Behaviormetrika
https://doi.org/10.1007/s41237-021-00152-w

NOTE

Robustness of factor solutions in exploratory factor analysis

David Goretzko[1] · **Markus Bühner**[1]

Received: 28 April 2021 / Accepted: 12 September 2021

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, an approach based on bootstrapping the factor retention process 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 and an information criterion 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.

Keywords Factor retention · Replication · Bootstrapping · Replication crisis

Communicated by Michio Yamamoto.

✉ David Goretzko
david.goretzko@psy.lmu.de

[1] Department of Psychology, Ludwig Maximilians University Munich, Leopoldstr. 13, 80802 Munich, Germany

Published online: 22 September 2021

1 Introduction

In recent years, the so-called replication crisis has shaken the social sciences in general and psychology in particular (e.g., Shrout and 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 et al. 2019) and plays a major role in questionnaire development and test construction. Determining the number of factors that should be retained in EFA is "likely to be the most important decision a researcher will make" (Zwick and 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 Jr and McCrae 1992), several studies reported difficulties in replicating this structure (e.g., Thalmayer et al. 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 et al. 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 and Roche 2012) or the empirical Kaiser criterion (EKC; Braeken and 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 et al. 1999; Goretzko et al. 2019), both CD and EKC showed higher accuracies in simulation studies for some data conditions (Auerswald and Moshagen 2019). The EFA 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 errors. Hence, a practical way to assess the robustness of a retention criterion's solution is proposed—bootstrapping. Bootstrapping have been already used in the context of structural equation modeling to estimate standard errors of parameters (Nevitt and Hancock 2001), to evaluate model fit statistics (Hancock and Liu 2012) or to get corrected p-values for the model test in cases where multivariate normality is violated (Nevitt and Hancock, 2001). Zientek and Thompson (2007) suggested to use bootstrapping in the context of EFA as well to obtain more stable results or to evaluate the replicability. As they did not focus on determining the number of factors (in fact they applied a bootstrapped Kaiser criterion, that has been shown to overestimate the number of factors, e.g., Jackson 1993), we want to evaluate

the usefulness of bootstrapping to the issue of factor retention in this study more closely.

1.1 Replicability and robustness

Throughout this paper, we discuss the replicability of the factor retention process and the robustness of different factor retention criteria. The latter does not refer to the term of robust statistics (robustness against outliers and distributional assumptions, for further readings, see Huber 1981), but rather describes the ability of the method to neglect noise (sampling error) and to provide estimates (here: the suggested number of factors) that do not change with minor, not important changes in the data (in the respective sample).

Replicability is understood in its narrowest sense in this article—the number of factors found in one sample should be replicated in another sample based on the same population (in this study we evaluate both a within-person replication assessing the dimensionality at different time points and a between-person replication comparing several samples from the same population). When it comes to the replicability of factor structures across populations (e.g., cross-cultural research, where measurement invariance is needed), a broader definition of replicability is used. We focus on the replicability of the number of factors among samples of the same population since it is necessary that a factor retention criterion provides replicable solutions in this narrow sense as a basis for broader replication as well. In other words, if a factor structure (or to be more precise the number of factors) is not replicable in samples from the same population, it will not be replicable across populations.

1.2 Factor retention criteria

In our study, we use four different factor retention criteria, PA, CD, EKC and a new machine learning approach- a tuned *xgboost* model (XGB; for the *xgboost* implementation, see Chen and Guestrin 2016; Chen et al. 2018; for the tuned XGB model, see Goretzko and Bühner 2020) as well as the Bayesian Information Criterion (BIC; Schwarz 1978) for comparison.

1.2.1 Parallel analysis

PA (Horn 1965) is based on a comparison of the empirical eigenvalues of the correlation matrix with eigenvalues of simulated or resampled data (for a comparison of these different implementations, see Lim and Jahng 2019). The traditional version of PA, for example, is based on the comparison of the empirical eigenvalues and the mean of S eigenvalues, where S is the number of simulated data sets. The first empirical eigenvalue is compared to the mean of these S first eigenvalues, the second empirical eigenvalue is compared to the mean of the S second eigenvalues and so on. PA suggests to retain factors as long as the empirical eigenvalue is greater than the reference eigenvalue.

