Psychological Test and Assessment Modeling, Volume 63, 2021 (1), 45–64

Determining the Number of Factors in Exploratory Factor Analysis – The Performance of Tests of Relative Model Fit

Kay Brauer[1] *& Jochen Ranger*[2]

Abstract

The number of factors in exploratory factor analysis is often determined with tests of model fit. Such tests can be employed in two different ways: Tests of global fit are used to compare factor models with increasing number of factors against a saturated model whereas tests of relative fit compare factor models against models with one additional factor. In both approaches, the number of factors is determined by choosing the simplest model that is not rejected by the test of model fit. Hayashi, Bentler, and Yuan (2007) recommend using tests of global fit because the tests of relative model fit tend to overfactoring. We investigate the performance of the tests of relative model fit. Overfactoring is prevented by using either a bootstrap implementation or a modification of the standard tests. The modification consists in testing each model against a restricted alternative that is identified under the null hypothesis. Simulation studies suggest that our tests of relative model fit perform well. Both implementations adhere to the nominal Type-I error rate closely and are more powerful than the tests of global fit. The application of the tests is illustrated in an empirical example.

Keywords: Exploratory Factor Analysis, Factor model, Fit tests; Model fit

[1] *Correspondence concerning this article should be addressed to:* Kay Brauer, Psychological Assessment and Differential Psychology, Department of Psychology, Martin-Luther-University Halle-Wittenberg, Halle, Germany; E-Mail: kay.brauer@psych.uni-halle.de

[2] Psychological Methods, Department of Psychology, Martin-Luther-University Halle-Wittenberg, Halle, Germany

Introduction

Factor analysis was developed more than a century ago (Cudeck & MacCallum, 2007) and has become one of the most popular statistical methods in psychology since then (Costello & Osborne, 2005; Fabrigar, Wegener, MacCallum, & Strahan, 1999; Golino & Epskamp, 2017). Foremost, factor analysis is used for identifying the underlying dimensions of instruments that aim at measuring latent constructs such as personality and intelligence (e.g., Carroll, 1993; Furr & Bacharach, 2014; Hopwood & Donnellan, 2010). In particular, factor analysis is used for identifying how many dimensions (i.e., number of factors) are needed to describe individual differences (e.g., structural models of broad and narrow personality traits) and for assessing whether theoretical model assumptions converge with empirical data. We will refer to this as the "number of factors problem". Frequently, the number of factors is determined by informal techniques that lack a sound statistical basis (Hayashi et al., 2007). However, there are scenarios in which two alternative factor solutions seem feasible such that a rigorous statistical comparison is required. In this study, we examine such approaches on basis of simulation studies and empirical data to extend the knowledge on determining the number of factors in factor analysis.

Factor Analytic Approaches and the Estimation of the Number of Factors

Two types of factor analysis are distinguished, exploratory factor analysis (EFA) and confirmatory factor analysis (CFA; Mair, 2018), which differ in their aims. EFA aims at finding a factor solution that describes the data well whereas CFA aims at evaluating whether an a-priori *assumed* factor structure and its implied covariance matrix are compatible with the *observed* covariance matrix. EFA and CFA differ mathematically in the way the factor loadings are restricted (Amemiya & Anderson, 1990) but both forms are interwoven. EFA is typically used for identifying potential factor solutions that describe the data well and CFA contributes to further validating the assumed structure (e.g., Furr & Bacharach, 2014). Here, the focus is entirely on EFA.

In EFA, the first and probably most fundamental research question refers to the number of factors to extract. This question has been addressed with different approaches in the past. Some of these approaches are based on the eigenvalues of the correlation matrix. Popular examples are the Kaiser criterion (Guttman, 1954; Kaiser, 1960), the scree test (Cattell, 1966; Lorenzo-Seva, Timmerman, & Kiers, 2011; Raiche, Walls, Magis, Riopel, & Blais, 2013; Zhu & Ghodsi, 2006), and parallel analysis (Green, Xu, & Thompson, 2018; Horn, 1965; Ruscio & Roche, 2012). However, those have been criticized (e.g., Furr & Bacharach, 2014; Goretzko, Pham, & Bühner, 2019); for example, the Kaiser criterion tends to overestimate the number of factors and the visual inspection of scree plots tends to rely on researchers' degrees of freedom in interpreting the cut-off. Other approaches are based on the indicators' intercorrelations that remain when the influences of the common factors have been partialled out (e.g., the Minimum Average Partial Correlation Test; Velicer, 1976) or exploratory graph analysis (Golino & Epskamp, 2017). Measures of model fit that describe how well a model represents the covariance matrix of the indicators, are also commonly used. In using them one chooses the factor model with the lowest number of factors that has still acceptable fit (e.g., Montoya & Edwards, 2020), compares the relative fit of different factor models (Akaike, 1992; Finch, 2019; Huang, 2017; Schwarz, 1978), or evaluates whether the loading matrix has a simple structure (Revelle & Rocklin, 1979; for an overview see e.g., Preacher, Zhang, Kim, & Mels, 2013). A discussion of the merits and drawbacks of the different approaches can be found in Furr and Bacharach (2014) and Goretzko et al. (2019).

