Psychological Test and Assessment Modeling, Volume 56, 2014 (3), 219-236

Evaluation of the χ^2-statistic and different fit-indices under misspecified number of factors in confirmatory factor analysis

Michael Themessl-Huber[1]

Abstract

Model evaluation is one of the most important parts in confirmatory factor analysis. There are different criteria to evaluate the fit of a model.

Using data simulation the type-I-risk and the type-II-risk of the χ^2-statistic were investigated. Therefore, a correct specified model and two models with misspecified number of factors were tested under different simulation conditions. The behavior of the *standardized root-mean-square residual (SRMR)*, the *root-mean-square error of approximation (RMSEA)* and the *comparative fit index* (CFI) were also investigated. Cut-off values provided by Hu and Bentler (1999) were used for these fit-indices. To compare different models the χ^2-difference-test or the F-statistic (Kubinger, Litzenberger, & Mrakotsky, 2006) can be used. The behavior of these methods was investigated too.

It was shown, that the χ^2-test did not hold the type-I-risk of 5 %. For the SRMR and the RMSEA different cut-off values should be used under present misspecification. The cut-off values for the CFI seem to be adequate. The F-test is an alternative to the χ^2-difference-test. It has the advantage, that it can be also used even if models are non-nested.

Keywords: confirmatory factor analysis; fit-indices; chi-square statistics; model comparison; F-test

[1] *Correspondence concerning this article should be addressed to:* Michael Themessl-Huber, PhD, Faculty of Psychology, University of Vienna, Austria; email: michael.themessl-huber@univie.ac.at

Introduction

Confirmatory factor analysis (cfa) can be used for a variety of purposes, such as psychometric evaluation, construct validation or for scale development to examine the latent structure of a test instrument (Brown, 2006).

One of the most important aspects in cfa is regarding model fit. There exist different criteria to evaluate the acceptability of the fitted cfa solution. Often global criteria are used for model assessment. Because of some disadvantages of the classical χ^2-statistic, many fit-indices have been developed. With increasing sample size, the value of the χ^2-statistic gets larger. This means, that models might be rejected although the differences between the input matrix and the model implied matrix are negligible. On the other side, a sufficient large sample is needed, so that distributional assumptions are fulfilled (Brown, 2006; Schermelleh-Engel, Moosbrugger, & Müller, 2003). Thus, the χ^2-statistic is strongly affected by sample size.

In contrast to the χ^2-statistic, which is judging exact model fit, fit-indices evaluate a model according to certain indicators. They can be categorized in absolute-fit-indices, incremental-fit-indices and fit-indices adjusting for model parsimony. Each type provides different information about model fit (Brown, 2006). For an overview of fit-indices see Hooper, Coughlan and Mullen (2008) or Schermelleh-Engel et al. (2003). Now, one index out of each category is presented:

1. Absolute fit-indices

   This class of indices evaluates a model without taking other models (more restricted) into account. Model fit is evaluated on an absolute level. The χ^2-statistic is an example for such an index. Another one is the *standardized root mean square residual* (SRMR). The SRMR is defined as the average discrepancy between the covariances in the input matrix and the model-implied matrix. Thus, it is derived from a residual covariance matrix.

$$SRMR = \sqrt{\frac{2\sum_{i=1}^{p}\sum_{j=1}^{i}\left[\frac{s_{ij}}{s_{ii}s_{jj}} - \frac{\widehat{\sigma_{ij}}}{s_{ii}s_{jj}}\right]^2}{p(p+1)}}$$

   p stands for the number of indicators, s_{ij} for the empirical covariances, $\widehat{\sigma_{ij}}$ for the reproduced covariances. The observed standard deviations are given by s_{ii} and s_{jj}.

   The SRMR takes values between 0 and 1. The lower the SRMR, the better the model fit (Brown, 2006; Schermelleh-Engel et al., 2003).

2. Incremental fit-indices

   A given model is evaluated in relation to a more restricted base model. For the base model often a “null model” or “independency model” is chosen, where all covariances among the observed variables are set to 0.

