication.* Washington, DC: American Psychological Association.

Yong, A. G., & Pearce, S. (2013). A beginners guide to factor analysis: Focusing on exploratory factor analysis. *Tutorials in Quantitative Methods for Psychology*, *9*(2), 79–94. doi:10.20982/tqmp.09.2.079

Zwick, W. R., & Velicer, W. F. (1986). Comparison of five rules for determining the number of components to retain. *Psychological Bulletin*, *99*(3), 432–442. doi:10.1037/0033-2909.99.3.432

10 Study 2

The article entitled “One model to rule them all? Using machine learning algorithms to determine the number of factors in exploratory factor analysis” submitted for publication is referred to as Study 2 throughout this thesis. It is presented hereinafter.

10.1 Abstract

Determining the number of factors is one of the most crucial decisions a researcher has to face when conducting an exploratory factor analysis. As no common factor retention criterion can be seen as generally superior, a new approach is proposed - combining extensive data simulation with state-of-the-art machine learning algorithms. First, data was simulated under a broad range of realistic conditions and three algorithms were trained using specially designed features based on the correlation matrices of the simulated data sets. Subsequently, the new approach was compared to four common factor retention criteria with regard to its accuracy in determining the correct number of factors in a large-scale simulation experiment. Sample size, variables per factor, correlations between factors, primary and cross-loadings as well as the correct number of factors were varied to gain comprehensive knowledge of the efficiency of our new method. A gradient boosting model outperformed all other criteria, so in a second step, we improved this model by tuning several hyperparameters of the algorithm and using common retention criteria as additional features. This model reached an out-of-sample accuracy of 99.3%. A great advantage of this approach is the possibility to continuously extend the data basis (e.g. using ordinal data) as well as the set of features to improve the predictive performance and to increase generalizability.

10.2 Introduction

Exploratory factor analysis (EFA) is a commonly used statistical method to explore latent psychological concepts. Its exploratory nature allows researchers to carve out the structure of new constructs, but also reflects a threat to the validity of its results as several methodological decisions have to be taken by the researchers - decisions that can strongly shape the outcome and future research in the respective field (Fabrigar, Wegener, MacCallum, & Strahan, 1999; Goretzko, Pham, & Bühner, 2019). The most crucial decision might be determining the number of factors that should be extracted. Extracting too few factors (underfactoring) or too many (overfactoring) can have adverse effects on the estimated factor scores (Fava & Velicer, 1996) and leads to estimation problems such as Heywood cases (De Winter & Dodou, 2012). Overfactoring is generally regarded as less critical since the actual relations between the manifest variables and the latent variables can be estimated more accurately than in cases of underfactoring (Fabrigar et al., 1999).

There are numerous ways to determine the number of factors. In some cases, theoretical considerations set the number of factors, but often such implications are missing and the number has to be estimated based on the empirical data. The majority of the various methods that have been developed for this issue evaluate the eigenvalue-structure of the item correlations. There are traditional approaches like the Scree test (Cattell, 1966), the Kaiser-Guttman rule (Kaiser, 1960) and the parallel analysis (Horn, 1965) as well as modern approaches like the comparison data (CD) approach (Ruscio & Roche, 2012) or the empirical Kaiser criterion (EKC, Braeken & Van Assen, 2017). Parallel analysis (PA) became some kind of gold-standard (e.g. Fabrigar et al., 1999; Goretzko et al., 2019) due to its robustness against varying distributional assumptions (Dinno, 2009) and its comparably good performance under various conditions such as sample sizes between 30 and 360 and number of variables between 9 and 72 (Peres-Neto, Jackson, & Somers, 2005; Zwick & Velicer, 1986). Nonetheless, new methods like CD (Ruscio & Roche, 2012), the hull method proposed by Lorenzo-Seva, Timmerman and Kiers (2011) and EKC (Braeken & Van Assen, 2017) showed superiority for specific data conditions[29]. A broad simulation study by Auerswald and Moshagen (2019) evaluated these (and other) criteria under various conditions (number of items: $4-60$, sample sizes: $100-1000$, number of factors: $1-5$, variables per factor: $4-12$, varying loading magnitudes and inter-factor correlations) and recommended combining different methods. They found combinations of PA (based on principal component analysis), EKC, the hull method and CD (when sample sizes were sufficiently large) to provide the best results.

10.2.1 Aim of the Study. Combining various methods is also recommended by several other authors (Fabrigar et al., 1999; Goretzko et al., 2019), yet this suggestion can be unsettling and frustrating for practitioners. For this reason, a new approach is tested in this study - combining extensive data simulation with machine learning algorithms to find a model that is predictive (can determine the number of factors correctly) under a broad range of conditions. We focus on two major approaches: random forests (Breiman, 2001)

[29] CD was superior to PA especially in conditions with few factors ($k < 5$), while the EKC was superior for conditions with oblique factor structures and the Hull method outperformed PA when overdetermination was high.

and extreme gradient boosting (Chen, He, Benesty, Khotilovich, & Tang, 2018) as well as an automatic gradient boosting approach (Thomas, Coors, & Bischl, 2018) since both random forests and gradient boosting are able to reflect non-linearities and complex interactions.

10.2.2 Random Forest. Random forests are based on n_{tree} bootstrap samples drawn from the empirical data. A regression or classification tree is grown on each bootstrap sample by recursive binary splitting. This growth-process stops when a predetermined number of observations is reached in each terminal node. The algorithm randomly uses m_{try} variables at each node, of which one variable is selected that allows for the best split (Breiman, 2001). The resulting trees are not pruned like single decision trees, because overfitting is prevented by averaging over the n_{tree} trees. These trees can vary a lot as only m_{try} of all features are used for each possible split and the bootstrapped samples do not consist of all original observations. The process of building a random forest by averaging the m_{try} trees provides reliable results (see James, Witten, Hastie, & Tibshirani, 2013 for a more detailed introduction). Common values for m_{try} are $\frac{p}{3}$ or $\sqrt{p}$ (with p being the number of variables or features). $\frac{p}{3}$ is preferred for regression task while $\sqrt{p}$ might be favorable for classification purposes (Breiman, 1999).

10.2.3 Extreme Gradient Boosting and Automatic Gradient Boosting. The principal idea of boosting is to sequentially perform a prediction task (regression or classification) with comparably simple (or "weak") methods like decision trees (Friedman, Hastie, & Tibshirani, 2000). Contrary to random forests or bagging approaches, no bootstrap samples are drawn. In fact, a number of decision trees (n_{tree}) is grown sequentially using the residuals of the complete model containing all previous trees. A shrinkage parameter (e.g. λ) also known as learning rate regulates the updating speed:

$$\hat{f}(x) \leftarrow \hat{f}(x) + \lambda \hat{f}^{new}(x)$$

where $\hat{f}(x)$ is the iteratively updated model, which yields $\hat{f}(x) = \sum_{n=1}^{n_{tree}} \lambda \hat{f}^{n}(x)$ for the complete boosted model. λ is usually chosen as $\lambda = 0.01$ or $\lambda = 0.001$ (James et al., 2013).

For gradient boosting this rate is not fixed for each new tree, but rather computed by minimizing the residuals in the respective step given a predefined loss-function (Friedman, 2001). As the *xgboost* algorithm (Chen & Guestrin, 2016; Chen et al., 2018) used for this study contains several hyperparameters (e.g. a learning rate that shrinks the step weights, a L_1-regularization and a L_2-regularization as well as tree-specific parameters like the minimal node size), tuning the *xgboost* model might be promising. Thomas et al. (2018) provided

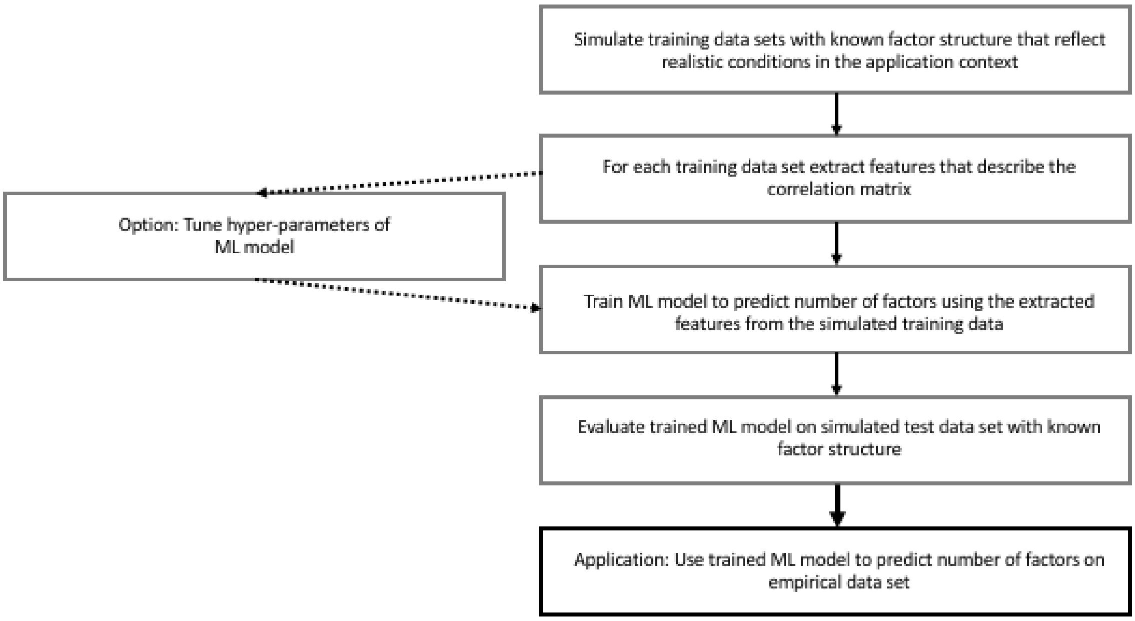
Simulate training data sets with known factor structure that reflect realistic conditions in the application context
For each training data set extract features that describe the correlation matrix
Option: Tune hyper-parameters of ML model
Train ML model to predict number of factors using the extracted features from the simulated training data
Evaluate trained ML model on simulated test data set with known factor structure
Application: Use trained ML model to predict number of factors on empirical data set

Figure 4. Study 2: Visualization of the new factor retention approach.

an automatized version - the automatic gradient boosting or *autoxgboost* which applies model-based optimization (Bayesian optimization) to find the best set of hyperparameters.

10.3 Methods

The idea of this study was to find a machine learning model that is predictive for the true number of factors (the number of latent dimensions underlying the data generating process) in the context of EFA. We therefore simulated several data sets with given factor structures that reflect realistic conditions in psychological research. Afterwards, three machine learning models (random forest, gradient boosting and automatic gradient boosting) were trained on this data (see Figure 4 for a flowchart demonstrating the approach). The performance of the resulting three trained models were compared to the performance of parallel analysis, the comparison data approach and the empirical Kaiser criterion as well as the common Kaiser criterion. This was done using new simulated data that also covered a broad range of conditions usually found in psychological literature.

10.3.1 Creating a Machine Learning Model as Factor Retention Criterion. The underlying data basis was simulated assuming multivariate normality. Sample sizes were between 200 and 1000, the true number of factors (k) ranged from 1 to 8 factors, variables per factor (vpf) varied between 3 and 10, factor correlations were set to values between 0 and 0.4, primary loadings ranged from 0.35 to 0.80 and cross-loadings from 0.00 and 0.20[30].