1.2.2 Empirical Kaiser criterion

EKC (Braeken and van Assen 2017) is a descendant of the Kaiser-Guttman rule or the Eigenvalue-greater-one rule (Kaiser 1960). Instead of comparing the empirical eigenvalues to one (and retaining all factors for which the associated eigenvalue is greater than one), EKC takes the sample size N, the number of manifest variables p and the strength of the other factors into account when calculating reference eigenvalues l_j^{REF} for an empirical eigenvalue λ_j:

$$l_j^{REF} = \max\left[\frac{p-\sum_{k=0}^{j-1}\lambda_k}{p-j+1}\left(1+\sqrt{\frac{p}{N}}\right)^2, 1\right] with\ \lambda_0 = 0.$$

EKC suggests to retain as many factors as there are eigenvalues greater than their respective reference eigenvalues l_j^{REF}.

1.2.3 Comparison data

CD (Ruscio and Roche 2012) is a variant of PA that integrates the simulation of comparison data (comparable to the simulated data in classical PA) and the model comparison perspective of structural equation modeling. For each possible number of factors, a population is simulated whose correlation matrix reproduces the empirical correlation matrix as closely as possible. Then numerous samples (the comparison data sets) are drawn from these populations and the root-mean-squared error (RMSE) between the sample eigenvalues and the empirical eigenvalues is calculated. Accordingly, if 100 samples are drawn, 100 RMSE values are calculated. The RMSE values of a one-factor model are compared to those of a two-factor model using a non-parametric Mann–Whitney-U significance test. This way n-factor models are compared to (n + 1)-factor models until no significant improvement (with regard to the RMSE values) is detected and n factors are retained.

1.2.4 Machine learning approach

XGB (Goretzko and Bühner 2020) is a completely new approach to the issue of factor retention combining data simulation and machine learning (ML) modeling. The idea of this method is to simulate various data sets that reflect all important data conditions of an application context and to calculate features for these simulated data sets that may be related to the dimensionality of the underlying factor structure such as eigenvalues or matrix norms of the correlation matrix. Since, the true number of factors is known for the simulated data sets, it is then possible to train a ML model that "learns" the relation between the extracted features and the number of factors and that can therefore be used to predict the dimensionality in EFA. Goretzko and Bühner (2020) provided a trained *xgboost* model that is able to predict the number of factors based on 184 different features (the *xgboost* algorithm is a rather complex ML algorithm that consists of numerous decision trees that are subsequently fitted to the residuals of the previous trees [idea of boosting simplified]

and that has many hyperparameters which influence the way these trees are grown and averaged, for more details, see Chen and Guestrin 2016).

1.3 Bootstrapping

Bootstrapping is a resampling strategy that was developed to assess the uncertainty of estimates when analytical solutions are not available (Efron and Tibshirani 1994). The ordinary non-parametric bootstrap is based on repeated case resampling. A bootstrap sample is created by drawing n_{obs} (which is the number of observations in the respective data set) times from the empirical data with replacement which means that every observation i ($i \in \{1, ..., n_{obs}\}$) has the same chance to be in a particular bootstrap sample multiple times, but not every observation will be drawn in each sample.[1] When repeating this procedure B times, one obtains B different bootstrap samples that consist of different subsets of the empirical data. Based on these B bootstrap samples, it is now possible to estimate a parameter of interest B times which yields B estimates (an empirical distribution of the estimation function) that can be used, for example, to quantify the standard error of this estimate.