The comparative fit-index is an often used index out of this class:

$$CFI = 1 - \frac{max\left(\chi_T^2 - df_T, 0\right)}{\max\left[(\chi_T^2 - df_T), (\chi_0^2 - df_0), 0\right]}$$

Where χ_T^2 is the χ^2-value of the target-model and χ_0^2 the value of the "null model". The degrees of freedom are given by df_T for the target-model and by df_0 for the null model. The CFI ranges from 0 to 1. Higher values indicate better model fit (Bentler, 1990; Schermelleh-Engel et al., 2003).

3. Parsimony correction

   These indices take the number of parameters of the model into account. There is a penalty term for too large models. Given two different models with the same fit on an absolute level, the one needing less parameter is preferred in this category. One index of this class is the *root mean square error of approximation* (RMSEA):

$$RMSEA = \sqrt{\max\left\{\frac{\chi_T^2 - df_T}{df_T(n-1)}, 0\right\}}$$

   The RMSEA is insensitive to sample size *n*. The upper range of the RMSEA is unbounded, but values greater than 1 are rarely observed. The lower the RMSEA, the better the model fit. A value of 0 indicates perfect fit (Brown, 2006; Steiger & Lind, 1980).

The sampling distribution of fit-indices is unknown, so cut-off values are needed to decide whether a model fits the data or not (Schermelleh-Engel et al., 2003). Given that many researchers evaluate their models using global fit indices, it is important to investigate the usefulness of such rules of thumbs.

In a simulation study, Hu and Bentler (1999) provided cut-off-values for some fit-indices under different simulation conditions. Factor correlations and/or factor loadings were misspecified. They suppose model fit, if:

SRMR < 0.11

CFI > 0.95

RMSEA < 0.08 for n < 250

RMSEA < 0.06 for n ≥ 250

Fit-indices are affected by various aspects such as sample size (Fan, Thompson, & Wang, 1999), model complexity or type of misspecification. Thus, the derived cut-off values depend on the simulation scenarios. Hu and Bentler (1999) were aware about the limitations of their study: "it is difficult to designate a specific cut-off value for each fit-

index because it does not work equally well with various conditions" (p. 27). In spite of the authors warnings, the proposed cut-offs were misinterpreted as "golden rules" by practitioners. Nevertheless, Beauducel and Wittmann (2005) recommended the same cut-off values for these indices.

In another simulation study was shown, that the values of the χ^2-statistics and the fit-indices are affected by factor loadings (Heene, Hilbert, Draxler, Ziegler, & Bühner, 2011). Hu and Bentler (1999) did not vary the factor loadings in their simulation, they used comparatively high loadings between 0.7 and 0.8.

Heene, Hilbert, Freudenthaler and Bühner (2012) investigated the usefulness of the introduced cut-off values in case of misspecified models evoked by unspecified correlated errors. They variied the sample size, the factor loadings and the strength of misspecification in their simulation. The SRMR and RMSEA accepted many false models. The CFI accepted slightly misspecified models.

In this work, the behavior of the χ^2-statistic and the presented fit-indices under misspecified number of factors are investigated. In a simulation, models with too few factors and models with too many factors are tested. The proportion of accepted models is an estimator for the type-II-risk. Also, correct specified models (number of factors is adequate) are tested to calculate the type-I-risk. The cut-off values recommended by Hu and Bentler (1999) are used for the fit-indices.

For model comparison, the two misspecified models are compared with the correct model. Therefore the χ^2-difference-test and the F-Test are used. The latter is an approach by Kubinger, Litzenberger and Mrakotsky (2006). The ratio of two independent random variables with a χ^2-distribution results in a F-distribution:

$$F = \frac{\chi_T^2 / df_T}{\chi_B^2 / df_B}$$

χ_T^2 and χ_B^2 are the χ^2-statistics of the compared cfa models, df_T and df_B are the number of degrees of freedom. The target model is the model under validation. If the F-statistic is significant, then the target model fits worse than the base model. Otherwise it does not fit significantly worse. By using the χ^2-difference-test (if possible), the target model would be the nested model and the base model would be the parent model. This F-statistic might not be exactly F-distributed, because it is possible, that the χ^2-statistics are not based on independent parameter estimations.

The application of the χ^2-difference-test is only possible, when models are nested. Using the F-test, also non-nested models can be compared.