A population correlation matrix was created for each data set based on the following decomposition:

$$\mathbf{\Sigma} = \mathbf{\Lambda}\mathbf{\Phi}\mathbf{\Lambda}^\top + \mathbf{\Psi}^2$$

with $\mathbf{\Lambda}$ being the true loading matrix, $\mathbf{\Phi}$ being the factor correlation matrix and $\mathbf{\Psi}^2$ being a diagonal matrix containing the unique variance of each variable. The true loading matrix contained all primary and cross-loadings drawn from different uniform distributions (e.g. when primary loadings should be high a uniform distribution between 0.65 and 0.80 was used). For $\mathbf{\Phi}$ a matrix that consists of the value one on the diagonal and equal values for the inter-factor correlations on the off-diagonal (0,0.1,0.2,0.3 or 0.4) was chosen, while $\mathbf{\Psi}^2$ was calculated from $\mathbf{\Sigma} = \mathbf{\Lambda}\mathbf{\Phi}\mathbf{\Lambda}^\top + \mathbb{1}_{p\times p} - diag(\mathbf{\Lambda}\mathbf{\Phi}\mathbf{\Lambda}^\top)$.

Data simulation and analysis were conducted with R (R Development Core Team, 2008) while the manuscript was written with the *papaja* package (Aust & Barth, 2018) and graphics were created with the *ggplot2* package (Wickham, 2016). We used the *mvtnorm* package (Genz et al., 2018) to simulate multivariate normal data with the respective correlation matrix Σ (consequently, all manifest variables had unit variance) and N observations (N was drawn from a uniform distribution between 200 and 1000).

10.3.2 Feature Engineering. 181 features were computed for each simulated data set to create the respective training data for the machine learning algorithms. The following features were used for training: the sample size N, the number of variables p, the number of eigenvalues that are greater than one, the relative proportion of the first eigenvalue, the relative proportion of the first two eigenvalues, the relative proportion of the first three eigenvalues, the number of eigenvalues that are greater than 0.7, the standard deviation of all eigenvalues, the number of eigenvalues that account for 50% of the variance, the number of eigenvalues that account for 75% of the variance, the L_1-norm of the correlation matrix, the Frobenius-norm of the correlation matrix, the maximum-norm of the correlation matrix, the spectral-norm of the correlation matrix, the average of the off-diagonal correlations, the number of correlations smaller or equal to 0.1, the average of

[30] Primary loadings were either small ($\lambda_{ij} \in [0.35; 0.50]$), medium ($\lambda_{ij} \in [0.50; 0.65]$) or high ($\lambda_{ij} \in [0.65; 0.80]$) and cross-loadings were either non-existing, small ($\lambda_{ij} \in [0.00; 0.10]$) or medium sized ($\lambda_{ij} \in [0.10; 0.20]$) representing different levels of communalities.

the initial communality estimates, the determinant of the correlation matrix, the measure of sampling adequacy (MSA after Kaiser, 1970), the Gini-coefficient (Gini, 1921) of the correlation matrix, the Kolm measure of inequality (Kolm, 1999) of the correlation matrix, all p eigenvalues as well as all p eigenvalues of the factor model[31].

We simulated 500000 data sets varying the sample size, primary and cross-loadings, the number of factors, the variables per factor and the factor correlations. For some of the random combinations, the resulting Σ-matrix was not positive semi-definite or the calculation of all features was not feasible, so our effective training sample consisted of 498971 simulated data sets with 181 features.

10.3.3 Model Training. Based on the simulated data, we used the machine learning framework *mlr* (Bischl et al., 2016) to train a random forest (*ranger*, Wright & Ziegler, 2017), the extreme gradient boosting model (*xgboost*, Chen et al., 2018) and the automatic gradient boosting model (*autoxgboost*, Thomas et al., 2018). Determining the true number of factors k was implemented as a classification task (multiclass) with no specific cost-matrix - the algorithms were trained to maximize the accuracy of the suggested factor solution which means that no differentiations were made among falsely classified cases (e.g. suggesting four factors when $k = 2$ had the same costs as suggesting five factors).

The *ranger* was applied with default settings ($n_{tree} = 500$, which seems to be a good trade-off between performance and the need for computational resources, see Genuer, Poggi, & Tuleau, 2008 and $m_{try} = floor(\sqrt{p})$ which is the rounded down square root of the number of features as recommended by Breiman, 1999 - in our case $m_{try} = floor(\sqrt{181}) = 13$). We also used the default settings of the *xgboost*[32], but set the number of iterations (the trees that are sequentially build on the residuals) to 500 for better comparison with the *ranger*. The *autoxgboost* was used with default settings as well. We only increased the time budget[33] of the algorithm from one hour to two hours.

We saved the three trained models to evaluate them on new data. The *ranger* model reached an in-sample accuracy of 97.2% while the *xgboost* model had an in-sample accuracy

[31] For both common eigenvalues and eigenvalues based on the factor model, the maximum number was 80 as $p = k * vpf$ could have been $p = 8 * 10$ in maximum. Missing values (for each simulated data set with $p < 80$) were coded with -1000.

[32] The default settings were used as we wanted to know whether the *xgboost* algorithm is useful at all. As there are many possible tuning parameters for this algorithm, the *autoxgboost* implementation was tested as well.

[33] The time budget is the maximum time that will be used for tuning the parameters of the underlying boosting algorithm via Bayesian optimization.

of 99.0% and the *autoxgboost* model an in-sample accuracy of 95.8%.[34]

10.3.4 Evaluation of the Machine Learning Models and four common Factor Retention Criteria. To evaluate the performance of the three models in more detail and on new data, we created several experimental conditions and simulated multivariate normal data. We varied the following conditions: sample size was $N = 250, 500$ or 1000, variables per factor were either $vpf = 4$,5 or 7, the true number of factors was $k = 1$,2,4 or 6, between factor correlations[35] were $\rho = 0, 0.1, 0.2, 0.3$ or 0.5 and loadings varied between 0.35 and 0.80 for primary loadings (different conditions with small ($\lambda_{ij} \in [0.35; 0.50]$), medium ($\lambda_{ij} \in [0.50; 0.65]$) and high ($\lambda_{ij} \in [0.65; 0.80]$) primary loadings) and 0.00 and 0.20 for cross-loadings (different conditions with no, small ($\lambda_{ij} \in [0.00; 0.10]$) and medium sized ($\lambda_{ij} \in [0.10; 0.20]$) cross-loadings). This gave us 3204 conditions in total as we excluded combinations that could yield improper solutions for Σ. Each condition was replicated 500 times, so 1512000 data sets were evaluated. Data simulation was conducted analogously to the simulation of the data basis for the training set.

We calculated all necessary features and saved the predictions of the three models for each data set. We also collected the suggested number of factors for four common factor retention criteria (Kaiser criterion, PA, CD, EKC) for comparison. Accuracies, ratios of under- or overfactoring as well as minima and maxima of the suggested number of factors per conditions were then calculated. For the sake of clarity, we used only four common criteria for the comparison that can be used as a baseline[36] for our new approach (e.g. foregoing the hull method which is superior to PA only in rather special conditions with high overdetermination and the minimum average partial test [MAP; Velicer, 1976] which is designed for principal component analyses rather than EFA in a narrow sense and which is not able to outperform PA [Zwick & Velicer, 1986]). For the same reason, we also focused on one implementation of the parallel analysis (using the 95% quantile of the eigenvalue distribution of random data for comparison as implemented in the *psych* package by Rev-

[34] The apparently lower accuracy (in-sample) of the *autoxgboost* compared with the *xgboost* indicates that the time budget was too short to find optimal settings for the hyperparameters of the gradient boosting algorithm.

[35] We evaluated oblique structures with (nearly) simple structure rather than the related orthogonal structures with higher cross-loadings as researchers almost always search for simple structure and many common rotation methods were designed to provide solutions with simple structure (Browne, 2001; Fabrigar et al., 1999; Goretzko et al., 2019).

[36] We focused on common factor retention criteria that are applied by practitioners and seemed to be promising for our data conditions. Including methods for this baseline made sense for those criteria that were shown to be superior to PA for some of these data conditions.

Table 8

Study 2: Accuracy of Retention Criteria averaged over All Conditions and for Different Factor Solutions separately

Method	Acc	Acc_1	Acc_2	Acc_4	Acc_6
xgboost	0.92886	0.99663	0.95883	0.90646	0.85002
ranger	0.91508	0.99625	0.94913	0.90314	0.80701
axgboost	0.85600	0.99583	0.93834	0.79607	0.68620
pa	0.82506	0.76951	0.90248	0.85971	0.76590
cd	0.81304	0.85624	0.90297	0.79599	0.69157
ekc	0.88432	0.99916	0.95634	0.82596	0.74983
kaiser	0.74644	0.96389	0.85003	0.64023	0.52161

Note. Acc is the overall accuracy of each method, whereas Acc_1 is the accuracy for single factor conditions, Acc_2 for conditions with two factors and so on.

elle, 2018), so all results concerning PA are related to a specific implementation and cannot be generalized to other types of parallel analyses. The simulation studies of Auerswald and Moshagen (2019) and Lim and Jahng (2019) provide further insights on these different types of parallel analysis.

10.4 Results

Averaged over all 3204 conditions, both trained models *ranger* and *xgboost* had a higher accuracy than the common factor retention criteria. When considering conditions with $k = 1, 2, 4$ and 6 separately, the *xgboost* model had the highest accuracy on average for two-factor, four-factor and six-factor solutions and fell short closely behind the EKC when one-factor solutions were evaluated (99.7% to 99.9%). While all retention criteria general had lower accuracies when k was higher, the *xgboost* model exclusively reached accuracies higher than 85% on average. Table 8 shows the accuracies of all methods averaged over all conditions as well as the accuracies for all different values of k separately.

Besides having the highest overall accuracy, the *xgboost* model showed no signs of biased estimation of the number of factors (estimated bias smaller than 0.01). The *ranger* showed no clear signs of bias in the estimation as well, yet it tended to overfactor when $k = 6$ (10.9% of the cases). In contrast, PA and CD tended to underfactor suggesting less than k factors for all values of k - a tendency that increased with higher values of k (e.g. when $k = 6$ PA underestimated k in 22.6% of the cases while CD did so in 23.3% of the cases). The Kaiser-Guttman-rule rather tended to overfactor (20.3% of the cases with $k = 4$ and 25.9% of the cases with $k = 6$), while EKC was nearly unbiased for $k = 1$ and $k = 2$ and suggested less than k factors when four or six factors (22.0% of the cases) were in the data generating model. In Table 9 the averaged suggested numbers of factors are presented for each criterion and the four different values of k, while Table 10 displays the respective proportions of under- and overfactoring.

Table 9

Study 2: Mean and Median Solution for each Number of Factors

Method	$Mean_1$	$Bias_1$	$Median_1$	$Mean_2$	$Bias_2$	$Median_2$	$Mean_4$	$Bias_4$	$Median_4$	$Mean_6$	$Bias_6$	$Median_6$
xgboost	1.00404	0.00404	1.00000	1.99281	-0.00719	1.99383	4.00270	0.00270	4.00247	6.00720	0.00720	6.02067
ranger	1.00465	0.00465	1.00000	1.98084	-0.01916	1.99506	4.01826	0.01826	4.00000	6.03094	0.03094	6.03682
axgboost	1.00491	0.00491	1.00000	1.99562	-0.00438	1.99383	4.08578	0.08578	4.05123	6.02002	0.02002	6.02196
pa	0.83837	-0.16163	0.77778	1.87148	-0.12852	1.89383	3.74010	-0.25990	3.78272	5.30216	-0.69784	5.33269
cd	1.16709	0.16709	1.00000	2.00030	0.00030	1.96235	3.79098	-0.20902	3.75432	5.45288	-0.54712	5.39664
ekc	1.00079	0.00079	1.00000	1.96730	-0.03270	1.97284	3.69512	-0.30488	3.71358	5.35788	-0.64212	5.40310
kaiser	1.03779	0.03779	1.03704	2.11237	0.11237	2.09136	4.16895	0.16895	4.16543	6.21663	0.21663	6.20155

Note. $Mean_1$ means the average suggested number of factors for conditions with one true factor, $Median_1$ is the respective median suggested number and $Bias_1$ the average deviance in these conditions.