In our study, we focused on two popular procedures that are based on statistical tests of model fit. The first procedure is based on tests of *global model fit*, which assess the compatibility of the implied covariance matrix with the empirical covariance matrix (e.g., Amemiya & Anderson, 1990; Bartlett, 1950; Foldnes, Foss, & Olsson, 2011; Lawley, 1943; Rao, 1955; Shapiro, 1986). These tests can be interpreted as a comparison between a specific factor model and a saturated factor model that is capable of reproducing the sample covariance matrix perfectly. Tests of global model fit can be used in order to identify the number of factors by testing several factor models with increasing number of factors. Among all factor models with an insignificant test result, the factor model with the lowest number of factors is chosen. Secondly, tests of *relative model fit* (Steiger, Shapiro, & Browne, 1985) compare the global fit of two nested models and assess whether the global fit of the two models differs systematically. Tests of relative model fit can be used for determining the number of factors by comparing models of increasing dimensionality: Beginning with a comparison of a 1- vs 2-factor model, one successively increases the number of factors until the test of relative model fit is insignificant. Since determining the number of factors can be interpreted as a model selection problem (Preacher et al., 2013), assessing relative fit seems the preferred approach. The fact that tests against specific alternatives are more powerful than tests against global alternatives additionally suggests using relative tests (Saris, Satorra, & van der Veld, 2009).

Model selection on basis of tests of relative model fit and all related model selection criteria is problematic because the regularity conditions assumed in maximum likelihood (ML) theory are violated when a factor model is tested against a model with a higher number of factors (see the next section for a more detailed description). In this case, the tests of relative model fit are liberal and, thus, simpler models are rejected too frequently, and the number of factors is overestimated. For this reason, Hayashi and colleagues (2007) recommend using the tests of global fit for the determination of the number of factors. We question this recommendation. In our study, we show that the tests of relative model fit *can* be used to determine the number of factors provided that they are implemented adequately. We suggest two solutions to the problem of overfactoring. The first solution is a modification of the test of relative model fit. Instead of comparing a factor model with an underidentified factor model with additional factors, we use an alternative model that is identified under the null hypothesis and provides a good approximation to a factor model with one additional factor. The second solution consists in a parametric bootstrap. In the following, we describe the two solutions in more detail and investigate their performance with respect to their size and power in a simulation study. Finally, we illustrate their application with real data.

Two Implementations of the Test of Relative Model Fit

In the test of relative model fit that are considered here, one compares a factor model with an extended version that contains one additional factor. No assumptions are made about the loadings of the additional factor. For the test, both models are fitted to the data, either via ML or Weighted Least Squares (WLS) estimation. In doing so, one determines those values of the model parameters that make the implied covariance matrix as similar to the observed covariance matrix as possible. The similarity of the two matrices is assessed with a discrepancy function that differs in ML and WLS estimation (Shapiro, 1986). The test of relative model fit is based on the values of the discrepancy function at the parameter estimates of the two models. If the factor model with the smaller number of factors holds exactly in the population, the two discrepancy values do not differ systematically. If the model with the additional factor holds, its discrepancy value is systematically lower. The scaled difference of the two discrepancy values provides a test statistic that is compared to a χ^2 distribution to evaluate its statistical significance. Although the test statistic is approximately distributed according to a central χ^2 distribution under the null hypothesis in CFA, this is not the case in EFA since the model of higher dimensionality is underidentified when the model of lower dimensionality holds. As the loadings of the additional factor are all zero in that case, the matrix of factor loadings does not have full rank. This violates the regularity conditions assumed in ML theory (Amemiya & Anderson, 1990; Geweke & Singleton, 1980). When the χ^2 distribution is used notwithstanding, the test becomes liberal (Hayashi et al., 2007). We suggest two solutions to this problem, based on (1) a model restriction and (2) a parametric bootstrap.

The Restricted Model Approach

Comparing a factor model with m factors (H_0-model) against a factor model with $m + 1$ factors (H_1-model) using a test of relative model fit is problematic. When the H_0-model holds, the loading matrix of the H_1-model does not have full column rank. This can be avoided by testing the H_0-model against an alternative H_1-model. The alternative H_1-model has to be identified under the null hypothesis (loading matrix of full column rank) and should provide a good approximation to the original H_1-model. Such an alternative H_1-model can be generated with the original H_1-model by restricting one specificity (i.e., specific variance) to zero. An example might help to clarify this: When a 1-factor model is tested against a 2-factor model, one could, for example, test the 1-factor model against the factor model given in Figure 3 (observable indicators are denoted as $x_1, x_2, \ldots, x_P$, factors as θ_1 and θ_2, and the residuals as $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_P$). For model identification, the loading of x_1 on θ_2 is set to zero (Algina, 1980). The restriction of the path $\varepsilon_2 - x_2$ restricts the specificity of x_2 to zero, which prevents that the loading matrix is rank-deficient when the H_0-model holds. In this case, the second factor takes over the role of the residual (specific factor) ε_2. The loadings of θ_2 in all further indicators will be estimated close to zero. When the H_0-model is wrong as the H_1-model with the additional factor holds, the model in Figure 3 is misspecified. The parameter estimates of the restricted model converge to those values that provide the best approximation to the covariance matrix implied by the true 2-factor model. As the misspecification is only local, the correctly specified part of the model becomes dominant in longer tests; note that the number of covariances increases quadratically with the number of indicators. As a consequence, the loadings of the second factor will deviate from zero.

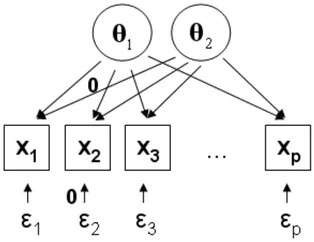

Figure 3
Restricted Exploratory Two-Factor Model with Specificity of Zero in the Second Indicator. *Note.* The loading of the first indicator on the second factor is restricted to zero to prevent the rotational indeterminacy of the model.

We suggest the following procedure to determine dimensionality. Starting from a comparison of a 1-factor model against a 2-factor model, one compares adjacent models with increasing number of factors consecutively. In each step, a model with m factors is tested against the adjacent model with $m + 1$ factor as follows: First, the model with $m + 1$ factors is fitted to the data with the restriction described above. For reasons given below, we suggest using the diagonally weighted least squares (DWLS) estimator. Having fitted the model, a Wald test is used to examine whether all loadings of the $(m+1)$-th factor are zero. In doing so, the loading of the indicator with zero specificity is excluded. Models of increasing number of factors are compared until the first test is insignificant. The number of factors of the more parsimonious model is then considered as the true number.