These two methods were also investigated.

Method

The simulation was performed with the statistic software R (R Development Core Team, 2012). The package "latent variable analysis" (lavaan; Rosseel, 2012) was used. In the Appendix can be found an exemplary extract to the R source-code of data generation and model fitting used in the present study.

Three types of cfa population models ("true models") were constructed. Each population model consists of 24 manifest variables, equally distributed over 3 factors. Such models are frequently used in psychological applications. For example, the 24 manifest variables could be the items of a questionnaire in which the items are assumed to load on 3 different factors (personality traits).

Population model with uncorrelated factors / without cross-loadings

The factor loadings for the 24 manifest variables were randomly drawn from a uniform distribution. Factor loadings have an impact on fit-indices and the power of the χ^2-statistic. Therefore, population models with low, medium, high and mixed factor loadings were investigated (cf. Heene et al., 2011). The corresponding loadings were drawn out of the intervals [0.3; 0.5], [0.5; 0.7], [0.7; 0.9] and [0.3; 0.9]. The population model with low loadings (completely standardized solution) is schematically depicted in Table 1. Adding 0.2 and 0.4 to the low loadings, one gets the values for the medium and high loadings, respectively.

The mixed factor loadings were given by:

Factor 1 (x1 to x8): 0.89; 0.54; 0.37; 0.34; 0.45; 0.78; 0.50; 0.88

Factor 2 (x9 to x16): 0.40; 0.58; 0.40; 0.44; 0.76; 0.36; 0.57; 0.35

Factor 3 (x17 to x24): 0.64; 0.31; 0.89; 0.49; 0.68; 0.48; 0.90; 0.84

Population model with correlated factors / without cross-loadings

This model is equivalent to the previous model, with the difference that factors are correlated:

$$\rho_{F1,F2} = 0.5 \mid \rho_{F1,F3} = 0.4 \mid \rho_{F2,F3} = 0.3$$

The same correlation coefficients were used in other simulation studies (Heene et al., 2011; Hu & Bentler, 1999).

Population model with uncorrelated factors / with cross-loadings

For this type of population model, primary and secondary loadings were specified (see Table 1). The loadings were randomly drawn from a uniform distribution. The boundary parameters for the primary loadings were given by 0.3 and 0.9. Cross-loadings were

drawn out of the interval [0; 0.2]. All manifest variables have loadings on every factor. The indicators x1 to x8 have primary loadings on Factor 1, x9 to x16 on Factor 2 and x17 to x24 on Factor 3. The rest are secondary loadings.

All models were tested for sample sizes of n = 150 and n = 600. The simulation design results in a total of 2 (correlation no/yes) x 4 (intervals of factor loadings) x 2 (sample sizes) = 16 conditions. Adding the population models with cross-loadings for both sample sizes, we get 18 simulation scenarios (population models). In each of these 18 conditions, three different models were tested: a 2-Factor model (misspecified, one factor too

Table 1:
Loading structure of population models

	Population model without cross-loadings, low loading condition			**Population model with cross-loadings, with uncorrelated factors**		
Indicator	**Factor 1**	**Factor 2**	**Factor 3**	**Factor 1**	**Factor 2**	**Factor 3**
x1	0.50	0.00	0.00	0.89	0.20	0.08
x2	0.38	0.00	0.00	0.52	0.02	0.14
x3	0.32	0.00	0.00	0.76	0.09	0.00
x4	0.31	0.00	0.00	0.79	0.11	0.17
x5	0.35	0.00	0.00	0.64	0.18	0.00
x6	0.46	0.00	0.00	0.71	0.03	0.04
x7	0.37	0.00	0.00	0.53	0.20	0.18
x8	0.49	0.00	0.00	0.58	0.19	0.12
x9	0.00	0.33	0.00	0.18	0.63	0.08
x10	0.00	0.39	0.00	0.19	0.85	0.09
x11	0.00	0.33	0.00	0.06	0.38	0.01
x12	0.00	0.35	0.00	0.17	0.72	0.19
x13	0.00	0.45	0.00	0.13	0.40	0.09
x14	0.00	0.32	0.00	0.10	0.66	0.19
x15	0.00	0.39	0.00	0.15	0.60	0.18
x16	0.00	0.32	0.00	0.03	0.84	0.13
x17	0.00	0.00	0.41	0.13	0.02	0.54
x18	0.00	0.00	0.30	0.14	0.10	0.32
x19	0.00	0.00	0.50	0.09	0.08	0.34
x20	0.00	0.00	0.36	0.14	0.18	0.58
x21	0.00	0.00	0.43	0.19	0.09	0.41
x22	0.00	0.00	0.36	0.05	0.17	0.63
x23	0.00	0.00	0.50	0.09	0.15	0.61
x24	0.00	0.00	0.48	0.19	0.16	0.38