Table 10

Study 2: Proportions of Under- and Overfactoring for the Different Number of Factors

Method	$\% - under_1$	$\% - over_1$	$\% - under_2$	$\% - over_2$	$\% - under_4$	$\% - over_4$	$\% - under_6$	$\% - over_6$
xgboost	0.000	0.003	0.025	0.016	0.052	0.041	0.071	0.079
ranger	0.000	0.004	0.036	0.015	0.045	0.052	0.084	0.109
axgboost	0.000	0.004	0.034	0.028	0.083	0.121	0.147	0.167
pa	0.198	0.033	0.085	0.013	0.133	0.008	0.226	0.008
cd	0.000	0.144	0.050	0.047	0.149	0.055	0.233	0.075
ekc	0.000	0.001	0.040	0.004	0.158	0.016	0.220	0.030
kaiser	0.000	0.036	0.040	0.110	0.157	0.203	0.220	0.259

Table 11

Study 2: Accuracy of Retention Criteria for Different Sample Sizes

Method	Acc_{250}	Acc_{500}	Acc_{1000}
xgboost	0.88754	0.93549	0.96355
ranger	0.88408	0.92132	0.93985
axgboost	0.81523	0.86006	0.89271
pa	0.75991	0.83494	0.88032
cd	0.74920	0.82486	0.86505
ekc	0.87652	0.88499	0.89145
kaiser	0.65478	0.75597	0.82856

Note. Acc_{250} is the mean accuracy for conditions with N = 250, Acc_{500} for conditions with N=500 and Acc_{1000} for conditions with N=1000.

All factor retention criteria improved their performance with increasing sample size. Especially the simulation based approaches PA and CD as well as the Kaiser criterion strongly benefited from greater samples. EKC, on the contrary, showed almost the same performance for all three values of N (mean accuracies: 87.7%, 88.5%, 89.1%). Table 11 displays the averaged accuracies for all sample sizes separately.

The accuracy of factor retention varied for different levels of variables per factor (the item-to-factor ratio) as well as for different combinations of primary and cross-loadings. Figures 5-8 show the performance of all methods for conditions with $k = 1$ (Figure 5), $k = 2$ (Figure 6), $k = 4$ (Figure 7) and $k = 6$ (Figure 8) and all combinations of these three variables (*vpf*, primary loadings and cross-loadings).

When $k = 1$ and primary loadings were high, all methods except CD achieved almost perfect accuracy, while PA failed to retain the correct number of factors more often than all other methods when primary loadings were small. Especially when $vpf = 4$ or 5, PA yielded an averaged accuracy below 30%. The three ML models, EKC and the Kaiser criterion achieved very high accuracies throughout all respective conditions.

When $k = 2$ and primary loadings were small, the Kaiser criterion and PA performed worse than the other retention criteria. While, in general, a higher number of variables

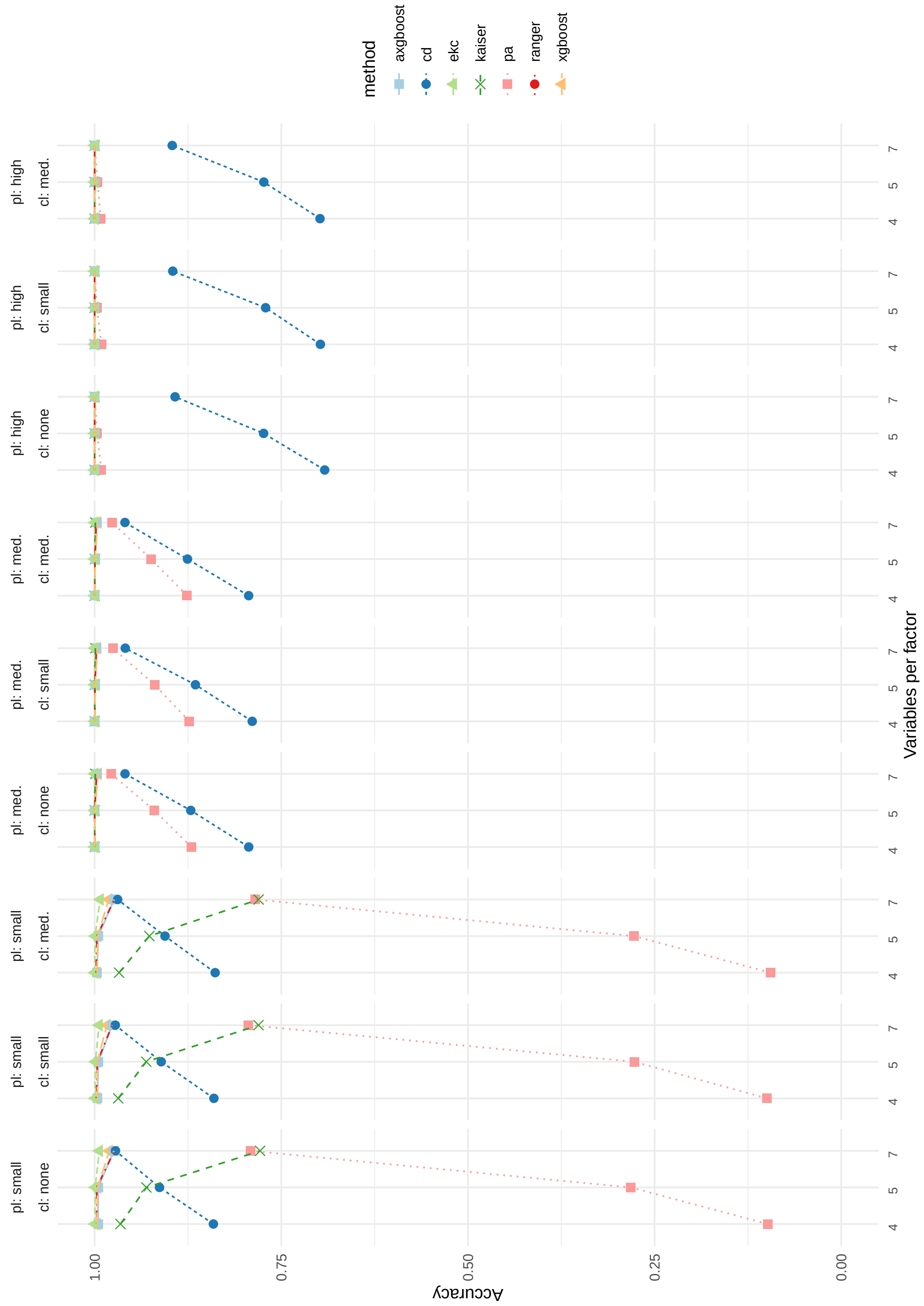

Figure 5. Study 2: Accuracy of retention criteria for conditions with one factor averaged over N.

per factor yielded better results, the Kaiser criterion and EKC showed opposing tendencies (when primary loadings were small) with worse results the more variables per factor were present. When cross-loadings were medium and primary loadings were small, all methods had averaged accuracies below 90% with EKC being the only exception when $vpf = 7$. In conditions with high primary loadings, all criteria performed reasonably well, yet CD failed to retain the true number of factors more often than the other methods (Figure 6).

Figure 7 and Figure 8, show the averaged accuracy for conditions with four and six true factors respectively. When $k = 4$ and primary loadings were small the Kaiser criterion performed quite poorly, while the empirical Kaiser criterion yielded high accuracies. The *xgboost* model was often superior to the other retention criteria, especially when cross-loadings got higher. When primary loadings were high all criteria performed better, yet the EKC showed some problems with relatively high cross-loadings and only four variables per factor. When $k = 6$ all methods lacked accuracy for conditions with small primary loadings and comparably high cross-loadings. Only the *xgboost* model reached an accuracy higher than 50% on average in these conditions. When $k = 6$, $vpf = 7$ and primary loadings were small (no cross-loadings), the Kaiser criterion was not able to retain the correct number of factors in a single data set. It rather suggested more than six factors for each case.

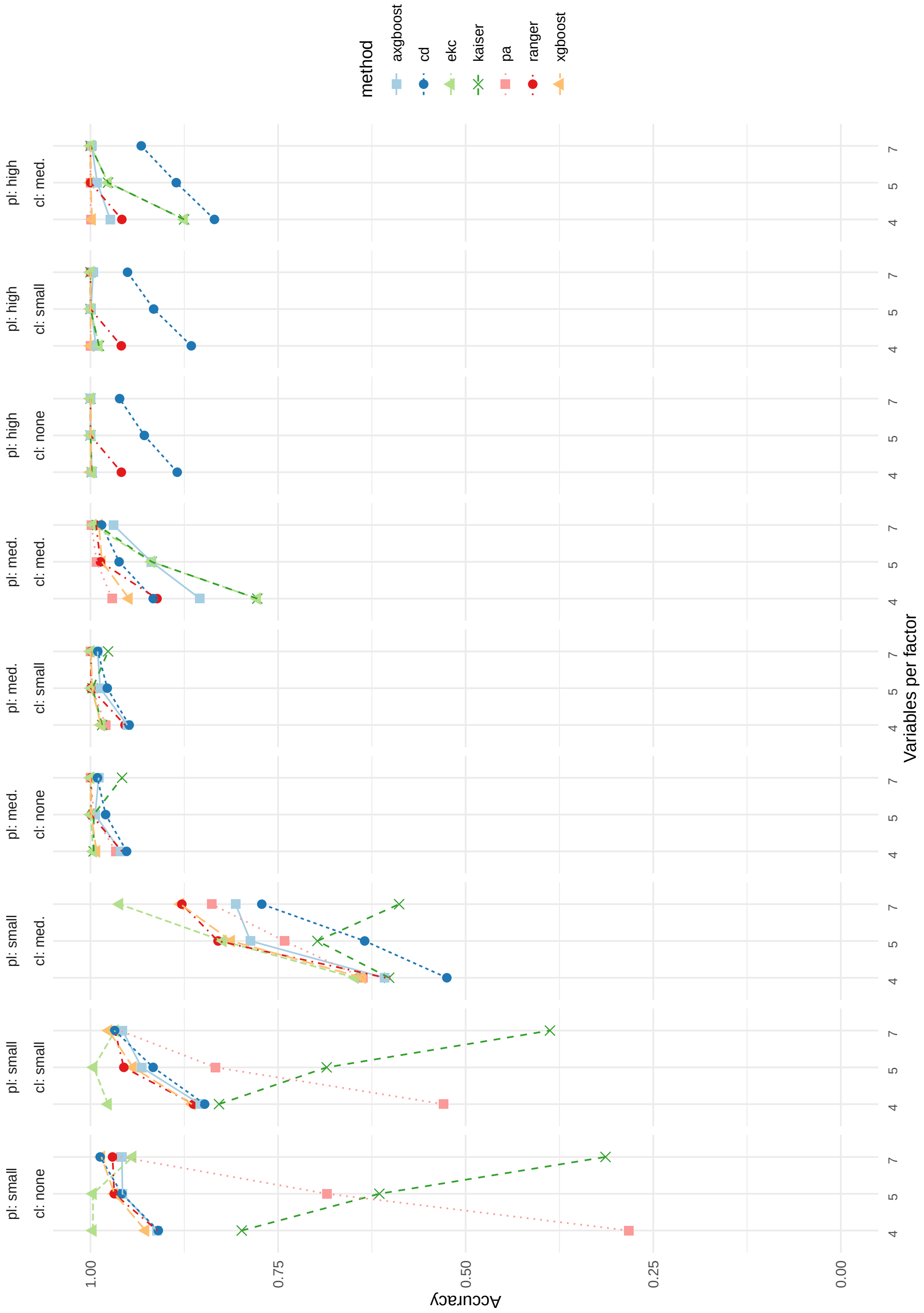

Figure 6. Study 2: Accuracy of retention criteria for conditions with two factors averaged over N and ρ.