We prefer DWLS over ML estimation because in DWLS one can control the impact the elements of the covariance matrix have on the parameter estimates by modifying the corresponding elements of the weight matrix. This allows to increase the power of the test. We suggest reducing the weight of the variance of the indicator with the restricted specificity, which reduces the impact of the misspecification of the restricted H_1-model under the H_1-hypothesis. A reasonable choice is to set the weight to 20 % of its original value. The power of the test depends on the choice of the indicator whose variance is set to zero. This is similar to the effects of the chosen identification restriction that is also known to affect the power of the tests of model fit (Millsap, 2001). We recommend choosing an indicator with low specificity.

The Bootstrap Approach

Bootstrapping is a viable approach to statistical inference in case this is difficult otherwise; for a general treatment of the bootstrap see Efron and Tibshirani (1993) and Davison and Hinkley (1997). Since its introduction into Structural Equation Models (Beran & Srivastava, 1985; Bollen & Stine, 1993), the bootstrap has been used for parameter inference (e.g., Boomsma, 1986; Lambert, Wildt, & Durand, 1991; Yuan & Hayashi, 2006), the evaluation of model fit (e.g., Cheng & Wu, 2017; Nevitt & Hancock, 2001; Yuan & Hayashi, 2003; Yung & Bentler, 1996; Zhang & Savalei, 2016), and power analyses (Yuan & Hayashi, 2003). The bootstrap is a resampling approach, that uses the distribution of estimators or test statistics over bootstrap samples as a proxy of the theoretical distribution. Depending on how the bootstrap samples are generated, two versions can be distinguished, the nonparametric bootstrap and the parametric bootstrap. In the nonparametric bootstrap, the bootstrap samples are generated by random draws with replacement from the original sample. The nonparametric bootstrap is used when the multivariate normality of the data is questionable. In the parametric approach, the bootstrap samples are generated with a standard distribution, typically the multivariate normal distribution. The parametric bootstrap is used in small samples when the data are not a good proxy of the true distribution function (for a critical review see Yung & Bentler, 1996). Here, we will employ the parametric bootstrap in order to compare the fit of an m-factor model against an $m + 1$-factor model which has – at least to our knowledge – not been done before.

In a first step, an m-factor model is fitted to the data of the original sample. Denote the sample size as n and the implied covariance matrix of the fitted factor model as $\hat{\Sigma}(m)$. Having fitted the model, one generates a large number of bootstrap samples of size n. Each bootstrap sample is generated by n independent random draws from a multivariate normal distribution with a mean of zero and covariance matrix $\hat{\Sigma}(m)$. That way, B bootstrap samples are generated. To all bootstrap samples, factor models with m factors and $m + 1$ factors are fitted. The relative model fit is assessed via a likelihood ratio (LR) test that compares the models with m and $m + 1$ factors. The distribution of the LR test statistic over the bootstrap samples is finally used to determine the rejection region of the test. In the simplest case, one uses a quantile of the bootstrap distribution as the critical level. With the bootstrap test of relative model fit, the number of factors is determined as described above. Starting from a comparison of a 1-factor against a 2-factor model, one successively increases the number of factors until the first test is not significant. The number of factors of the more parsimonious model is then considered as the true number of factors.

Study 1 (Simulations)

We conducted two simulation studies to evaluate the proposed approaches. First, we evaluated the tests of model fit with respect to their nominal Type-I error rate (Study 1a). Then, we assessed their power (Study 1b).

Simulation Study 1a

In the first simulation study, the same model was used for generating and analyzing the data. Two simulation conditions were considered. In the *first* simulation condition, the data were generated according to a 1-factor model as follows. For each fictitious test taker, a factor score was drawn from the standard normal distribution. Individual responses were then generated according to a factor model with normally distributed residuals. In line with previous simulation studies (Hayashi et al., 2007; Hu & Bentler, 1999), the loadings were set to 0.707 in half of the indicators and to 0.816 in the remaining ones. The variances of the residuals were chosen such that the variances of the indicators were 1. The chosen parameter values implied communalities of .50 and .66. Simulation samples were generated for a test with 12, 24, or 36 indicators. The sample size was $n = 125$, $n = 250$, and $n = 1000$. For each of the 3×3 combinations of test length and sample size, 500 simulation samples were generated. In the *second* simulation condition, the data were generated according to a 2-factor model. Factor scores were randomly drawn from a standard bivariate normal distribution with covariance of zero. Residuals were assumed to be normally distributed. The loading matrix of the 2-factor model had a simple structure. Half of the indicators had nonzero loadings on the first factor and half of the indicators had nonzero loadings on the second factor. The loadings on each factor alternated between the values 0.707 and 0.816 for indicators with non-zero loadings. The chosen parameter values implied communalities of .50 and .66. As in the first simulation condition, we considered three different lengths of the test (12/24/36 indicators) and three sample sizes ($n = 125/250/1000$). For each of the 3×3 combinations of test length and sample size, 500 simulation samples were generated.

Three factor models were fitted to the simulated samples. First, we fitted an unrestricted factor model with the correct number of factors to the data via ML estimation. Here, as well as in the following, we denote models as unrestricted when no restrictions are made beyond the ones required to scale the latent variables and to prevent rotational indeterminacies. Second, we fitted an unrestricted factor model with one additional factor via ML estimation. Finally, we fitted a restricted factor model with one additional factor via DWLS estimation. The restriction consisted in setting one specificity to zero, as described in the previous section. The weight matrix of the DWLS estimator was modified by reducing the weight of the variance in the indicator with zero specificity by 80 %. When estimating the models, the residual variances were restricted to be non-negative (see Savalei & Kolenikov, 2008).