Assuming that we are interested in an estimate of the population mean of a certain variable and we have four observations of this variable ($x^\top = (1,2,3,4)$), then our point estimate would be $\bar{x} = 2.5$. Using the five bootstrap samples $x_1^{*\top} = (2,2,1,1)$, $x_2^{*\top} = (3,1,4,3)$, $x_3^{*\top} = (4,4,4,2)$, $x_4^{*\top} = (1,2,3,3)$ and $x_5^{*\top} = (2,4,3,1)$, we obtain five estimates that can be used to build a 60%-confidence interval ([2.25;2.75] based on the 20%-percentile and the 80%-percentile of the bootstrap sample estimates).

Transferred to the issue of replicability or robustness of factor retention criteria, 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. However, replicability has no value in itself, of course. Robust and replicable factor solutions that misrepresent the underlying relations on population level are by no means desirable, but when comparing different factor retention criteria that showed comparably good performances in simulation studies focusing on the approximation of the data generating process (e.g., Auerswald and Moshagen 2019) which means trying to find the "true" dimensionality of a latent variable model that is assumed to be the data-generating cause, it might be reasonable to trust the more robust criterion. In this study, we evaluate the robustness (and the replicability) of factor retention criteria on real empirical data trying to complement the findings of Monte Carlo simulation studies, as both the replication goal and the approximation goal should be considered likewise in the factor retention process.

[1] The probability that an observation is included in the bootstrap sample equals $1 - (\frac{n_{obs}-1}{n_{obs}})^{n_{obs}}$ which is roughly $\frac{2}{3}$. In other words, every bootstrap sample contains roughly $\frac{2}{3}$ of the original observations on average. In total, $\frac{(2n_{obs}-1)!}{n_{obs}!(n_{obs}-1)!}$ different bootstrap samples are possible.

2 Methods

To illustrate how to use bootstrapping for the evaluation of the robustness of factor retention criteria and to investigate the relation between robustness and potential replicability, 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 et al. 2019; third data set: Stachl et al. 2017) and four samples of the *10 Item Big Five Inventory* (BFI-10; Rammstedt et al. 2013) 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. Since the *Phonestudy* data were collected for mobile sensing studies using smartphone logging data, the participants are comparably young (mean age: $M_1 = 24$, $M_2 = 23$, $M_3 = 24$) and well educated due to the recruitment procedures in academic contexts. In case of the BFI-10, we have one set of participants closely representing the German population, that were asked to fill out the questionnaire four times (waves *bd*, *cd*, *dd*, *ed* of the panel), so our four samples predominantly consist 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). Since all the *Phonestudy* data were collected in a comparable study setting, the three data sets can be interpreted as different cohorts of the same study and therefore used for a replication attempt. The questions for the four different waves of the panel study were mainly the same and instructions did not vary among the different measurements, so the BFI-10 data are repeated measures that can be used for a within-person replication study. These two different projects (*Phonestudy* and *GESIS* panel) where chosen for this study because they represent these two different replication contexts (within-person and between-person). While a within-person replicability speaks for a reliability of the questionnaire (related to the idea of the re-test reliability), a successful between-person replication can be seen as an indicator of factorial validity.

2.1 Aim of the analysis

In this paper, above all, we want to evaluate the relation between the robustness of a factor retention solution (defined as the stability of the suggested number of factors across bootstrap samples) and its replicability in empirical data sets. Furthermore, by applying different factor retention criteria to the empirical data sets, we are able to compare them with regard to their robustness and replicability. If robustness is

a good indicator for replicability, it might be a good idea to rely on robust factor retention criteria rather than on those that show little stability (in case that the robust criterion shows high accuracies in simulation studies as well).

3 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 (default settings with $\alpha = 0.30$ for the internal Mann–Whitney-*U* tests and 500 simulated data sets for the "comparison" approach), EKC and XGB as well as with a model comparison approach using the BIC. Afterwards, 100 bootstrap samples (ordinary non-parametric bootstrapping as described above) were drawn (using the *boot* package, Canty and 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 (in other words we evaluated the link between robustness indicated by the stability across bootstrap samples and replicability). 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. To quantify the relationship between the stability across bootstrap samples and actual replicability, we introduced two robustness metrics (volatility of solutions across bootstrap samples and a *rate of consistency* that is the percentage of bootstrap samples that yielded the same results as the empirical data set) and used them as independent variables in a generalized linear model (GLM; Nelder and Wedderburn 1972) to predict the probability of exact replication (logistic regression) as well as in a second GLM to predict the absolute error of replication (Poisson regression).