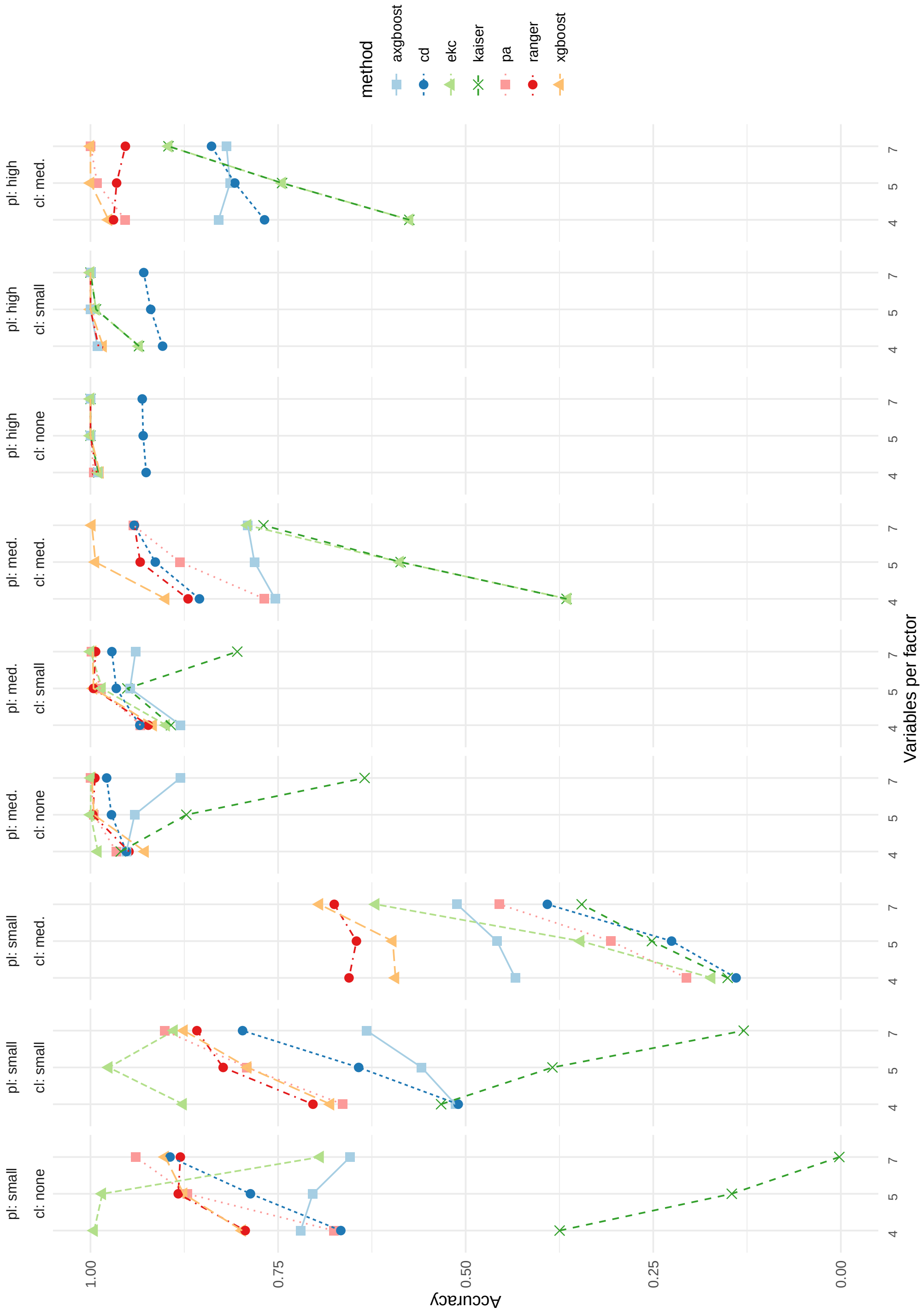

Figure 7. Study 2: Accuracy of retention criteria for conditions with four factors averaged over N and ρ.

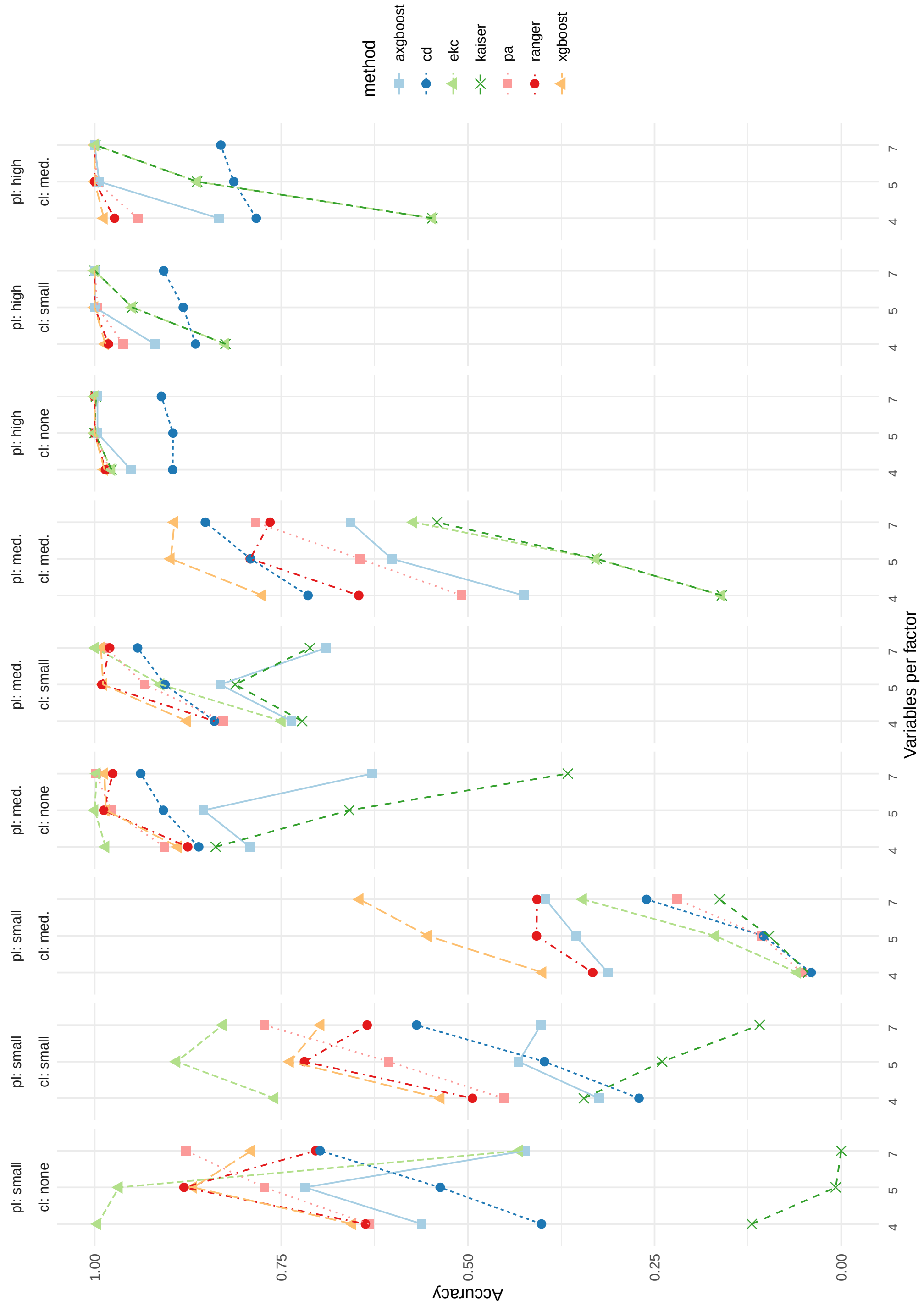

Figure 8. Study 2: Accuracy of retention criteria for conditions with six factors averaged over N and ρ.

10.4.1 Additional Conditions. As the data sets used for this evaluation were based on loading matrices containing solely positive loadings, we added 16 data conditions with both negative and positive loadings to find out whether our new approach (focusing on the *xgboost* model as it outperformed both the *ranger* and the *autoxgboost*) can handle respective data. We also evaluated the performance of the *xgboost* model compared to the four common retention criteria under a condition based on a random intercept model with ten variables loading equally on the first factor and unequally on a second factor. This random intercept condition can provide first insights on how the *xgboost* model behaves in conditions fundamentally different to conditions based on a (near) simple structure. Table 12 shows that the *xgboost* model performed quite well under these additional conditions being able to retain the correct number of factors almost in every case, while the EKC struggled with the random intercept model (1% accuracy).

Table 12

Study 2: Accuracy of Factor Retention Criteria for Conditions with Negative and Positve Loadings (A) and a Data Condition based on a Random Intercept Model (B)

N	vpf	k	ρ	primary load.	cross-load.	Acc_{xgb}	Acc_{ekc}	Acc_{cd}	Acc_{pa}	Acc_{kc}
A										
500	5	2	0.0	high	medium	1.00	1.00	0.89	1.00	1.00
500	5	2	0.0	high	none	1.00	1.00	0.94	1.00	1.00
500	5	2	0.0	low	medium	1.00	0.98	0.98	0.87	0.88
500	5	2	0.0	low	none	1.00	1.00	0.96	0.81	0.70
500	5	2	0.3	high	medium	1.00	1.00	0.85	1.00	1.00
500	5	2	0.3	high	none	1.00	1.00	0.93	1.00	1.00
500	5	2	0.3	low	medium	1.00	0.95	0.99	0.94	0.80
500	5	2	0.3	low	none	1.00	1.00	0.98	0.90	0.77
500	5	4	0.0	high	medium	1.00	1.00	0.95	1.00	1.00
500	5	4	0.0	high	none	1.00	1.00	0.86	1.00	1.00
500	5	4	0.0	low	medium	1.00	0.91	0.99	0.99	0.60
500	5	4	0.0	low	none	1.00	1.00	1.00	0.89	0.01
500	5	4	0.3	high	medium	1.00	1.00	0.93	1.00	1.00
500	5	4	0.3	high	none	1.00	1.00	0.97	1.00	1.00
500	5	4	0.3	low	medium	0.95	0.64	0.97	0.97	0.38
500	5	4	0.3	low	none	0.98	0.65	0.95	0.90	0.00
B										
500	NA	2	0.0	equal	unequal	1.00	0.01	0.96	1.00	0.88

Note. vpf = Variables per factor, k = number of factors. A: Data conditions with both negative and positive loadings. B: Data condition based on random intercept model with all variables loading equally on the first factor and unequally on the second factor.