little), a 3-Factor model (correct specified) and a 4-Factor model (misspecified, one factor too many). For every simulation 10,000 replications were chosen, so all in all 540,000 (18 x 3 x 10,000) models were fitted.

Data on the manifest variables were generated from the population models using the `simulateData` procedure (lavaan; Rosseel, 2012). The simulated observations were multivariate, normally distributed. These raw data served as input for the cfa and was used to evaluate correct and misspecified models with the χ^2-statistic and the fit-indices.

The correct specified 3-Factor model is always in accordance with the corresponding population model. Even cross-loadings are correctly specified.

Depending on whether factor correlations are omitted or permitted in the population models, they are also omitted or permitted in the constructed models. One exception affects the 4-Factor model. Here, the correlation between the additional fourth factor and the other factors is always set to 0. Otherwise many improper solutions would occur, because the parameter estimation of the correlation coefficient between the third and fourth factor would often exceed the value of 1.

The structure and the pattern of the indicator-factor loadings of the misspecified 2-Factor-model and the misspecified 4-Factor model are shown in Table 2. In the 2-Factor-model the indicators which originally loaded on Factor 3, are now loading on Factor 1. In the 4-Factor-model, one factor was split.

The χ^2-statistic and fit-indices were calculated for every fitted model. Thus, in every simulation condition there are 10,000 values for every model and every statistic.

In all conditions and for all fit-statistics, the proportion of accepted 2-Factor, 3-Factor and 4-Factor models was determined.

The proportion of accepted 2-Factor and 4-Factor models is an estimator for the type-II-risk, because these models are misspecified and should be rejected. The proportion of rejected 3-Factor models is an estimator for the type-I-risk. For the χ^2-test the significance level was always 5 %. For the fit-indices, the cut-off values provided by Hu and Bentler (1999) were used. Furthermore, mean values and standard deviation of the χ^2-statistik and fit-indices were calculated for every simulated scenario.

Because of the misspecification, it is possible that improper solutions occur. In this case, the affected model gets rejected without calculating any fit-statistics, because improper solutions indicate gross misspecification (Brown, 2006).

The χ^2-difference-test could be applied, if models were nested. Therefore, the degrees of freedom in the 3-Factor model (parent model) had to differ from those in the 2-Factor and 4-Factor model. This was the case in the population model with cross-loadings. The significance level was 5 %. The proportion of significantly better fitting 3-Factor models in comparison with 2-Factor or 4-Factor models was calculated.

In the same manner the proportion of significantly better fitting 3-Factor models (base-models) compared to the 2-Factor or 4-Factor models was calculated for the F-statistic ($\alpha = 0.05$). For calculation of the F-statistic, the degrees of freedom in the compared models do not have to differ. That means that also non-nested models can be compared when using the F-statistic.

Table 2:
Misspecified models

2-Factor model		4-Factor model			
Factor 1	Factor 2	Factor 1	Factor 2	Factor 3	Factor 4
x1		x1			
x2		x2			
x3		x3			
x4		x4			
x5		x5			
x6		x6			
x7		x7			
x8		x8			
	x9		x9		
	x10		x10		
	x11		x11		
	x12		x12		
	x13		x13		
	x14		x14		
	x15		x15		
	x16		x16		
x17				x17	
x18				x18	
x19				x19	
x20				x20	
x21					x21
x22					x22
x23					x23
x24					x24

Results

The number of improper solutions turned out to be small. In most conditions, the proportion of improper solutions was 0. The clearly highest percentage of improper solutions was 10.5 % and was observed in the condition with the 4-Factor model, n = 150 and mixed loadings.