We used R (Version 4.0.0; R Core Team 2018) and the R-packages *data.table* (Version 1.12.8; Dowle and Srinivasan 2018), and *papaja* (Version 0.1.0.9997; Aust and Barth 2018) for all our analyses and the preparation of the manuscript.

4 Results

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 first BFI-10 data set and two factors for the third 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. The BIC-based model comparison approach, on the contrary, suggested five factors across all four samples and all bootstrap samples. Table 1 shows the solutions of the four retention criteria and the BIC for the four initial BFI-10 data sets as well as summary statistics for the respective bootstrap samples.

PA showed the lowest replicability—for the first wave (BFI_1) PA suggested three factors, for the second wave one factor, for the third wave two factors and for the fourth wave one factor. Across the 100 bootstrap samples of the first wave, PA yielded 2.98 factors on average ($SD = 0.887$), yet in just 40 of these bootstrap samples three factors as in the empirical data set were suggested. The percentages of bootstrap samples, for which PA implied the same dimensionality as for the empirical data set, were 48, 44, 29 for the second, third and fourth wave respectively. This so-called *rate of consistency* was higher for all other factor retention criteria that also showed perfect replication rates.

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 BFI-10 data. Mostly six facets per factor were suggested, but the results varied according to the retention criterion, the data set and the respective factor. BIC, EKC and XGB tended to show fewer differences between the bootstrapped solutions, whereas CD yielded the highest variance (or standard deviation) between the bootstrap samples for all combinations of data sets and factors.

Table 2 shows the solutions of the four retention criteria and the BIC for the five factors *openness*, *conscientiousness*, *extraversion*, *neuroticism* and *agreeableness* of three empirical BFSI data sets separately as well as summary statistics for the respective bootstrap samples. XGB showed the highest replicability, as it suggested six facets for the *openness* and the *conscientiousness* factor for all three data sets, while yielding six facets for two data sets and seven facets for one data set for the *extraversion*, *neuroticism* and *agreeableness* factors. All other factor retention criteria did not provide the same estimate for the number of facets for all three data sets once—PA, for example, suggested three different number of facets for all factors except *neuroticism*.

4.3 Robustness and replicability

We used a GLM with binomial family and logit link to model whether the number of factors was exactly replicated in the next data set 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 BFI-10 data set, the results of the third with those of the fourth BFI-10 data set, and the same for the three BFSI data sets. We modeled the

Table 1 Suggested number of factors of the four retention criteria and the BIC approach, means and standard deviations of the suggested number of factors across Bootstrap samples as well as percentages of Bootstrap samples with the same factor solution as the respective Empirical BFI-10 data set

Criterion	BFI_1	BFI_2	BFI_3	BFI_4	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
BIC	5	5	5	5	5.00	0.000	100	5.00	0.000	100	5.00	0.000	100	5.00	0.000	100

BFI_1 means the BFI-10 data set of the first wave (bd) in the panel, BFI_2 the second BFI-10 data set (wave cd) and so on. M_{BFI1} is the average number of factors suggested for the bootstrap samples of the first BFI-10 data set, SD_{BFI1} is the respective standard deviation across the bootstrap samples and $\%_{BFI1}$ indicates the *rate of consistency* which is the percentage of bootstrap samples for which the same number of factors was suggested as for the empirical data set

Table 2 Suggested number of factors of the four retention criteria and the BIC approach, means and standard deviations of the suggested number of factors across Bootstrap samples as well as percentages of Bootstrap samples with the same factor solution as the respective Empirical BFSI data set