10.4.2 Feature Importance. While common retention criteria are derived from statistical theory (e.g. EKC) or are grounded on it, the machine learning models only use components of these criteria (e.g. eigenvalues) as features and provide a prediction for the dimensionality *k* based on rather complex interaction patterns. Despite their black box character, it is possible to calculate measures of feature importance indicating the influence of features on the prediction. In case of the best performing machine learning model, the *xgboost* model, both “inequality”-measures - Kolm measure and Gini coefficient - had the highest importance followed by several corrected eigenvalues of the factor model and the averaged value of the bivariate correlations (Table 13).

Table 13

Study 2: Feature Importance: 15 Most Important Features of the Xgboost Model

Feature	Type	Importance
Kolm	inequality measure	0.248
Gini	inequality measure	0.103
fa_eigval2	eigenvalue factor model	0.085
fa_eigval3	eigenvalue factor model	0.082
fa_eigval8	eigenvalue factor model	0.081
fa_eigval4	eigenvalue factor model	0.079
fa_eigval6	eigenvalue factor model	0.072
fa_eigval7	eigenvalue factor model	0.071
fa_eigval5	eigenvalue factor model	0.045
eigval2	eigenvalue	0.041
eigval3	eigenvalue	0.011
avgcor	correlation size	0.010
N	sample	0.008
fa_eigval9	eigenvalue factor model	0.006
eigval5	eigenvalue	0.005

Note. fa_eigval2 means the second eigenvalue of the factor model, while eigval2 is the second eigenvalue of the correlation matrix. avgcor is the averaged inter-item correlation.

10.5 Tuning the Best Model

Using default settings in this study, the *xgboost* model outperformed both the random forest and all common retention criteria. Although, the in-sample accuracy (the accuracy based on the training data) had already been quite high, tuning hyperparameters of the algorithm might raise the predictive power of the model, so we decided to improve the final model in a second step. In addition to hyperparameter-tuning, the results of PA, CD and EKC were added as features for training the machine learning model in this second step as well.

Table 14

Study 2: Hyperparameter Space for Tuning

Name	lower	upper	log2_scale
eta	0.01	0.20	FALSE
gamma	-7.00	6.00	TRUE
max_depth	3.00	20.00	FALSE
colsample_bytree	0.50	1.00	FALSE
lambda	-10.00	10.00	TRUE
subsample	0.50	1.00	FALSE

Note. log2_scale means that values are transformed according to the binary logarithm. Parameter names are identical to names in xgboost implementation in R.

10.5.1 Hyperparameter Tuning. Data were simulated as described before. In total, we used 246355[37] simulated data sets to reduce computational costs. 70% of these data sets served as training data. As the *xgboost* implementation allows for tuning several hyperparameters, we used six out of eight parameters that were defined as the simple space for the *autoxgboost* algorithm (Thomas et al., 2018). Table 14 shows these six parameters with the lower and upper bounds chosen for tuning the model.

Again, we applied the *mlr* framework (Bischl et al., 2016) to train the model and to tune the six parameters. 80% of the data from the training set (137959 data sets) were used as the actual training set while 20% served as an internal test set (34489 data sets) for an early stopping rule as implemented in the *autoxgboost* algorithm (Thomas et al., 2018). We set the *early_stopping_rounds* argument of this implementation to five[38] and used the fast histogram optimized algorithm (https://github.com/dmlc/xgboost/issues/1950) as the tree construction algorithm to save computation time.

For tuning purposes, we decided not to rely on a predefined grid, but use the iterative racing procedure (*irace*) by López-Ibáñez, Dubois-Lacoste, Pérez Cáceres, Stützle and Birattari (2016). It allows for automatic parameter configuration as it samples possible pa-

[37] 250000 data sets were simulated originally, but in 3645 cases either population correlation matrices were not semi-positive definite or internal simulations of comparison data (CD approach) were causing errors.

[38] This means that when no improvement in the performance measure is indicated for five iterations the algorithm stops.

rameter configurations from iteratively updated distributions in the parameter space. With regard to the sample size of the training set and the idea that the training data were simulated as “representative” for real-world scenarios, we used a holdout set ($\frac{1}{3}$ of the actual training set which equaled 45986 data sets) for the tuning procedure.

After six iterations with 2483 so-called experiments (configurations tested), the following hyperparameter set was selected $\eta = 0.158$, $\gamma = 0.015$, $max\_depth = 4$, $colsample\_bytree = 0.789$, $\lambda = 0.005$ and $subsample = 0.812$. The tuned *xgboost* model reached an out-of-sample accuracy of 99.3% on the test set (30% of the 246355 data sets = 73907 data sets)[39].

10.6 Discussion

In this study, we present a new approach to determine the number of factors in EFA. We combined different machine learning algorithms with a large data simulation to build a model that can predict the true number of factors based on features of the empirical correlation matrix. The used *xgboost* model was able to constantly outperform the common retention criteria, even under conditions that were outside the range that the model was trained on (e.g. $\rho = 0.5$ and the random intercept model). Since the simulation study had to focus on some general data conditions in order to ensure an adequate scope, we were not able to evaluate the performance of the *xgboost* model under all potential conditions. Therefore, some condition variations, such as different numbers of variables for each factor, were not considered. However, many conditions not covered by the evaluation study were still included in the test set for the tuned *xgboost* model. Thus, the new approach should be able to deal with this kind of data. Hence, our study can be seen as a proof of concept. While all common factor retention criteria showed some tendencies of bias (i.e. over- or underfactoring) the *xgboost* model estimated the number of factors without bias (for all k). This is a great advantage of the trained model as all common retention criteria perform poorly under some circumstances which is why several authors recommend combinations of different criteria (Auerswald & Moshagen, 2019; Fabrigar et al., 1999; Goretzko et al., 2019). The machine learning models, are not theoretically founded like the EKC, for example. Nevertheless, they are able to reflect the complex relations between the number of factors

[39] We applied a slightly different importance measure to look at the feature importance of the tuned *xgboost* model. All three added factor retention criteria were among the 20 most important features and helped to improve the prediction performance.

and the data characteristics (in this case described by 181/184 features) almost perfectly as demonstrated by the tuned *xgboost* model with its out-of-sample accuracy of 99.3%.

The dependency on the simulated data basis can be seen as a weakness of the new approach. When empirical data is fundamentally different to this data basis, model predictions are probably invalid and not trustworthy. However, this study showed that the performance of the *xgboost* model was quite good in conditions not completely covered in the data basis it was trained on. In fact, when the data basis is sufficiently large and all possible data conditions are included, the machine learning models (the *xgboost* model in particular) are able to outperform all common criteria. So providing a wide-ranging training set allows us to rely on a single model rather than combining several criteria. One specific advantage of this approach is that the data basis can be extended easily and the model can be improved if necessary when specific conditions have to be considered that have been left out previously. Further research should also focus on the evaluation of the model under other data conditions (for example extending our additional analyses: conditions with different numbers of variables for each factor, more complex factor structures like the random intercept model or more conditions with negative loadings).

It might also be possible to further improve the model by adding new features (as we did in the tuned version adding the solutions of PA, CD and EKC as features) and extending the data basis for specific applications (e.g. panel data with far more factors and variables). So far, the data basis is solely based on data following a multivariate normal distribution, yet data in psychological research is often of ordinal nature, so in a next step a model trained on ordinal data has to be developed as well. The procedure can easily be applied to both ordinal data and other somehow exceptionally distributed data (for example count data from observational studies). Accordingly, the new approach provides a framework that is less dependent on distributional assumptions than other criteria (e.g. EKC and CD relying on normally distributed items).

Another advantage of this approach is the possibility to get not just an estimate for k but also a probability estimate for several values of k. With this option, the *xgboost* model provides an implicit uncertainty measure that enables the user (researcher) to assess how convincing a particular solution is. Common retention criteria, on the contrary, only return an estimate for k or in case of the Scree test an ambiguous plot that has to interpreted[40], but usually no information reflecting the estimation uncertainty (due to sampling error) is

[40] Note that there are ways to objectify this procedure, like the Cattell-Nelson-Gorsuch approach (e.g. Nasser, Benson, & Wisenbaker, 2002).

given. Our study showed that the accuracy of all factor retention criteria is influenced by such sampling error as reflected by the comparably poor performance of PA, CD and the Kaiser criterion when $N = 250$ - a sample size that is not necessarily reached in current research (Goretzko et al., 2019).

10.6.1 Understanding the Black Box. Practitioners might be bothered by the black box character of the model which hampers its interpretability. Hence, one can use tools like the local interpretable model-agnostic explanations (LIME; Ribeiro, Singh, & Guestrin, 2016) for each new empirical data set that is evaluated with the *xgboost* model. We present a short example on how this could work. For this purpose, we chose a data set containing 1369 observations of 50 items of a *BIG5*-inventory constructed by Goldberg (1990) that can be retrieved from https://openpsychometrics.org/_rawdata/ (version of 11/8/2018). Applying the tuned *xgboost* model to this data yielded six factors (estimated probability for $k = 6$ was 97.3%) while CD suggested five, EKC six, PA eight and the Kaiser-Guttman rule nine factors. Approximating the complex *xgboost* model locally with (generalized) linear models, LIME provides the best features to explain this six factor solution[41]. Nine out of the ten most important features explaining the six factor solution were different eigenvalues with the sixth eigenvalue of the factor model being greater than 0.5115 having the highest explanatory power and $p = 50 > 35$ being third (Table 15). There were also several features having negative explanatory power which means that these features and its values would speak against the final prediction of the complex model based on the simple approximation (e.g. the second eigenvalue being greater than 2.4). Approximating the complex model locally using (generalized) linear models provides insights on how the complex model estimates the number of factors k. Even though LIME might improve the interpretability of the black box model, researchers have to be cautious as the (local) approximation with simple models might not fully reflect the complex interactions among some features that indicate a specific k-factor solution (here: rather weak $r^2 = 0.235$ of the explaining model).

[41] Note that the estimated r^2 of the model used for the approximation was 0.235 in this case.

Table 15

Study 2: Explaining the Xgboost Model for Exemplary Data

Feature	Weight	Feature Value	Explanation
fa_eigval6	0.193	0.66	fa_eigval6 > 0.51
eigval6	0.076	1.52	eigval6 > 1.26
p	0.058	50.00	p > 35
fa_eigval5	0.042	1.98	fa_eigval5 > 0.81
fa_eigval4	0.036	2.44	fa_eigval4 > 1.07
eigva22	0.025	0.63	eigva22 > 0.51
eigva15	0.022	0.80	eigva15 > 0.67
eigva29	0.019	0.55	eigva29 > 0.33
frobnorm	0.014	11.70	frobnorm > 11.01
N	0.014	1,369.00	N > 799

Note. fa_eigval6 means the sixth eigenvalue of the factor model, frobnorm represents the Frobenius norm of the correlation matrix

10.6.2 Conclusion. This study shows that the new approach combining extensive data simulation and machine learning techniques to determine the number of factors provides very good results, outperforming common criteria. Based on data that cover a wide range of conditions, the new approach promises to tackle the ambiguous decision of how many factors to extract in EFA. Extending the data basis as well as the features might improve the method even further. Further research could also evaluate other ML algorithms, even though the performance of the tuned *xgboost* model seems to be tough to beat. Adaptation to ordinal data will follow, so that Likert-type data will be specifically accounted for.

10.7 References

Auerswald, M., & Moshagen, M. (2019). How to determine the number of factors to retain in exploratory factor analysis: A comparison of extraction methods under realistic conditions. *Psychological Methods*, *24*(4), 468–491. doi:10.1037/met0000200

Aust, F., & Barth, M. (2018). *papaja: Create APA manuscripts with R Markdown.* Retrieved from https://github.com/crsh/papaja

Bischl, B., Lang, M., Kotthoff, L., Schiffner, J., Richter, J., Studerus, E., … Jones, Z. M. (2016). Mlr: Machine learning in R. *The Journal of Machine Learning Research*, *17*(1), 5938–5942.