Having fitted the models, several tests of model fit were performed. The first test was a LR test of relative model fit that compared the baseline model against the model with one additional factor. This is the liberal test that was criticized by Hayashi et al. (2007). The LR test of relative model fit was also implemented in a parametric bootstrap version. The second test was the LR test of global model fit that compares the baseline model against a saturated model that is capable of reproducing the sample covariance matrix perfectly. We considered the test in its standard implementation and in Bartlett's (1950) corrected version. Finally, we examined the Wald test that tested whether the loadings of the additional factor all deviate from zero. For this test, we used the DWLS estimates from the restricted model. Loadings were excluded when they were set to zero for sake of identification or when they corresponded to the indicator with specificity of zero. An overview of the tests is given in Table 7.

Table 7
Overview of the Tests Considered in the Simulation Study

Label	Estimator	Test	Comparison	Approach
LR-C	ML	LR	m-Factor vs. $(m+1)$-Factor Model	Standard
LR-B	ML	LR	m-Factor vs. $(m+1)$-Factor Model	Bootstrap
LR-S	ML	LR	m-Factor vs. Saturated Model	Standard
LR-S'	ML	LR	m-Factor vs. Saturated Model	Bartlett
Wald-R	DWLS	Wald	Loadings of $(m+1)$-th Factor against Zero	Standard

Note. ML = Maximum-Likelihood. DWLS = Diagonally Weighted Least Squares. LR = Likelihood-Ratio.

All tests were performed on $\alpha = .01$, $\alpha = .05$, and $\alpha = .10$. The empirical rejection rates are reported in Table 8 for data sets with a one-factor structure and in Table 9 for data with a two-factor structure. Note that the empirical rejection rate should be close to the nominal Type-I error rate α.

With respect to the LR-C test and the LR-S' test, our findings (see Table 8 and 9) are similar to those of Hayashi and colleagues (2007). The LR-C test is liberal and exceeds the nominal Type-I error rate tremendously. It amounts up to 0.988 in samples of $n = 125$ subjects and $I = 36$ indicators. The poor performance, however, is universal, and not related to unfavorable conditions like small samples and long tests. The LR-S', on the other hand, closely adheres to the nominal Type-I error rate. The good performance of the test is partly due to the Bartlett correction. The standard LR-S test without correction is less well behaved and tends to be liberal in small samples and long tests. The LR-B performs well in all conditions. The Wald-R test performs less well. It is too liberal in samples of $n = 125$ and $n = 250$ subjects, irrespective of the number of indicators. However, the test performs well in samples of $n = 1000$ subjects. In most conditions—the case of 24 indicators and 125 subjects being an exception—the Wald-R test performs better than the standard LR-S test.

Table 8
Type-I Error Rates of Tests of Model Fit for Different Levels (α), Sample Sizes (n) and Number of Indicators (I) When Testing the Fit of a Factor Model with one Factor

		$n = 125$			$n = 250$			$n = 1000$		
Test	α	$I = 12$	$I = 24$	$I = 36$	$I = 12$	$I = 24$	$I = 36$	$I = 12$	$I = 24$	$I = 36$
LR-C	.10	.566	.878	.988	.550	.878	.986	.564	.920	.988
	.05	.364	.712	.892	.348	.724	.936	.382	.770	.934
	.01	.112	.290	.508	.106	.322	.598	.118	.298	.614
LR-S	.10	.154	.410	.808	.140	.252	.410	.132	.126	.154
	.05	.094	.276	.716	.072	.152	.300	.056	.070	.070
	.01	.032	.092	.474	.016	.044	.098	.018	.016	.028
LR-S’	.10	.100	.092	.108	.106	.116	.096	.124	.112	.096
	.05	.060	.062	.070	.054	.054	.044	.054	.058	.048
	.01	.010	.010	.018	.012	.006	.010	.014	.016	.008
LR-B	.10	.122	.104	.106	.108	.114	.102	.136	.098	.096
	.05	.058	.036	.072	.062	.052	.056	.076	.044	.028
	.01	.004	.008	.014	.018	.020	.010	.016	.016	.002
Wald-R	.10	.174	.532	.270	.122	.092	.196	.094	.114	.116
	.05	.106	.510	.198	.054	.054	.116	.048	.054	.056
	.01	.042	.452	.100	.010	.028	.046	.012	.008	.006

Note. Results are based on 500 simulation samples; a description of the tests is given in Table 7.

Simulation Study 1b

In the second simulation study, we investigated the power of the tests. Data were generated for a test comprising $I = 24$ indicators and $n = 250$ subjects, as these were

Table 9

Type-I Error Rates of Tests of Model Fit for Different Levels (α), Sample Sizes (n) and Number of Indicators (I) When Testing the Fit of a Factor Model with two Factors