	$BFSI_1$	$BFSI_2$	$BFSI_3$	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
BIC	6	7	4	7.35	0.702	13	7.32	0.665	46	5.24	0.754	13
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
BIC	4	4	2	5.31	0.761	12	5.35	0.757	13	3.70	0.823	6
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
BIC	5	5	4	6.16	0.677	13	6.40	0.899	17	5.05	0.626	15
Agreeableness												
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	4	4	6.40	0.711	6	6.00	0.636	63	5.11	0.827	26
BIC	4	5	2	5.98	0.778	2	5.99	0.718	24	4.11	0.931	4

Table 2 (continued)

	$BFSI_1$	$BFSI_2$	$BFSI_3$	M_{BFSI1}	SD_{BFSI1}	$\%_{BFSI1}$	M_{BFSI2}	SD_{BFSI2}	$\%_{BFSI2}$	M_{BFSI3}	SD_{BFSI3}	$\%_{BFSI3}$
Neuroticism												
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	6	4	6.15	0.642	13	5.33	0.697	12	5.08	0.720	19
BIC	4	5	3	5.51	0.810	9	5.77	0.617	33	3.95	0.809	28

$BFSI_1$ is the data set of Schoedel et al. (2018), $BFSI_2$ is the data set of Schuhwerk et al. (2019), $BFSI_3$ is the data set of Stachl et al. (2017). M_{BFSI1} is the average number of factors suggested for the bootstrap samples of the first BFSI data set, SD_{BFSI1} is the respective standard deviation across the bootstrap samples and $\%_{BFSI1}$ indicates the *rate of consistency* which is the percentage of bootstrap samples for which the same number of factors was suggested as for the empirical data set

relation between robustness and replicability for all factor retention criteria combined, so that five (number of criteria + BIC approach) times 13 (number of replications considered: three replications of the BFI-10 data and two replications of each of the five factors of the BFSI data)—65—instances were used for the analysis. 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 measures of robustness of the proposed factor solution. The absolute difference in the suggested number of factors between two consecutive data sets (e.g., the first BFSI data set compared with the second BFSI data set) served as a second measure of "replicability" of the proposed factor solutions (absolute replication error). A second GLM with Poisson family and log link was used for this dependent variable analogous to the first model with the standard deviation of the bootstrapped factor retention solutions and the *rate of consistency* as independent variables.

The results of the GLM analyses support the descriptive observations that factor retention criteria, that were more stable across bootstrap samples, were more likely to yield replicable results. With respect to exact replication (the first GLM), higher standard deviations for the suggested number of factors across the bootstrap samples were associated with a lower probability of replication [$b = -0.69$, 95% CI (-3.14, 1.21), $z = -0.65$, $p = 0.517$], whereas the percentage of bootstrap samples with the same solution as the initial data set (*rate of consistency*) was positively linked to this probability [$b = 0.05$, 95% CI (0.02, 0.09), $z = 3.09$, $p = 0.002$]. Results of the second GLM indicated that 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 [$b = 0.66$, 95% CI (0.13, 1.14), $z = 2.57$, $p = 0.010$]. 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, -0.01), $z = -2.81$, $p = 0.005$].

4.4 Comparing the criteria

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 for the BFI-10 data, BIC, 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 of consecutive number of factors or mean replication error: 0.38). PA had the lowest mean and median *rate of consistency* as well as the worst replicability rate of 7.69 %. CD yielded the most volatile results (highest mean and median standard deviation across the bootstrap samples),

Table 3 Means and medians of standard deviations of the suggested number of factors and rates of consistency over all data sets for the four factor retention criteria and the BIC approach 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
BIC	0.568	0.702	37.08	17	38.46	1.077

M_{SD} and Md_{SD} are the mean and median of the standard deviations of the suggested number of factors across the bootstrap samples over all data sets, $M_{\%}$ and $Md_{\%}$ are the mean and median of the *rate of consistency* over all data sets. $\%_{Replicable}$ is the percentage of exact replications, whereas, $M_{abs.Difference}$ is the mean absolute error of the replication of the factor retention process

which can be linked to the highest mean replication error (especially caused by the facet *conscientiousness* of the BFSI data sets, see Table 1). Table 3 provides an overview of these robustness and replicability measures for the four retention criteria as well as the BIC approach.