Braeken, J., & Van Assen, M. A. (2017). An empirical Kaiser criterion. *Psychological Methods*, *22*(3), 450–466. doi:10.1037/met0000074

Breiman, L. (1999). *Random forest.* Retrieved from http://machinelearning202.pbworks.com/w/file/fetch/60606349/breiman_randomforests.pdf

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32.

Browne, M. W. (2001). An overview of analytic rotation in exploratory factor analysis. *Multivariate Behavioral Research*, *36*(1), 111–150. doi:10.1207/S15327906MBR3601_05

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

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). New York, NY, USA: ACM. doi:10.1145/2939672.2939785

Chen, T., He, T., Benesty, M., Khotilovich, V., & Tang, Y. (2018). Xgboost: Extreme gradient boosting. R package version 0.6. 4.1.

De Winter, J. C., & Dodou, D. (2012). Factor recovery by principal axis factoring and maximum likelihood factor analysis as a function of factor pattern and sample size. *Journal of Applied Statistics*, *39*(4), 695–710. doi:10.1080/02664763.2011.610445

Dinno, A. (2009). Exploring the sensitivity of Horn's parallel analysis to the distributional form of random data. *Multivariate Behavioral Research*, *44*(3), 362–388.

doi:10.1080/00273170902938969

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

Fava, J. L., & Velicer, W. F. (1996). The effects of underextraction in factor and component analyses. *Educational and Psychological Measurement*, *56*(6), 907–929. doi:10.1177/0013164496056006001

Friedman, J. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, *29*(5), 1189–1232. Retrieved from https://www.jstor.org/stable/2699986

Friedman, J., Hastie, T., & Tibshirani, R. (2000). Additive logistic regression: A statistical view of boosting (with discussion and a rejoinder by the authors). *The Annals of Statistics*, *28*(2), 337–407. doi:10.1214/aos/1016218223

Genuer, R., Poggi, J.-M., & Tuleau, C. (2008). Random forests: Some methodological insights. *arXiv Preprint arXiv:0811.3619.*

Genz, A., Bretz, F., Miwa, T., Mi, X., Leisch, F., Scheipl, F., & Hothorn, T. (2018). *mvtnorm: Multivariate normal and t distributions.* Retrieved from https://CRAN.R-project.org/package=mvtnorm

Gini, C. (1921). Measurement of inequality of incomes. *The Economic Journal*, *31*(121), 124–126.

Goldberg, L. R. (1990). An alternative "description of personality": The big-five factor structure. *Journal of Personality and Social Psychology*, *59*(6), 1216–1229. doi:10.1037/0022-3514.59.6.1216

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

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, *30*(2), 179–185. doi:10.1007/BF02289447

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical*

Learning. New York, NY: Springer.

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

Kaiser, H. F. (1970). A second generation little jiffy. *Psychometrika*, *35*(4), 401–415. doi:10.1007/BF02291817

Kolm, S.-C. (1999). The rational foundations of income inequality measurement. In *Handbook of income inequality measurement* (pp. 19–100). Dordrecht, NL: Springer.

Lim, S., & Jahng, S. (2019). Determining the number of factors using parallel analysis and its recent variants. *Psychological Methods*, *24*(4), 452–467. doi:10.1037/met0000230

Lorenzo-Seva, U., Timmerman, M. E., & Kiers, H. A. L. (2011). The hull method for selecting the number of common factors. *Multivariate Behavioral Research*, *46*(2), 340–364. doi:10.1080/00273171.2011.564527

López-Ibáñez, M., Dubois-Lacoste, J., Pérez Cáceres, L., Stützle, T., & Birattari, M. (2016). The irace package: Iterated racing for automatic algorithm configuration. *Operations Research Perspectives*, *3*, 43–58. doi:10.1016/j.orp.2016.09.002

Nasser, F., Benson, J., & Wisenbaker, J. (2002). The performance of regression-based variations of the visual scree for determining the number of common factors. *Educational and Psychological Measurement*, *62*(3), 397–419. doi:10.1177/00164402062003001

Peres-Neto, P. R., Jackson, D. A., & Somers, K. M. (2005). How many principal components? Stopping rules for determining the number of non-trivial axes revisited. *Computational Statistics & Data Analysis*, *49*(4), 974–997.

R Development Core Team. (2008). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org

Revelle, W. (2018). *Psych: Procedures for psychological, psychometric, and personality research*. Evanston, Illinois: Northwestern University. Retrieved from https://CRAN.R-project.org/package=psych

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Model-agnostic interpretability of machine learning. *arXiv Preprint arXiv:1606.05386*.

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), 282–292. doi:10.1037/a0025697

Thomas, J., Coors, S., & Bischl, B. (2018). Automatic gradient boosting. *arXiv Preprint arXiv:1807.03873*.

Velicer, W. F. (1976). Determining the number of components from the matrix of partial correlations. *Psychometrika*, *41*(3), 321–327. doi:10.1007/BF02293557

Wickham, H. (2016). *Ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York. Retrieved from https://ggplot2.tidyverse.org

Wright, M. N., & Ziegler, A. (2017). ranger: A fast implementation of random forests for high dimensional data in C++ and R. *Journal of Statistical Software*, *77*(1), 1–17. doi:10.18637/jss.v077.i01

Zwick, W. R., & Velicer, W. F. (1986). Comparison of five rules for determining the number of components to retain. *Psychological Bulletin*, *99*(3), 432–442. doi:10.1037/0033-2909.99.3.432

11 Study 3

The article entitled “Two factors, or rather four? Robustness of factor solutions in exploratory factor analysis” submitted for publication is referred to as Study 3 throughout this thesis. It is presented hereinafter.

11.1 Abstract

Replicability has become a highly discussed topic in psychological research. The debates focus mainly on significance testing and confirmatory analyses, whereas exploratory analyses such as exploratory factor analysis are more or less ignored, although hardly any analysis has a comparable impact on entire research areas. Determining the correct number of factors for this analysis is probably the most crucial, yet ambiguous decision - especially since factor structures have often been not replicable. Hence, a new approach is proposed to evaluate the robustness of factor retention criteria against sampling error and to predict whether a particular factor solution may be replicable. We used three samples of the *Big Five Structure Inventory* and four samples of the *10 Item Big Five Inventory* to illustrate the relationship between stable factor solutions across bootstrap samples and their replicability. In addition, we compared four factor retention criteria in terms of their stability on the one hand and their replicability on the other. Based on this study, we want to encourage researchers to make use of bootstrapping to assess the stability of the factor retention criteria they use and to compare these criteria with regard to this stability as a proxy for possible replicability.

11.2 Introduction

In recent years, the so-called replication crisis has shaken the social sciences in general and psychology in particular (e.g. Shrout & Rodgers, 2018). Several replication projects (e.g. Aarts et al., 2015; Camerer et al., 2018) showed that many published effects cannot be replicated and urged a reform of research practices. Replicability is not only a problem within the (confirmatory) framework of hypothesis testing, which is mainly affected by p-hacking, publication bias and underpowered studies (Asendorpf et al., 2013), but also crucial for exploratory analyses that shape entire research areas. One prominent example for such an analysis is exploratory factor analysis (EFA), which is widely used to assess the dimensionality and structure of psychological constructs (Goretzko, Pham, & Bühner, 2019). Determining the number of factors that should be retained in EFA is “likely to be the most important decision a researcher will make” (Zwick & Velicer, 1986), because its implications are extremely far-reaching. The most prominent example in psychological research might be the dimensionality of personality. Although it has been widely agreed to describe personality with the five-factor model (“BIG5”, e.g. Costa & McCrae, 1992), several studies reported difficulties in replicating this structure (e.g. Thalmayer, Saucier, & Eigenhuis, 2011).

Therefore, when conducting an EFA and determining the number of factors that should be retained, the goal of replicability should be considered alongside the goal of approximating the data generating process (Preacher, Zhang, Kim, & Mels, 2013). Common factor retention criteria such as the Scree test (Cattell, 1966), the Kaiser-Guttman rule (Kaiser, 1960) and parallel analysis (PA; Horn, 1965) as well as modern approaches like the comparison data (CD) approach (Ruscio & Roche, 2012) or the empirical Kaiser criterion (EKC; Braeken & Van Assen, 2017) have been developed to primarily serve the approximation goal and focus less on the replication goal. While PA has become some kind of gold-standard for factor retention (Fabrigar, Wegener, MacCallum, & Strahan, 1999; Goretzko et al., 2019), both CD and EKC showed higher accuracies in simulation studies for some data conditions (e.g. Auerswald & Moshagen, 2019). The literature clearly lacks a focus on replicability, though, as called for by Preacher et al. (2013) or Osborne and Fitzpatrick (2012). For this reason, we want to evaluate the relationship between replicability in the context of factor retention and the robustness of common criteria against sampling error. Hence, a practical way to assess the robustness of a retention criterion’s solution is proposed - bootstrapping.

Bootstrapping is a resampling strategy that was developed to assess the uncertainty of estimates when analytical solutions are not available (Efron & Tibshirani, 1994). Transferred to the issue of replicability or robustness of factor retention criteria, this means that bootstrapping allows us to assess the influence of (small) changes in the empirical data on the outcome of these criteria. Conversely, we expect that small and/or few changes in the suggested factor solutions for different bootstrap samples will be an indicator for (closer) replicability. In addition, when comparing criteria, it may be preferable to use those that have minor differences between the bootstrap samples and thus promise more robust solutions.

11.3 Methods

To illustrate how to use bootstrapping for the evaluation of the robustness of factor retention criteria, we used three different samples of the *Big Five Structure Inventory* (BFSI, Arendasy, 2009) that were provided by Stachl et al. (2018) and collected within the *Phonestudy* project (first data set: Schoedel et al., 2018; second data set: Schuwerk, Kaltefleiter, Au, Hoesl, & Stachl, 2019; third data set: Stachl et al., 2017) and four samples of the *10 Item Big Five Inventory* (BFI-10; Rammstedt, Kemper, Klein, Beierlein, & Kovaleva, 2017) that were collected within the *GESIS* panel (GESIS, 2018). The BFSI consists of 300 items that measure the typical five factors (*openness*, *emotional stability/ neuroticism*, *extraversion*, *conscientiousness* and *agreeableness*), which can be described by six facets each. We evaluated the 60 items assigned to each factor separately focusing on the dimensionality of the respective trait (e.g. determining how many facets can be found for *extraversion*). Contrary, the BFI-10 consists of 10 items also measuring these five factors without further facets. Accordingly, we evaluated the dimensionality of the questionnaire as a whole and applied the retention criteria to all ten items.

The first sample of the BFSI contains $N = 312$ observations, the second sample of the BFSI counts $N = 256$ observations and the third sample has $N = 120$ observations. In case of the BFI-10, we have one set of participants, that were asked to fill out the questionnaire four times (waves *bd*,*cd*,*dd*,*ed* of the panel), so our four samples predominantly consists of the same persons (sample sizes are $N_1 = 4888$, $N_2 = 4249$,$N_3 = 3797$,$N_4 = 3448$ using only complete cases of the BFI-10 items in each wave).