To estimate the type-I-risk, the proportion of rejected 3-Factor models (correct specified) was calculated in each condition (see Table 3).

Table 3:
Type-I-risk under different simulation scenarios. Mean values of the χ^2-statistic and fit-indices (standard deviation) and [proportion of type-I-errors] for correct specified 3-Factor models for every simulation condition

	Sample size	
Simulation condition	n = 150	n = 600
Uncorrelated		
Chi-square[a]		
Low	270.455 (24.142) [0.208]	256.313 (22.604) [0.075]
Medium	270.985 (24.207) [0.214]	256.418 (22.644) [0.078]
High	271.151 (24.232) [0.215]	256.507 (22.623) [0.075]
Mixed	271.100 (24.290) [0.218]	256.464 (22.534) [0.076]
SRMR		
Low	0.072 (0.004) [0.001]	0.036 (0.002) [0.000]
Medium	0.069 (0.006) [0.000]	0.035 (0.003) [0.000]
High	0.066 (0.011) [0.002]	0.033 (0.006) [0.000]
Mixed	0.070 (0.006) [0.000]	0.035 (0.003) [0.000]
CFI		
Low	0.912 (0.078) [0.613]	0.987 (0.017) [0.050]
Medium	0.975 (0.023) [0.161]	0.997 (0.005) [0.000]
High	0.991 (0.008) [0.000]	0.999 (0.002) [0.000]
Mixed	0.981 (0.018) [0.063]	0.997 (0.003) [0.000]
RMSEA		
Low	0.020 (0.014) [0.001]	0.006 (0.006) [0.000]
Medium	0.020 (0.014) [0.000]	0.006 (0.006) [0.000]
High	0.020 (0.014) [0.000]	0.006 (0.006) [0.000]
Mixed	0.020 (0.014) [0.000]	0.006 (0.006) [0.000]
Correlated		
Chi-square[b]		
Low	267.305 (24.119) [0.206]	253.248 (22.480) [0.074]
Medium	267.946 (24.206) [0.214]	253.463 (22.477) [0.074]
High	267.972 (24.166) [0.212]	253.400 (22.458) [0.076]
Mixed	267.964 (24.261) [0.216]	253.376 (22.352) [0.077]
SRMR		
Low	0.068 (0.003) [0.001]	0.034 (0.002) [0.000]
Medium	0.061 (0.004) [0.000]	0.030 (0.002) [0.000]
High	0.047 (0.004) [0.000]	0.024 (0.002) [0.000]
Mixed	0.061 (0.004) [0.000]	0.031 (0.002) [0.000]

CFI		
Low	0.920 (0.072) [0.589]	0.988 (0.016) [0.034]
Medium	0.976 (0.022) [0.139]	0.997 (0.004) [0.000]
High	0.992 (0.008) [0.000]	0.999 (0.001) [0.000]
Mixed	0.982 (0.017) [0.051]	0.998 (0.003) [0.000]
RMSEA		
Low	0.020 (0.014) [0.001]	0.006 (0.006) [0.000]
Medium	0.020 (0.014) [0.000]	0.006 (0.006) [0.000]
High	0.020 (0.014) [0.000]	0.006 (0.006) [0.000]
Mixed	0.020 (0.014) [0.000]	0.006 (0.006) [0.000]
Cross-loadings		
Chi-square[c]	224.296 (22.146) [0.254]	211.049 (20.694) [0.097]
SRMR	0.041 (0.003) [0.000]	0.020 (0.001) [0.000]
CFI	0.983 (0.014) [0.025]	0.998 (0.003) [0.000]
RMSEA	0.023 (0.014) [0.000]	0.007 (0.007) [0.000]

Note. [a]df =252. [b]df = 249. [c]df = 204.

In the small sample size condition (n = 150) the χ^2-test did not hold the nominal type-I-risk of 5 %. Depending on the condition, between 20.6 % and 25.4 % of correct specified models were rejected. In some way, the RMSEA and the SRMR showed better results. The proportion of wrong-rejection rate was nearly 0. The mean values of both fit-indices were always below their cut-off values (RMSEA < 0.08 und SRMR < 0.11). The CFI seems to have difficulties to accept correctly specified models when factor loadings are low or medium. In the low-loading condition, the percentage of wrong-rejection rate lay between 0.589 and 0.613, in the medium-loading condition between 0.139 and 0.161. For low loadings and correlated factors the mean value of the CFI was 0.92 and thus below the cut-of-value of 0.95.