		$n = 125$			$n = 250$			$n = 1000$		
Test	α	$I = 12$	$I = 24$	$I = 36$	$I = 12$	$I = 24$	$I = 36$	$I = 12$	$I = 24$	$I = 36$
LR-C	.10	.478	.886	.988	.476	.884	.986	.484	.904	.992
	.05	.296	.726	.904	.276	.704	.924	.320	.730	.958
	.01	.070	.312	.542	.076	.270	.598	.106	.292	.646
LR-S	.10	.142	.406	.810	.120	.224	.422	.112	.134	.144
	.05	.080	.268	.702	.058	.126	.294	.056	.062	.070
	.01	.022	.118	.496	.014	.040	.118	.008	.010	.012
LR-S’	.10	.086	.114	.110	.092	.102	.106	.106	.110	.092
	.05	.048	.050	.048	.042	.058	.054	.054	.050	.038
	.01	.014	.008	.016	.012	.012	.008	.008	.008	.006
LR-B	.10	.094	.118	.090	.102	.094	.114	.142	.096	.082
	.05	.048	.058	.048	.048	.034	.070	.064	.054	.042
	.01	.008	.014	.010	.016	.001	.018	.020	.010	.008
Wald-R	.10	.212	.643	.364	.130	.152	.250	.124	.116	.116
	.05	.136	.590	.282	.084	.110	.182	.056	.056	.068
	.01	.078	.536	.158	.018	.070	.084	.006	.008	.014

Note. Results are based on 500 simulation samples; a description of the tests is given in Table 7.

the conditions under which most tests of model fit adhered to the Type-I error rate in Study 1a. In order to investigate the power, the model used for data analysis was misspecified. We considered different forms of misspecification in two simulation conditions.

The *first* simulation condition was similar to Study 1a. We, however, replaced each factor by two factors with a correlation coefficient of $\rho = 0.95$. The 1-factor model of the first simulation condition in Study 1a was replaced by a model with two oblique factors and a simple structure. The loadings of the factors were identical to the loadings of Study 1a (0.707 and 0.816 in half of the indicators each). The 2-factor model of the second simulation condition in Study 1a was replaced by a factor model with four oblique factors and simple structure. The four factors could be grouped into correlated pairs. Factors in different groups were uncorrelated. The loadings were identical to Study 1a. By this proceeding, we increased the number of factors without changing the uniqueness of the indicators in comparison to the first simulation study. A correlation of $\rho = 0.95$ was chosen in order to simulate a difficult detection problem.

In the *second* simulation condition, we assumed an additional local factor. The data were generated similar to Study 1a with one exception. In addition to the one or two factors of the original model, we introduced an additional local factor. The local factor was uncorrelated with the main factors of the model and had loadings of 0.30 on the last three indicators. The data had thus a two dimensional or a three-dimensional structure depending on the condition.

Data were generated as described in the previous section. For each condition, 500 samples were simulated. The data that were generated with a 2-factor model, were analyzed with a 1-factor model; the data that were generated with a 3-factor or 4-factor model, were analyzed with a 2-factor model. The models were fitted to the data as in the previous section. Then, the tests of model fit were performed. In the second simulation study, we only considered the LR test of global model fit in the version of Bartlett (LR-S'), the test of relative model fit with bootstrapped p-values (LR-B), and the Wald test based on the DWLS estimates of the restricted model (Wald-R). We did not consider the remaining tests because they were liberal. All tests were performed on $\alpha = .01$, $\alpha = .05$, and $\alpha = .10$. For both simulation conditions, the rejection rates of the tests (power) are reported in Table 10.

The LR-S' test of global model fit that was recommended by Hayashi et al. (2007) has high power in the first simulation condition and moderate power in the second simulation condition, irrespective of the number of factors. The power, however, is comparatively lower than the power of the LR-B test. The bootstrap test of relative model fit (LR-B) performs much better in both simulation conditions. The bootstrap test is so powerful that the misspecified factor structure is detected in almost all samples with $\alpha = .10$. Findings are mixed in data generated with two factors. The Wald test is rather powerful in the first simulation condition but lacks power in the second. This might be due to the weak effects of the local factor that are not strong enough to overpower the variance restriction.

Table 10
Power of Tests of Model Fit to Detect Additional Factors for Different Factor Models, Levels of α and two Forms of Misspecified Factor Structure

Data: Two Factors / Model: One Factor						
	Correlated Factors			Local Factor		
Test	$\alpha = .01$	$\alpha = .05$	$\alpha = .10$	$\alpha = .01$	$\alpha = .05$	$\alpha = .10$
LR-S’	.524	.740	.828	.236	.512	.614
LR-B	.948	.982	.988	.714	.846	.908
Wald-R	.802	.832	.848	.098	.146	.194
Data: Three or Four Factors / Model: Two Factors						
	Correlated Factors			Local Factor		
Test	$\alpha = .01$	$\alpha = .05$	$\alpha = .10$	$\alpha = .01$	$\alpha = .05$	$\alpha = .10$
LR-S’	.198	.444	.572	.272	.494	.614
LR-B	.338	.562	.678	.642	.788	.862
Wald-R	.320	.386	.458	.160	.192	.260

Note. Results are based on 500 simulation samples; a description of the tests is given in Table 7.

Study 2 (Empirical Application)

We examined the performance of the tests by analyzing openly available real-life data (Price, 2012). The data set contained scores of 1,000 participants (53.3 % females; age: $M = 46.4$, $SD = 23.8$ years) in seven tests of intelligence. Previous analyses suggested that the data were compatible with a 2-factor model (i.e., supposedly representing the domains of crystallized and fluid intelligence). Here, we reanalyze the data in order to investigate whether the introduced tests would support this conclusion.

For this purpose, we fitted unrestricted 1-factor, 2-factor, and 3-factor models (ML estimation) as well as restricted 2-factor and 3-factor models (DWLS estimation) to the data. We then performed the tests described above. Each of the unrestricted factor models was tested for global fit with the standard test (LR-S) and with the modified Bartlett test (LR-S’). To assess the fit, the unrestricted 1-factor model was compared to the unrestricted 2-factor model and the unrestricted 2-factor model to the unrestricted 3-factor model. This was achieved by the (incorrect) standard test (LR-C) and the implementation with bootstrapped p-values (LR-B). For the restricted 2-factor and 3-factor model, we tested whether the free loadings on the last factor deviated significantly from zero (Wald-R). Results for the tests are displayed in Table 11.