11.3.1 Data Analysis. For all 19 data sets (four BFI-10 samples and three BFSI samples with five factors each) we assessed the dimensionality with PA (default settings in the *psych* package in R [Revelle, 2018] using the 95% quantile of the random eigenvalue distribution and the *Minres* algorithm as extraction method), CD (with default settings: $\alpha = 0.30$ for the internal Mann-Whitney-U tests and 500 simulated data sets for the "comparison" approach) and EKC as well as a new machine learning approach - a tuned xgboost model (for the tuned XGB model, see Goretzko & Bühner, 2019; for the general xgboost implementation, see Chen & Guestrin, 2016; Chen, He, Benesty, Khotilovich, & Tang, 2018). Afterwards 100 bootstrap samples were drawn (using the *boot* package, Canty & Ripley, 2019) for each data set and all four factor retention criteria were applied to each of these bootstrap samples. We compared the range of proposed solutions between data sets and between retention criteria, and evaluated whether robust solutions (less fluctuation in bootstrap samples) were promising with regard to the replication purpose. We used each wave of the panel data as a replication data set for the previous one. In the case of the BFSI, the second data set ($N = 256$) was used as the replication data set for the first ($N = 312$) and the third data set ($N = 120$) was used as the replication data set for the second.

We used R (Version 3.5.1; R Core Team, 2018) and the R-packages *automatic* (Lang et al., 2014), *data.table* (Version 1.11.8; Dowle & Srinivasan, 2018), *ggplot2* (Version 3.1.0; Wickham, 2016), *mlr* (Version 2.13; Bischl et al., 2016, 2017), *mlrmbo* (Bischl et al., 2017), *multilabel* (Probst, Au, Casalicchio, Stachl, & Bischl, 2017), *openml* (Casalicchio et al., 2017), *papaja* (Version 0.1.0.9842; Aust & Barth, 2018), and *ParamHelpers* (Version 1.11; Bischl et al., 2018) for all our analyses and the preparation of the manuscript.

11.4 Results

11.4.1 BFI-10. The application of the four retention criteria (XGB, PA, CD and EKC) to the four BFI-10 data sets mostly yielded one-factor solutions. XGB, CD and EKC suggested one factor in all four cases, while PA proposed three factors for the second BFI-10 data set and two factors for the fourth empirical data set. Moreover, EKC and XGB provided one factor solutions for all 100 bootstrap samples of all four original data sets ($4 * 100$ data sets), whereas CD did so in 94%, 98%, 96% and 95% of the cases. PA had the highest volatility among the bootstrapped samples and contradicted its solution when comparing the original data set with the bootstrapped samples (e.g. for just 13% of the bootstrap samples of the second data set, PA suggested a one-factor solution, as it did for the original data set, but suggested three factors 48 times). Table 16 shows the solutions of the four retention criteria for the four initial BFI-10 data sets as well as summary statistics for the respective bootstrap samples.

Table 16

Study 3: Solutions of the Four Retention Criteria, Bootstrapped Means and Standard Deviations as well as Percentages of Bootstrap Samples with the Same Factor Solution as the Respective Empirical BFI-10 Data Set

Criterion	BFI1	BFI2	BFI3	BFI4	M_{BFI1}	SD_{BFI1}	$\%_{BFI1}$	M_{BFI2}	SD_{BFI2}	$\%_{BFI2}$	M_{BFI3}	SD_{BFI3}	$\%_{BFI3}$	M_{BFI4}	SD_{BFI4}	$\%_{BFI4}$
XGB	1	1	1	1	1.00	0.000	100	1.00	0.000	100	1.00	0.000	100	1.00	0.000	100
PA	3	1	2	1	2.98	0.887	40	2.41	0.753	48	2.49	0.689	44	2.07	0.868	29
CD	1	1	1	1	1.07	0.293	94	1.03	0.222	98	1.05	0.261	96	1.07	0.326	95
EKC	1	1	1	1	1.00	0.000	100	1.00	0.000	100	1.00	0.000	100	1.00	0.000	100

Note. BFI1 means the BFI-10 data set of the first wave (bd) in the panel, BFI2 the second BFI-10 data set (wave cd) and so on.

11.4.2 BFSI. Since the three BFSI data sets consisted of far fewer observations $(312; 256; 120)$, yet more variables ($p = 60$ compared to $p = 10$ in case of the BFI-10), the factor retention results were considerably more volatile than the results for the BFSI-10 data. Mostly six facets per factor were suggested, but the results varied according to the retention criterion, the data set and the big 5 factor. EKC and XGB tended to show fewer differences between the bootstrapped solutions, whereas PA or CD yielded the highest variance (or standard deviation) between the bootstrap samples for all combinations of data sets and factors.

Table 17 illustrates the relationship between this dispersion and the likelihood of replicability, as CD and PA tended to suggest different factor solutions across the three data sets more often than XGB and EKC.

Table 17

Study 3: Solutions of the Four Retention Criteria, Bootstrapped Means and Standard Deviations as well as Percentages of Bootstrap Samples with the Same Factor Solution as the Respective Empirical BFSI Data Set

	BFSI1	BFSI2	BFSI3	M_{BFSI1}	SD_{BFSI1}	$\%_{BFSI1}$	M_{BFSI2}	SD_{BFSI2}	$\%_{BFSI2}$	M_{BFSI3}	SD_{BFSI3}	$\%_{BFSI3}$
Openness												
XGB	6	6	6	7.09	0.975	43	6.96	0.898	42	6.42	0.684	69
PA	8	7	6	9.62	0.736	4	8.67	0.817	4	8.02	0.932	0
CD	6	6	4	7.00	1.214	25	6.77	1.262	24	5.67	1.544	11
EKC	6	5	4	6.16	0.545	74	5.90	0.577	22	5.36	0.732	8
Conscientiousness												
XGB	6	6	6	7.07	0.998	46	6.94	0.962	49	6.69	0.940	64
PA	8	6	4	11.09	1.055	0	9.47	1.049	0	7.66	1.165	0
CD	5	6	1	4.65	2.320	16	5.62	2.019	17	3.20	2.020	38
EKC	4	4	3	4.92	0.734	30	4.94	0.528	17	4.18	0.716	10
Extraversion												
XGB	6	7	6	7.06	0.952	42	7.36	0.835	18	6.65	0.809	56
PA	8	7	6	9.24	1.046	20	8.17	0.829	20	7.14	0.921	25
CD	6	7	6	6.56	1.380	23	6.16	1.791	27	5.47	1.956	32
EKC	5	5	4	5.23	0.529	67	5.16	0.662	63	4.85	0.609	27
Neuroticism												
XGB	6	7	6	7.08	0.907	37	6.93	0.935	13	6.11	0.399	92
PA	8	8	6	11.90	1.275	0	9.97	1.087	8	9.01	1.259	2
CD	4	7	4	6.07	2.016	22	6.55	1.777	20	4.61	1.524	33
EKC	5	6	4	6.40	0.711	6	6.00	0.636	63	5.11	0.827	26
Agreeableness												
XGB	7	6	6	7.05	0.977	14	6.69	0.929	63	6.28	0.653	83
PA	7	8	6	10.90	1.210	0	9.11	0.909	22	8.69	1.116	1
CD	7	5	4	6.62	1.482	15	6.08	1.637	34	5.35	1.321	18
EKC	5	4	4	6.15	0.642	13	5.33	0.697	12	5.08	0.720	19

Note. BFSI1 is first BFSI data set of Schoedel et al., 2018, while BFSI2 is the data set of Schuwerk et al., 2019.

11.4.3 Robustness and Reproducibility. We used a generalized linear model (GLM; Nelder & Wedderburn, 1972) with binomial family and logit link to model the probability of replicating the number of factors when comparing the results of the first BFI-10 data set with the results of the second, the results of the second with those of the third, the results of the third with those of the fourth, and the same for the three BFSI data sets. The standard deviation of the suggested number of factors of the respective 100 bootstrap samples as well as the percentage of bootstrap solutions being equal to the outcome of the initial data set (referred to as the *rate of consistency*) served as independent variables in our model. Both the standard deviation and this *rate of consistency* can be seen as indicators of the robustness of the proposed factor solution and also as a measure of confidence for the EFA user. The absolute difference in the suggested number of factors between two consecutive data sets served as a second measure of the "replicability" of the proposed factor solutions. A second GLM with Poisson family and log link was used for this dependent variable analogous to the first model with the standard deviation and the *rate of consistency* as independent variables.

The results of the GLM analyses support the descriptive observations that retention criteria, that were more stable across bootstrap samples, were more likely to yield replicable results. With respect to exact replication (first GLM), higher standard deviations for the suggested number of factors across the bootstrap samples were associated with a lower probability of replication ($b = -0.82$, 95% CI [-3.39, 1.24], $z = -0.72$, $p = .471$), whereas the percentage of bootstrap samples with the same solution as the initial data set was positively linked to this probability ($b = 0.04$, 95% CI [0.01, 0.08], $z = 2.44$, $p = .015$)[42]. We modeled the difference of the suggested number of factors between two consecutive data sets (e.g. BFI-10 of the first and the second wave of the panel) with the second GLM. Again, the higher the standard deviations for the proposed number of factors across the bootstrap samples were, the less accurate the replication was - illustrated here by a positive association with the dependent variable ($b = 0.79$, 95% CI [0.22, 1.33], $z = 2.82$, $p = .005$). With an increasing *rate of consistency*, a smaller deviation of the proposed number of factors from two consecutive data sets was associated ($b = -0.02$, 95% CI [-0.03, 0.00], $z = -2.11$, $p = .035$).

[42] Significance testing of parameters within the GLM are of little meaning in this case, as the number of observations and thus the statistical power is quite low. We therefore consider these analyses as rather descriptive.

Table 18

Study 3: Means and Medians of Standard Deviations and Rates of Consistency *over all Data Sets for the Four Retention Criteria as well as the Means of both Replicability Measures (Dependent Variables of the GLM Analyses)*

Retention Criterion	M_{SD}	Md_{SD}	$M_{\%}$	$Md_{\%}$	$\%_{Replicable}$	$M_{abs.Difference}$
XGB	0.721	0.929	51.31	43	61.54	0.385
PA	0.949	0.909	16.15	8	7.69	1.308
CD	1.360	1.482	39.31	24	30.77	1.462
EKC	0.482	0.577	51.31	63	46.15	0.615

11.4.4 Comparing the Criteria. Although it is not the focus of this article, both the standard deviation of the bootstrap results and the *rate of consistency* can be used to compare the retention criteria with regard to their robustness against sampling errors. While in the case of the BFI-10 data, EKC and XGB had a *rate of consistency* of 100% and thus no variance in the bootstrap results, all criteria were much more volatile for the $BFSI$ data sets, which can be explained by the far smaller sample sizes and the higher number of items ($p = 60$ vs. $p = 10$).

EKC provided the most robust results (smallest mean and median standard deviation as well as highest mean and median *rates of consistency*). In terms of replicability, however, XGB yielded better results on average (highest replicability rate with 61.54% and the smallest mean absolute difference between the number of factors for consecutive data sets: 0.38). PA had the lowest mean and median *rate of consistency*, which is reflected in the worst replicability rate of 7.69%. CD yielded the most volatile results (highest mean and median standard deviation across the bootstrap samples), which can be linked to the highest mean absolute difference of the proposed number of factors between consecutive data sets (especially caused by the facet *conscientiousness* of the BFSI data sets, see table 16). Table 18 provides an overview of these robustness and replicability measures for the four retention criteria.

11.5 Discussion

The present study examines the relationship between the robustness of factor retention criteria and the replicability of their solutions. Bootstrapping of the initial empirical data sets is proposed as an easy-to-use method to evaluate the robustness of the factor retention process and to provide a proxy for replicability. The study results show some promising patterns, since criteria in specific cases with high robustness tended to show higher replicability rates and provided more consistent results across the data sets that were used for replication.