If sample size is larger (n = 600), the χ^2-test accepted more correct specified models. The wrong-rejection rate was between 0.074 and 0.097. So the nominal type-I-risk is lower, but still tops 5 %. The SRMR and RMSEA accepted every model. The mean values of the fit-indices are far smaller than the cut-off values. With larger sample size, the behavior of the CFI is improving. Only between 0 % and 5 % of models were not accepted. The mean values were between 0.988 and 0.999 and are therefore larger than the cut-off value.

The proportion of accepted 2-Factor or 4-Factor models is an estimator for the type-II-risk. On the basis of similar results for both models, only the findings for the 4-Factor models are presented. In Table 4 mean values, standard deviations and proportions of accepted models are depicted for the 4-Factor model.

In the low loading condition and with a sample size of n = 150, the χ^2-test accepted between 20.6 % and 23.7 % of false specified models. Hence, the power is at least 0.763. In all other conditions, the χ^2-test rejected the false specified models.

Table 4:
Type-II-risk under different simulation scenarios for the 4-Factor model. Mean values of the χ^2-statistic and fit-indices (standard deviation) and [proportion of type-II-errors] for false specified 4-Factor models for every simulation condition

	Sample size	
Simulation condition	n = 150	n = 600
Uncorrelated		
Chi-square[a]		
Low	307.843 (26.547) [0.237]	405.687 (32.425) [0.000]
Medium	381.955 (29.989) [0.000]	700.372 (42.076) [0.000]
High	529.918 (32.772) [0.000]	1291.168 (50.489) [0.000]
Mixed	471.557 (32.155) [0.000]	1057.488 (48.216) [0.000]
SRMR		
Low	0.084 (0.005) [0.938]	0.055 (0.003) [1.000]
Medium	0.113 (0.007) [0.315]	0.095 (0.004) [1.000]
High	0.169 (0.008) [0.000]	0.159 (0.003) [0.000]
Mixed	0.128 (0.006) [0.002]	0.112 (0.003) [0.224]
CFI		
Low	0.762 (0.094) [0.031]	0.817 (0.033) [0.000]
Medium	0.848 (0.031) [0.000]	0.866 (0.011) [0.000]
High	0.890 (0.012) [0.000]	0.896 (0.005) [0.000]
Mixed	0.805 (0.024) [0.000]	0.818 (0.009) [0.000]
RMSEA		
Low	0.037 (0.010) [0.938]	0.032 (0.003) [1.000]
Medium	0.058 (0.007) [0.999]	0.054 (0.003) [0.979]
High	0.086 (0.005) [0.115]	0.083 (0.002) [0.000]
Mixed	0.076 (0.006) [0.654]	0.073 (0.002) [0.000]
Correlated		
Chi-square[b]		
Low	308.305 (26.920) [0.206]	418.223 (33.215) [0.000]
Medium	382.407 (30.005) [0.000]	711.751 (42.552) [0.000]
High	528.423 (32.712) [0.000]	1295.357 (50.510) [0.000]
Mixed	470.666 (32.257) [0.000]	1065.802 (47.895) [0.000]
SRMR		
Low	0.086 (0.006) [0.960]	0.061 (0.004) [1.000]
Medium	0.125 (0.011) [0.062]	0.113 (0.006) [0.294]
High	0.196 (0.014) [0.000]	0.191 (0.007) [0.000]
Mixed	0.143 (0.010) [0.000]	0.132 (0.005) [0.000]