All tests clearly reject the hypothesis that one factor is capable of modeling the covariance matrix. The Wald-R test has the highest test statistic and the lowest degrees of freedom. The 2-factor model is not in conflict with the data. None of the tests can be rejected on

Table 11
Test Statistic (χ^2), Degrees of Freedom and p-Values of the Tests When Testing the Global and Relative Fit of Factor Models with Different Number of Factors

	Factors								
	1			2			3		
Test	χ^2	*df*	p	χ^2	*df*	p	χ^2	*df*	p
LR-C	494.07	6	$<.001$	7.44	5	.189	–	–	–
LR-S	503.53	14	$<.001$	9.47	8	.305	2.02	3	.567
LR-S’	501.10	14	$<.001$	9.41	8	.309	2.01	3	.569
LR-B	494.07	–	$<.001$	7.44	–	.296	–	–	–
Wald-R	862.07	5	$<.001$	8.08	4	.088	–	–	–

Note. $n = 1000$; a description of the tests is given in Table 7.

$\alpha = .05$. The Wald-R test has again the lowest p-value. It has almost the same value than the standard tests of relative model fit, but fewer degrees of freedom. The bootstrap version of the test of relative model fit is more similar to the tests of global model fit. Taking all findings together, there is little support that a 3-factor model is superior to a 2-factor model with respect to model fit. Given the high power of the new tests, this further supports the notion that only two latent constructs are measured.

Discussion

Determining the number of factors is the first and most important question in EFA. To address this problem, numerous approaches have been suggested. Although several of these approaches work well in practice, only few have a sound statistical basis (Fabrigar et al., 1999; Hayashi et al., 2007; Schmitt, 2011). Among them are the tests of model fit that evaluate whether the observed covariance matrix is compatible with the assumed factor model.

There are two different ways to employ the tests of model fit in order to identify the underlying number of factors. The first way consists of testing factor models with increasing number of factors against the saturated model that is capable of reproducing the observed covariance matrix perfectly. The simplest model capable of representing the covariances is chosen as the correct model. The second way consists in comparing a factor model against an extended model with one additional factor. The number of factors is increased until the extended model’s fit does not significantly differ from the reduced model’s fit. In both procedures, the problem of multiple conditional testing is ignored.

Theoretical considerations and empirical findings suggest that only the first procedure is justifiable when the standard LR test is used. The second procedure is problematic as tests of relative model fit based on the standard LR test do not perform regularly because the matrix of the factor loadings is rank-deficient under the null hypothesis; note that this complication not only impairs the tests of relative model fit. Practically all approaches based on the χ^2 approximation (e.g., RMSEA) or the asymptotic theory of ML estimation will be affected (Huang, 2017). For this reason, Hayashi and colleagues (2007) recommend using tests of global fit. This recommendation is partly supported by our simulation studies. As the tests of global fit control for the Type-I error rate well, the procedure guards against overfactoring. A drawback of the tests of global fit is the rather unspecified alternative hypothesis. This is disadvantageous for the power of the tests. High power, however, is very important in the present context. Due to the nature of the hypothesis, a nonsignificant test of model fit does not necessarily imply that the model is capable of representing the data (Yuan, Chan, Marcoulides, & Bentler, 2016). A nonsignificant test result is only informative when the test has sufficient power against a relevant set of alternative models.

In this paper, we investigated the performance of tests of relative fit. Overfactoring was prevented by two approaches: First, by bootstrapping the distribution of the test statistic. Although we used a parametric bootstrap, one could also use the model-based nonparametric bootstrap (Bollen & Stine, 1993), at least in large samples (Ichikawa & Konishi, 1995). Secondly, by testing a factor model against a restricted version of the model with one additional factor. The restriction was chosen such that the model is identified under the H_0-hypothesis and provides an approximation to the unrestricted factor model with one additional factor. The performance of the two approaches was investigated in a simulation study. The study corroborates that tests of relative model fit perform better than tests of global model fit when implemented adequately. Using the bootstrap test for the relative model comparisons results in very high power. Under some conditions, the rejection rates of the bootstrap test are four times higher than the rejection rates of the test of global model fit. Findings are less promising for the test against the restricted H_1-model. Although this test is powerful in models with an additional global factor, it has low power in a model with an additional local factor. From a mathematical point of view, this is clearly a disadvantage. From a practitioner's perspective, one can debate whether the detection of a local factor is wanted. Tests should be powerful against relevant alternatives. The restricted test is constructed in such a way that only global factors can be detected. Hence, it might test for precisely the alternative one is interested in. The relation between the power for a specific hypothesis and the way the restriction is implemented, however, needs further investigation.

We want to note two potential limitations. First, our simulation study was limited with regard to the assumption of comparatively high loadings. We chose this scenario in line with previous simulation studies (Hayashi et al., 2007; Hu & Bentler, 1999). The simulation study should be supplemented by further simulation studies in which factor loadings are varied and also cover a lower range of loadings (Themessl-Huber, 2014). We also did not consider alternative tests of relative model fit like, for example, the F-test employed by Kubinger, Litzenberger, and Mrakotsky (2006; see also Themessl-

Huber, 2014). Second, although we have demonstrated that a series of local model comparisons has higher power than a series of global model comparisons, this is foremost a mathematical advantage. Whether the gain in power has relevance in practice, where all models are probably false, is uncertain. It has to be noted that the capability to consistently reject a H_0-hypothesis when it is false and the sample increases beyond bound is not a defect of hypothesis testing but a strength, provided the right question has been posed. Although perfect model fit is unrealistic and every model can be rejected with a sufficiently large sample size, this does not invalidate model testing in general. When a m-factor model has been rejected, it is always possible to test the m-factor model for approximate model fit. One could, for example, test whether m and $m+1$-factor models differ by a certain amount. The bootstrap method of Yuan and colleagues' (2007) might be used for this by combining the two implied covariance matrices when bootstrapping the distribution of relative model fit tests. Alternatively, one could simply test whether the loadings of one factor are greater than a certain threshold. This, however, requires that the loading matrix is rotated to m dominant factors and one minor factor.