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 chosen as an easy-to-use method to evaluate the robustness (or stability) of the factor retention process that seems to be a good proxy for replicability. The study results showed 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 the 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 BFSI data. Several authors discussed this relationship between robustness and sample sizes for EFA in general (e.g., Osborne and Fitzpatrick 2012) and various simulation studies showed the need for larger samples to achieve higher accuracy/precision in EFA (see Goretzko et al. 2019 for an overview or MacCallum et al. 1999 for a simulation study). Regarding factor retention criteria, Auerswald and Moshagen (2019) (among others) found that they consistently perform better at higher sample sizes and although their focus lay 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. It was striking that (except PA for two waves) all factor retention criteria (not the BIC approach, though) suggested one factor for all four BFI-10 data sets, even though the BFI-10 claims to measure the "BIG5" with two indicators per factor. This small overdetermination (two manifest variables per latent factor) prevents a comprehensive confirmatory analysis of the factorial structure (Reilly 1995) and is seen as too small for EFA as well (e.g., Fabrigar et.

al. 1999), so the factorial validity of the BFI-10 remains unclear. However, for our study the questionnaire can still be used for the evaluation of the robustness and replicability of the factor retention process (the results should not be interpreted against the background of theoretical considerations, though).

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 an extensive simulation study of Goretzko and Bühner (2020). 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 rather than in the solutions PA and CD produced. However, 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. In practice, of course, replications of EFA results are usually conducted using confirmatory factor analysis (CFA), so the factor retention process is not replicated in general. However, when the number of factors suggested by a factor retention criterion is not replicable which means that the initial factor solution is not replicable, it cannot be expected that the CFA model with the same number of factors will show an acceptable fit to the data. Thus, the method applied to determine the number of factors should be replicable as the suggested number of factors is necessary to be "correct" for both the initial analysis and the subsequent replication.

The study should be considered purely descriptive, as the number of observations for the GLM analyses is rather small ($N = 65$). As mentioned above, this small number leads to an insufficient statistical power and does not allow cross-validation. With an α-level 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. 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 Tables 1 and 2) and the signs of GLM parameter estimates that met our expectations are indicators that robustness and 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)—particularly with regard of the 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 cohorts were sampled, making it a between-person replication.

Therefore, bootstrapping can be used to assess the robustness of the factor retention process against small data changes and seems to be a good proxy for replicability in the narrowest sense which means the replicability in samples from the same population (see also the section "Replicability and Robustness"). This robustness of the factor retention (or rather its replicability in a population) is a necessary but not sufficient condition for replications across populations (generalizability of the factor structure) which is of interest in context such as questionnaire development for cross-culture comparisons (i.e., measurement invariance across populations). Since

this study focuses solely on the factor retention process, it is also worth mentioning that the replicability of a factor structure goes beyond replicating the number of factors (even though replicability of the dimensionality is the basis for a successfully replicated factor structure)—inter-factor correlations, loading patterns and factor scores have to be regarded as well. Zientek and Thompson (2007) suggested bootstrapped EFA or PCA for these evaluations which of course can be combined with bootstrapped factor retention.