CFI		
Low	0.771 (0.087) [0.026]	0.820 (0.030) [0.000]
Medium	0.852 (0.029) [0.000]	0.869 (0.011) [0.000]
High	0.892 (0.012) [0.000]	0.898 (0.004) [0.000]
Mixed	0.812 (0.023) [0.000]	0.824 (0.009) [0.000]
RMSEA		
Low	0.039 (0.010) [0.960]	0.033 (0.003) [1.000]
Medium	0.059 (0.007) [1.000]	0.056 (0.003) [0.942]
High	0.086 (0.005) [0.090]	0.084 (0.002) [0.000]
Mixed	0.077 (0.006) [0.638]	0.074 (0.002) [0.000]
Cross-loadings		
Chi-square[c]	421.308 (35.739) [0.000]	857.727 (56.800) [0.000]
SRMR	0.161 (0.019) [0.001]	0.151 (0.010) [0.000]
CFI	0.871 (0.023) [0.000]	0.883 (0.009) [0.000]
RMSEA	0.067 (0.007) [0.943]	0.063 (0.003) [0.107]

Note. [a]*df =252.* [b]*df = 249.* [c]*df = 252.*

For the SRMR, the proportion of wrongly accepted models in the condition with low and medium loadings is surprisingly high. This is the case for both sample sizes. In these scenarios the proportion of accepted models is between 0.062 and 1. The mean values ranged from 0.055 to 0.125.

With a sample size of n = 150, the RMSEA had problems to detect the misspecified 4-Factor models. For example, in the mixed loading condition the percentage of wrongly accepted models is between 63.8 % and 94.3 %. The problem held on for the larger sample size (n = 600) in the low and medium loading conditions. The mean values for n = 150 were between 0.037 and 0.086, for n = 600 between 0.032 and 0.084.

The CFI completely rejected the false specified models in most conditions. Only in two scenarios the failure-to-reject rate was greater than 0. For n = 150 and low loadings, the proportion of accepted models was 0.031 in the uncorrelated and 0.026 in the correlated population model. The mean values of the CFI ranged from 0.762 to 0.898. They were beyond the cut-off value.

To compare different models, the F-test and the χ^2-difference-test were used. The degrees of freedom differed in the population model with cross-loadings for the three models. Hence, the χ^2-difference-test could be applied. The 3-Factor model served as parent model. The F-test could be computed in every scenario, because models do not have to be nested. For calculation of the F-test, the 3-Factor model served as the base model. It was tested whether the misspecified 2-Factor and 4-Factor models fit significantly worse than the 3-Factor model. The proportions of rejections are presented in Table 5.

Table 5:
F-test and χ^2-difference-test. Proportions of rejected 2-Factor and 4-Factor models in comparison with 3-Factor models

Simulation-condition	F-test		χ^2-difference-test	
	2-Factor model	4-Factor model	2-Factor model	4-Factor model
Uncorrelated				
n = 150				
Low	0.536	0.091		
Medium	1.000	0.996		
High	1.000	1.000		
Mixed	1.000	1.000		
n = 600				
Low	1.000	1.000		
Medium	1.000	1.000		
High	1.000	1.000		
Mixed	1.000	1.000		
Correlated				
n = 150				
Low	0.123	0.113		
Medium	1.000	0.997		
High	1.000	1.000		
Mixed	1.000	1.000		
n = 600				
Low	1.000	1.000		
Medium	1.000	1.000		
High	1.000	1.000		
Mixed	1.000	1.000		
Cross-loadings				
n = 150	1.000	1.000	1.000	1.000
n = 600	1.000	1.000	1.000	1.000

In nearly all conditions the F-test completely rejected the misspecified 2-Factor and 4-Factor models. Only in the sample size condition n = 150, the 4-Factor model was often accepted when loadings were low. The F-test as well as the χ^2-difference-test rejected all misspecified models in comparison with the correct specified model in the population model with cross-loadings.

Only those 3-Factor models were considered for model comparison, which were accepted by the χ^2-test. The results are nearly the same, if all models (inclusive rejected 3-Factor models by the χ^2-test) were considered.

If the 2-Factor or the 4-Factor model is serving as base model for the F-test, the proportion of rejected 3-Factor models is 0 in all conditions. Hence, no correct specified model was rejected by comparison with the other two models.

Discussion

The χ^2-test did not hold the nominal type-I-risk of 5 %, especially not for n = 150. The χ^2-statistic is based on a Maximum-Likelihood estimation. These estimators have excellent asymptotical properties, but the behavior by small sample sizes may be problematic. For a correct specified model, the expected value for the χ^2-statistic equals the number of degrees of freedom. But only with a sufficiently large sample size, the test statistic follows a χ^2-distribution (Bollen, 1990). In the present paper this problem has been quantified.