A practical implication of our findings concerns the widespread strategy of justifying multidimensionality of assessment instruments. Frequently, a proposed m-factor model is tested against a restricted 1-factor model in order to justify the assumption of $1 + m$ dimensions but without stepwise comparisons. This procedure tends to overproduce significant results that indicate the rejection of the lower-dimensional model due to the comparatively high sensitivity of the tests and, thus, leads to the assumption of the multidimensional H_1-model. However, based on our findings, we suggest that stepwise comparisons (i.e., comparing 1-, $1 + m$-, $1 + m + 1$-factor models etc.) should be considered when one is interested in whether a model of lower dimensionality is better suited than a proposed m-factor model.

In conclusion, the search for the number of factors is a problem with many facets (Preacher et al., 2013). Results from statistical tests of model fit are only one aspect that should be considered. This, however, does not signify that the statistical techniques do not have to be mathematically sound or that researchers are given a *carte blanche* to abandon good statistical practice. When tests of model fit are considered, they should adhere to the nominal Type-I error rate and be as powerful as possible. The testing strategies suggested here might be such methods.

Acknowledgment

Both authors contributed equally to this work.

References

Akaike, H. (1992). Information theory and an extension of the Maximum Likelihood principle. In S. Kotz & N. Johnson (Eds.), *Breakthroughs in statistics, Vol I, Foundations and basic theory* (pp. 610–624). Springer.

Algina, J. (1980). A note on identification in the oblique and orthogonal factor analysis model. *Psychometrika*, *45*, 393–396. https://doi.org/10.1007/BF02293911

Amemiya, Y., & Anderson, T. (1990). Asymptotic chi-square tests for a large class of factor analysis models. *The Annals of Statistics*, *18*, 1453–1463. https://doi.org/10.1214/aos/1176347760

Bartlett, M. (1950). Tests of significance in factor analysis. *British Journal of Mathematical and Statistical Psychology*, *3*, 77–85. https://doi.org/10.1111/j.2044-8317.1950.tb00285.x

Beran, R., & Srivastava, M. (1985). Bootstrap tests and confidence regions for functions of a covariance matrix. *The Annals of Statistics*, *13*, 95–115. https://doi.org/10.1214/aos/1176346579

Bollen, K., & Stine, R. (1993). Bootstrapping goodness-of-fit measures in structural equation models. In K. Bollen & J. Long (Eds.), *Testing structural equation models* (pp. 111–135). SAGE.

Boomsma, A. (1986). On the use of bootstrap and jackknife in covariance structure analysis. In F. DeAntoni, N. Lauro, & A. Rizzi (Eds.), *Compstat* (pp. 205–210). Physica.

Carroll, J. (1993). *Human cognitive abilities: A survey of factor-analytic studies*. Cambridge University Press.

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

Cheng, C., & Wu, H. (2017). Confidence intervals of fit indexes by inverting a bootstrap test. *Structural Equation Modeling*, *24*, 870–880. https://doi.org/10.1080/10705511.2017.1333432

Costello, A., & Osborne, J. (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://pareonline.net/genpare.asp?wh=0&abt=10

Cudeck, R., & MacCallum, R. (2007). *Factor analysis at 100: Historical developments and future directions*. https://doi.org/10.4324/9780203936764

Davison, A., & Hinkley, D. (1997). *Bootstrap methods and their application*. https://doi.org/10.1017/CBO9780511802843

Efron, B., & Tibshirani, R. (1993). *An introduction to the bootstrap*. Chapman and Hall/CRC.

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

Finch, H. W. (2019). Using fit statistic differences to determine the optimal number of factors to retain in an exploratory factor analysis. *Educational and Psychological Measurement*, *80*, 217–241. https://doi.org/10.1177/0013164419865769

Foldnes, N., Foss, T., & Olsson, U. (2011). Residuals and the residual-based statistic for testing goodness of fit of structural equation models. *Journal of Educational and Behavioral Statistics*, *37*, 367–386. https://doi.org/10.3102/1076998611411920

Furr, R., & Bacharach, V. (2014). *Psychometrics*. SAGE.

Geweke, J., & Singleton, K. (1980). Interpreting the likelihood ratio statistic in factor models when sample size is small. *Journal of the American Statistical Association*, *75*, 133–137. https://doi.org/10.1080/01621459.1980.10477442

Golino, H., & Epskamp, S. (2017). Exploratory graph analysis: A new approach for estimating the number of dimensions in psychological research. *PLoS ONE*, *12*. https://doi.org/10.1371/journal.pone.0174035

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

Green, S., Xu, Y., & Thompson, M. (2018). Relative accuracy of two modified parallel analysis methods that use the proper reference distribution. *Educational and Psychological Measurement*, *78*, 589–604. https : / / doi . org / 10 . 1177 / 0013164417718610

Guttman, L. (1954). Some necessary conditions for common-factor analysis. *Psychometrika*, *19*, 149–161. https://doi.org/10.1007/BF02289162

Hayashi, K., Bentler, P., & Yuan, K.-H. (2007). On the likelihood ratio test for the number of factors in exploratory factor analysis. *Structural Equation Modeling*, *14*, 505–526. https://doi.org/10.1080/10705510701301891

Hopwood, C., & Donnellan, M. (2010). How should the internal structure of personality inventories be evaluated? *Personality and Social Psychology Review*, *14*, 332–346. https://doi.org/10.1177/1088868310361240