Higher robustness and replicability rates were recorded for the BFI-10 panel data, which can be explained by the much larger sample sizes compared to the BFI data. Several authors discussed this relationship between robustness and sample sizes for EFA in general (e.g. Osborne & Fitzpatrick, 2012) and various simulation studies showed the need for larger samples to achieve a higher accuracy/precision in EFA (see Goretzko et al., 2019 for an overview; and MacCallum, Widaman, Zhang, & Hong, 1999 for a comprehensive simulation study). Several studies (e.g. Auerswald & Moshagen, 2019) found that retention criteria consistently perform better at higher sample sizes and although these studies predominantly focus on the approximation goal and not on the replication goal, it seems reasonable to assume that higher sample sizes also benefit the replicability of factor retention criteria since the impact of of sampling error decreases with increasing sample sizes.

Comparing the retention criteria, EKC and XGB provided more robust and replicable results on average than PA and CD. These advantages with regard to the replicability goal are in line with the higher overall accuracy by both XGB and EKC in a simulation study of Goretzko and Bühner (2019)[43]. Although we do not know the true dimensionality, since this study is based on empirical data, the result patterns strengthen confidence in the suggested number of factors provided by XGB and EKC[44] rather than in the solutions PA and CD produced.

The study should be considered purely descriptive, as the number of observations for the GLM analyses is rather small ($N = 52$). As mentioned in the footnote above, this small number leads to an insufficient statistical power[45] and does not allow cross-validation. Nonetheless, from a descriptive point of view, a positive relationship can be assumed between the robustness and the replicability of factor retention criteria. Both the face validity (regarding the result patterns in table 16 and table 17) and the signs of GLM parameter estimates that met our expectations are indicators that robustness and

[43] The referenced manuscript is Study 2 of this thesis.

[44] The results of XGB seem to be more in line with the theoretical assumptions of the BFSI - namely six facets per factor - than the results of the EKC.

[45] With an α of five percent, three out of four coefficients of interest would be classified as significant anyway. However, this does not mean that the true effects are necessarily large enough, so that our power was sufficiently high. We therefore refrain from interpreting the hypothesis tests for the GLM coefficients.

replicability are positively related.

The empirical data sets had quite different characteristics (BFI-10 data with great N and small p and BFSI data with small N and rather large p), which differed particularly in the type of replication context. The panel data (BFI-10 data) consists of the same participants, making it a within-person replication scenario, while in the BFSI data sets different people were collected, making these evaluations a between-person replication. Therefore, one can presume that the established relation between robustness and replicability of factor retention criteria can be found in various data conditions.

11.6 Conclusion

The present study demonstrates a positive relation between the robustness of factor retention criteria and the replicability of their solutions. Using bootstrap samples of the empirical data set, it is possible to evaluate the robustness of a given solution, either by looking at the standard deviation of the bootstrap solutions or by computing the *rate of consistency* (as described above). We want to encourage researchers to include bootstrapping in their analyses, as individual point estimates of the number of factors based on one empirical data set do not reflect the uncertainty of this estimate and the possible vulnerability to sampling error. This idea aims in the same direction as splitting the empirical data set and evaluating the factor retention criteria on both subsets in order to gain confidence in the stability of the proposed factor solution (Fabrigar et al., 1999; Goretzko et al., 2019). Relying on bootstrapped samples instead of splitting the empirical data can be beneficial for small samples (as demonstrated for the BFSI data in this study). When evaluating the robustness of the criteria, a comparison among them is imperative, because the stability measures cannot be interpreted absolutely (unless all bootstrap samples provide the same solution, then the standard deviation would be 0 and the rate of consistency would be 100%). Both Fabrigar et al. (1999) and Goretzko et al. (2019) recommend comparing methods and also considering combinations of criteria as suggested by Auerswald and Moshagen (2019). Ultimately, the users of EFA should not only focus on the goal of approximation, but also consider the goal of replication, where bootstrapping and the evaluation of the robustness of factor solutions might be a good starting point.

11.7 References

Aarts, A., Anderson, J., Anderson, C., Attridge, P., Attwood, A., Axt, J., … others. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), 943–950. doi:10.1126/science.aac4716

Arendasy, M. (2009). *BFSI: Big-Five Struktur-Inventar (Test & Manual).* Mödling: SCHUHFRIED GmbH.

Asendorpf, J. B., Conner, M., De Fruyt, F., De Houwer, J., Denissen, J. J., Fiedler, K., … others. (2013). Recommendations for increasing replicability in psychology. *European Journal of Personality*, *27*(2), 108–119. doi:10.1002/per.1919

Auerswald, M., & Moshagen, M. (2019). How to determine the number of factors to retain in exploratory factor analysis: A comparison of extraction methods under realistic conditions. *Psychological Methods*, *24*(4), 468–491. doi:10.1037/met0000200

Aust, F., & Barth, M. (2018). *papaja: Create APA manuscripts with R Markdown.* Retrieved from https://github.com/crsh/papaja

Bischl, B., Lang, M., Kotthoff, L., Schiffner, J., Richter, J., Studerus, E., … Jones, Z. M. (2016). mlr: Machine learning in r. *Journal of Machine Learning Research*, *17*(170), 1–5. Retrieved from http://jmlr.org/papers/v17/15-066.html

Bischl, B., Lang, M., Richter, J., Bossek, J., Horn, D., & Kerschke, P. (2018). *ParamHelpers: Helpers for parameters in black-box optimization, tuning and machine learning.* Retrieved from https://CRAN.R-project.org/package=ParamHelpers

Bischl, B., Richter, J., Bossek, J., Horn, D., Thomas, J., & Lang, M. (2017). MlrMBO: A modular framework for model-based optimization of expensive black-box functions. *arXiv Preprint arXiv:1703.03373.*

Braeken, J., & Van Assen, M. A. (2017). An empirical Kaiser criterion. *Psychological Methods*, *22*(3), 450–466. doi:10.1037/met0000074

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T.-H., Huber, J., Johannesson, M., … others. (2018). Evaluating the replicability of social science experiments in nature and science between 2010 and 2015. *Nature Human Behaviour*, *2*(9), 637–644. doi:10.1038/s41562-018-0399-z

Canty, A., & Ripley, B. D. (2019). *Boot: Bootstrap R (S-Plus) functions.* Retrieved from https://cran.r-project.org/web/packages/boot/boot.pdf

Casalicchio, G., Bossek, J., Lang, M., Kirchhoff, D., Kerschke, P., Hofner, B., … Bischl, B. (2017). OpenML: An r package to connect to the machine learning platform openml. *Computational Statistics*, 1–15.

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

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). New York, NY, USA: ACM. doi:10.1145/2939672.2939785

Chen, T., He, T., Benesty, M., Khotilovich, V., & Tang, Y. (2018). Xgboost: Extreme gradient boosting. R package version 0.6. 4.1.

Costa Jr, P. T., & McCrae, R. R. (1992). Four ways five factors are basic. *Personality and Individual Differences*, *13*(6), 653–665. doi:10.1016/0191-8869(92)90236-I

Dowle, M., & Srinivasan, A. (2018). *Data.table: Extension of 'data.frame'*. Retrieved from https://CRAN.R-project.org/package=data.table

Efron, B., & Tibshirani, R. J. (1994). *An introduction to the bootstrap.* Boca Raton, FL: CRC Press.

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

GESIS. (2018). GESIS Panel - Standard Edition (Version 25.0.0, Data file ZA5665). GESIS Data Archive: Cologne. doi:10.4232/1.13158

Goretzko, D., & Bühner, M. (2019). One model to rule them all? Using machine learning algorithms to determine the number of factors in exploratory factor analysis. *Manuscript Submitted for Publication.*

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

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psy-*

chometrika, *30*(2), 179–185. doi:10.1007/BF02289447

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

Lang, M., Kotthaus, H., Marwedel, P., Weihs, C., Rahnenfuehrer, J., & Bischl, B. (2014). Automatic model selection for high-dimensional survival analysis. *Journal of Statistical Computation and Simulation*, *85*(1), 62–76.

MacCallum, R. C., Widaman, K. F., Zhang, S., & Hong, S. (1999). Sample size in factor analysis. *Psychological Methods*, *4*(1), 84–89. doi:10.1037/1082-989X.4.1.84

Nelder, J. A., & Wedderburn, R. W. (1972). Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, *135*(3), 370–384. doi:10.2307/2344614

Osborne, J. W., & Fitzpatrick, D. C. (2012). Replication analysis in exploratory factor analysis: What it is and why it makes your analysis better. *Practical Assessment, Research & Evaluation*, *17*(14/15), 1–8.

Preacher, K. J., Zhang, G., Kim, C., & Mels, G. (2013). Choosing the optimal number of factors in exploratory factor analysis: A model selection perspective. *Multivariate Behavioral Research*, *48*(1), 28–56. doi:10.1080/00273171.2012.710386

Probst, P., Au, Q., Casalicchio, G., Stachl, C., & Bischl, B. (2017). Multilabel classification with r package mlr. *arXiv Preprint arXiv:1703.08991.*

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

Rammstedt, B., Kemper, C. J., Klein, M. C., Beierlein, C., & Kovaleva, A. (2017). A short scale for assessing the big five dimensions of personality: 10 item Big Five Inventory (BFI-10). *Methods, Data, Analyses*, *7*(2), 233–249. doi:10.12758/mda.2013.013

Revelle, W. (2018). *Psych: Procedures for psychological, psychometric, and personality research.* Evanston, Illinois: Northwestern University. Retrieved from https://CRAN.R-project.org/package=psych

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

chological Assessment, *24*(2), 282–292. doi:10.1037/a0025697

Schoedel, R., Au, Q., Völkel, S. T., Lehmann, F., Becker, D., Bühner, M., … Stachl, C. (2018). Digital Footprints of Sensation Seeking. *Zeitschrift Für Psychologie*, *226*(4), 232–245. doi:10.1027/2151-2604/a000342

Schuwerk, T., Kaltefleiter, L. J., Au, J.-Q., Hoesl, A., & Stachl, C. (2019). Enter the wild: Autistic traits and their relationship to mentalizing and social interaction in everyday life. *Journal of Autism and Developmental Disorders*, 1–16. doi:10.1007/s10803-019-04134-6

Shrout, P. E., & Rodgers, J. L. (2018). Psychology, science, and knowledge construction: Broadening perspectives from the replication crisis. *Annual Review of Psychology*, *69*(1), 487–510. doi:10.1146/annurev-psych-122216-011845

Stachl, C., Hilbert, S., Au, J.-Q., Buschek, D., De Luca, A., Bischl, B., … Bühner, M. (2017). Personality traits predict smartphone usage. *European Journal of Personality*, *31*(6), 701–722. doi:10.1002/per.2113

Stachl, C., Schoedel, R., Au, Q., Völkel, S., Buschek, D., Hussmann, H., & Bühner, M. (2018). The phonestudy project. *Open Science Framework.* doi:10.17605/OSF.IO/UT42Y

Thalmayer, A. G., Saucier, G., & Eigenhuis, A. (2011). Comparative validity of brief to medium-length big five and big six personality questionnaires. *Psychological Assessment*, *23*(4), 995–1009. doi:10.1037/a0024165

Wickham, H. (2016). *Ggplot2: Elegant graphics for data analysis.* Springer-Verlag New York. Retrieved from http://ggplot2.org

Zwick, W. R., & Velicer, W. F. (1986). Comparison of five rules for determining the number of components to retain. *Psychological Bulletin*, *99*(3), 432–442. doi:10.1037/0033-2909.99.3.432