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*. We want to encourage researchers to include bootstrapping in their analyses, since 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 may be a better option for small samples, i.e., cases in which subsamples become too small for factor analytic methods, even though bootstrapping also benefits from greater samples and yields more trustworthy results with increasing sample sizes. When evaluating the robustness of the criteria, a comparison among them is imperative, because the stability measures cannot be interpreted absolutely (if 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) also recommend comparing methods and evaluate 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 on the goal of replication, where bootstrapping and the evaluation of the robustness of factor solutions might be a good start. However, it has to be stated again that replicability should not be an end in itself since replicating under- or overfactoring is not desirable at all. Accordingly, comparing the robustness of different factor retention criteria should always be accompanied by a reference to simulation studies (such as Auerswald and Moshagen 2019 or Goretzko and Bühner, 2020) that evaluate the accuracy of the respective factor retention methods. Thus, we would recommend to assess the robustness of several factor retention methods that have shown high accuracy in simulation studies with data conditions similar to the respective empirical data using bootstrapping. The results of the factor retention criteria showing the highest robustness (e.g., the highest *rate of consistency*) can be seen as more trustworthy and should be focused on when combining the results of the different methods. When combining the results (or setting the number of factors according to a specific criterion), the suggested number of factors based on

the empirical data set should be used as all methods yielded higher numbers on the bootstrapped samples on average (which could be seen as a sign of overfactoring).

Funding Open Access funding enabled and organized by Projekt DEAL.

Data availability No data were collected for this study. The data that support the findings of this study were provided by Stachl et al. (2018) and collected within the *Phonestudy* project (OSF projects at https://doi.org/10.17605/OSF.IO/UT42Y). Further data of the *GESIS* panel (GESIS 2018—https://doi.org/10.4232/1.13158) were used in this study. The panel data are not publicly available, but can be requested by researchers.

Declarations

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

Ethical Statement Data collected in two external research programs (*GESIS* panel and *Phonestudy* project, see above) were reanalyzed. No datasets were generated for this study (i.e., no surveys were conducted). Hence, no ethical approval was required.

Informed consent No new data were collected for this study. For both the *Phonestudy* project and the *GESIS* panel informed consent was obtained from the participants during data collection.

References

Aarts A, Anderson J, Anderson C, Attridge P, Attwood A, Axt J et al (2015) Estimating the reproducibility of psychological science. Science 349(6251):943–950

Arendasy M (2009) BFSI: Big-Five Struktur-Inventar (test & manual). Mödling, Schuhfried GmbH

Asendorpf JB, Conner M, De Fruyt F, De Houwer J, Denissen JJ, Fiedler K et al (2013) Recommendations for increasing replicability in psychology. Eur J Pers 27(2):108–119

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. Psychol Methods 24(4):468–491

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

Braeken J, Van Assen MA (2017) An empirical Kaiser criterion. Psychol Methods 22(3):450–466

Camerer CF, Dreber A, Holzmeister F, Ho T-H, Huber J, Johannesson M et al (2018) Evaluating the replicability of social science experiments in nature and science between 2010 and 2015. Nat Hum Behav 2(9):637–644

Canty A, Ripley BD (2019) Boot: Bootstrap R (S-Plus) functions. https://cran.r-project.org/web/packages/boot/boot.pdf

Cattell RB (1966) The scree test for the number of factors. Multivar Behav Res 1(2):245–276