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

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*, 1–55. https://doi.org/10.1080/10705519909540118

Huang, P. (2017). Asymptotics of AIC, BIC, and RMSEA for model selection in structural equation modeling. *Psychometrika*, *82*, 407–426. https://doi.org/10.1007/s11336-017-9572-y

Ichikawa, M., & Konishi, S. (1995). Application of the bootstrap methods in factor analysis. *Psychometrika*, *60*, 77–93. https://doi.org/10.1007/BF02294430

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

Kubinger, K. D., Litzenberger, M., & Mrakotsky, C. (2006). Practised intelligence testing based on a modern test conceptualization and its reference to the common intelligence theories. *Learning and Individual Differences*, *16*, 175–193. https://doi.org/10.1016/j.lindif.2005.08.001

Lambert, Z., Wildt, A., & Durand, R. (1991). Approximating confidence intervals for factor loadings. *Multivariate Behavioral Research*, *26*, 421–434. https://doi.org/10.1207/s15327906mbr2603_3

Lawley, D. (1943). The application of the maximum likelihood method to factor analysis. *British Journal of Psychology*, *33*, 172–175. https://doi.org/10.1111/j.2044-8295.1943.tb01052.x

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

Mair, P. (2018). *Modern psychometrics with R*. https://doi.org/10.1007/978-3-319-93177-7

Millsap, R. (2001). When trivial constraints are not trivial: The choice of uniqueness constraints in confirmatory factor analysis. *Structural Equation Modeling*, *8*, 1–17. https://doi.org/10.1207/S15328007SEM0801_1

Montoya, A., & Edwards, M. (2020). The poor fit of model fit for selecting number of factors in exploratory factor analysis for scale evaluation. *Educational and Psychological Measurement*. https://doi.org/10.1177/0013164420942899

Nevitt, J., & Hancock, G. (2001). Performance of bootstrapping approaches to model test statistics and parameter standard error estimation in structural equation modeling. *Structural Equation Modeling*, *8*, 353–377. https://doi.org/10.1207/S15328007SEM0803_2

Preacher, K., 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*, 28–56. https://doi.org/10.1080/00273171.2012.710386

Price, L. (2012). *Psychometric methods: Theory into practice*. Guilford.

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

Rao, C. (1955). Estimation and tests of significance in factor analysis. *Psychometrika*, *20*, 93–111. https://doi.org/10.1007/BF02288983

Revelle, R., & Rocklin, T. (1979). Very simple structure: An alternative procedure for estimating the optimal number of interpretable factors. *Multivariate Behavioral Research*, *14*, 403–414. https://doi.org/10.1207/s15327906mbr1404_2

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*, 282–292. https://doi.org/10.1037/a0025697

Saris, W., Satorra, A., & van der Veld, W. (2009). Testing structural equation models or detection of misspecifications? *Structural Equation Modeling*, *16*, 561–582. https://doi.org/10.1080/10705510903203433

Savalei, V., & Kolenikov, S. (2008). Constrained versus unconstrained estimation in structural equation modeling. *Psychological Methods*, *13*, 150–170. https://doi.org/10.1037/1082-989X.13.2.150

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

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*, 461–464. https://doi.org/10.1214/aos/1176344136

Shapiro, A. (1986). Asymptotic theory of overparameterized structural models. *Journal of the American Statistical Association*, *81*, 142–149. https://doi.org/10.1080/01621459.1986.10478251

Steiger, J., Shapiro, A., & Browne, M. (1985). On the multivariate asymptotic distribution of sequential chi-square statistics. *Psychometrika*, *50*, 253–264. https://doi.org/10.1007/BF02294104

Themessl-Huber, M. (2014). Evaluation of the χ^2-statistic and different fit-indices under misspecified number of factors in confirmatory factor analysis. *Psychological Test and Assessment Modeling*, *56*, 219–236.

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

Yuan, K.-H., Chan, W., Marcoulides, G., & Bentler, P. (2016). Assessing structural equation models by equivalence testing with adjusted fit indexes. *Structural Equation Modeling*, *23*, 319–330. https://doi.org/10.1080/10705511.2015.1065414

Yuan, K.-H., & Hayashi, K. (2003). Bootstrap approach to inference and power analysis based on three statistics for covariance structure models. *British Journal of Mathematical and Statistical Psychology*, *56*, 93–110. https://doi.org/10.1348/000711003321645368

Yuan, K.-H., & Hayashi, K. (2006). Standard errors in covariance structure models: Asymptotics versus bootstrap. *British Journal of Mathematical and Statistical Psychology*, *59*, 397–417. https://doi.org/10.1348/000711005X85896

Yuan, K.-H., Hayashi, K., & Yanagihara, H. (2007). A class of population covariance matrices in the bootstrap approach to covariance structure analysis. *Multivariate Behavioral Research*, *42*, 261–281. https://doi.org/10.1080/00273170701360662

Yung, Y.-F., & Bentler, P. (1996). Bootstrapping techniques in analysis of mean and covariance structures. In G. Marcoulides & R. Schumacker (Eds.), *Advanced structural equation modeling: Issues and techniques* (pp. 195–226). https://doi.org/10.4324/9781315827414

Zhang, X., & Savalei, V. (2016). Bootstrapping confidence intervals for fit indexes in structural equation modeling. *Structural Equation Modeling*, *23*, 392–408. https://doi.org/10.1080/10705511.2015.1118692

Zhu, M., & Ghodsi, A. (2006). Automatic dimensionality selection from the scree plot via the use of profile likelihood. *Computational Statistics & Data Analysis*, *51*, 918–930. https://doi.org/10.1016/j.csda.2005.09.010