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. ACM, pp 785–794
Chen T, He T, Benesty M, Khotilovich V, Tang Y (2018) Xgboost: Extreme gradient boosting. R package version 0.6. 4.1
Costa PT Jr, McCrae RR (1992) Four ways five factors are basic. Personality Individ Differ 13(6):653–665
Dowle M, Srinivasan A (2018) Data.table: Extension of 'data.frame'. https://CRAN.R-project.org/package=data.table
Efron B, Tibshirani RJ (1994) An introduction to the bootstrap. CRC Press, Boca Raton FL
Fabrigar LR, Wegener DT, MacCallum RC, Strahan EJ (1999) Evaluating the use of exploratory factor analysis in psychological research. Psychol Methods 4(3):272–299
GESIS (2018) GESIS Panel - Standard Edition (Version 25.0.0, Data file ZA5665). Cologne, GESIS Data Archive
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 25(6):776–786
Goretzko D, Pham TTH, Bühner M (2019) Exploratory factor analysis: Current use, methodological developments and recommendations for good practice. Curr Psychol. https://doi.org/10.1007/s12144-019-00300-2
Hancock GR, Liu M (2012) Bootstrapping standard errors and data-model fit statistics in structural equation modeling. In: Hoyle RH (ed) Handbook of structural equation modeling. The Guilford Press, pp 296–306
Horn JL (1965) A rationale and test for the number of factors in factor analysis. Psychometrika 30(2):179–185
Huber PJ (1981) Robust Statistics. Hoboken, NJ, John Wiley & Sons, Inc.
Jackson DA (1993) Stopping rules in principal components analysis: a comparison of heuristical and statistical approaches. Ecology 74(8):2204–2214
Kaiser HF (1960) The application of electronic computers to factor analysis. Educ Psychol Measur 20(1):141–151
Lim S, Jahng S (2019) Determining the number of factors using parallel analysis and its recent variants. Psychol Methods 24(4):452–467
MacCallum RC, Widaman KF, Zhang S, Hong S (1999) Sample size in factor analysis. Psychol Methods 4(1):84–89
Nelder JA, Wedderburn RW (1972) Generalized linear models. J R Stat Soc 135(3):370–384
Nevitt J, Hancock GR (2001) Performance of bootstrapping approaches to model test statistics and parameter standard error estimation in structural equation modeling. Struct Equ Model 8(3):353–377
Osborne JW, Fitzpatrick DC (2012) Replication analysis in exploratory factor analysis: What it is and why it makes your analysis better. Pract Assess Res Eval 17(15):1–8
Preacher KJ, Zhang G, Kim C, Mels G (2013) Choosing the optimal number of factors in exploratory factor analysis: a model selection perspective. Multivar Behav Res 48(1):28–56
Rammstedt B, Kemper C, Klein MC, Beierlein C, Kovaleva A (2013) Eine kurze Skala zur Messung der fünf Dimensionen der Persönlichkeit: Big-five-inventory-10 (bfi-10). Methoden, Daten, Analysen 7(2):233–249
Reilly T (1995) A necessary and sufficient condition for identification of confirmatory factor analysis models of factor complexity one. Sociol Methods Res 23(4):421–441
R Core Team (2018) R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/
Revelle W (2018) Psych: Procedures for psychological, psychometric, and personality research. Illinois, Northwestern University, Evanston
Ruscio J, Roche B (2012) Determining the number of factors to retain in an exploratory factor analysis using comparison data of known factorial structure. Psychol Assess 24(2):282–292
Schoedel R, Au JQ, Völkel ST, Lehmann F, Becker D, Bühner M, Stachl C (2018) Digital footprints of sensation seeking. Zeitschrift Für Psychologie 226(4):232–245
Schuwerk T, Kaltefleiter LJ, Au JQ, Hoesl A, Stachl C (2019) Enter the wild: autistic traits and their relationship to mentalizing and social interaction in everyday life. J Autism Dev Disord 49:1–16
Schwarz GE (1978) Estimating the dimension of a model. Ann Stat 6(2):461–464
Shrout PE, Rodgers JL (2018) Psychology, science, and knowledge construction: broadening perspectives from the replication crisis. Annu Rev Psychol 69(1):487–510
Stachl C, Hilbert S, Au JQ, Buschek D, De Luca A, Bischl B, Bühner M (2017) Personality traits predict smartphone usage. Eur J Pers 31(6):701–722

Stachl C, Schoedel R, Au JQ, Völkel ST, Buschek D, Hussmann H, Bühner M (2018) The phonestudy project. Open Sci Framework. https://doi.org/10.17605/OSF.IO/UT42Y

Thalmayer AG, Saucier G, Eigenhuis A (2011) Comparative validity of brief to medium-length big five and big six personality questionnaires. Psychol Assess 23(4):995–1009

Zientek LR, Thompson B (2007) Applying the bootstrap to the multivariate case: bootstrap component/factor analysis. Behav Res Methods 39(2):318–325

Zwick WR, Velicer WF (1986) Comparison of five rules for determining the number of components to retain. Psychol Bull 99(3):432–442