For n = 600, the χ^2-test still exceeded the nominal type-I-risk, but not that much. That means, the value of the statistic is moving closer to the expectation value (the number of degrees of freedom), when sample size gets larger. In the low loading and small sample size condition, the χ^2-test accepted some false specified models.

It is a difficult task to find the optimal sample size for the χ^2-statistic. Is the sample size too small, the statistic is not χ^2-distributed, is the sample size too large, lowest model deviations will lead to rejection of the model. Muthén & Muthén (2002) recommend a sample size of n = 150 for multivariate, normally distributed data and correct specified models. This sample size did not lead to best results in this study. The optimal sample size depends on model complexity (Kenny & McCoach, 2003).

The fit-indices SRMR and RMSEA almost accepted all correct specified models, the wrong-rejection rate is nearly 0. On the other hand, both indices did not reject many false specified models. The failure-to-reject rate is very high. All in all, they rejected few models. The used cut-off values seem not to be the best choice under present misspecification. Maybe the threshold values for the cut-offs should be lowered for both indices. Then less (false specified) models would be accepted.

The CFI rejected many correct specified models in the condition with small sample size and low loadings (> 58 %). In the higher loading conditions, the results are better. For false specified models, the CFI showed the best behavior of all statistics. In most scenarios all misspecified models were rejected. The cut-off values for the CFI are appropriate.

All statistics have problems to detect misspecified models, when factor loadings are low. At first glance, one will think that these low loadings are not representative. But Peterson (2000) showed in a meta analysis that the mean value of factor loadings in psychology questionnaire is 0.32, with 50 % of loadings between 0.23 and 0.37. That means, the low loadings in this study are common in psychology research.

Hu and Bentler (1999) proposed a 2-index strategy. The combination of two fit-indices should be used for model evaluation. This approach is based on the assumption that some indices are more sensitive against certain misspecification than others. But Fan and Sivo (2005) showed that this assumption is not always met.

For model comparison the F-test is an alternative to the χ^2-difference-test, because it can be also used, if the number of degrees of freedom is the same in the compared models. Models do not have to be nested. The researcher can decide which model serves as base model or target model, respectively. Given that the χ^2-difference-test could just be computed in the population model with cross-loadings, it is difficult to compare both methods. In this simulation condition no differences between the methods were found. Both tests rejected all false specified models.

Although three different population models were used and the sample size and factor loadings were varied, the results are bound to the specific simulation conditions. Future research could investigate the influence of the number of manifest and latent variables. Those parameters were not varied in this study. Nonetheless, some recommendations for practitioners are given.

From the three fit-indices, the CFI showed the most reliable results. Therefore the use of the CFI is recommended. For model comparison, the F-test is a helpful tool, especially when models are non-nested. In this case, the χ^2-difference-test can not be applied at all.

To check if the number of replications (10,000) per simulation condition was large enough to get reliable results, one condition was also executed with 100,000 replications. For both numbers of replications, the results were nearly the same. At times, there were small differences in the third position after decimal point, but this is negligible.

References

Beauducel, A., & Wittmann, W. W. (2005). Simulation study on fit indexes in CFA based on data with slightly distorted simple structure. *Structural Equation Modeling, 12*, 41-75, DOI: 10.1207/s15328007sem1201_3.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin, 107*, 238-246.

Bollen, K. A. (1990). Overall fit in covariance structure models: Two types of sample size effects. *Psychological Bulletin, 107*, 256-259.

Brown, T. A. (2006). *Confirmatory Factor Analysis for Applied Research.* New York: Guilford.

Fan, X., & Sivo, S. A. (2005). Sensitivity of fit indexes to misspecified structural or measurement model components: Rationale of two-index strategy revisited. *Structural Equation Modeling, 12*, 343-367, DOI: 10.1207/s15328007sem1203_1.

Fan, X., Thompson, B., & Wang, L. (1999). Effects of sample size, estimation methods, and model specification on structural equation modeling fit indexes: A Multidisciplinary Journal. *Structural Equation Modeling, 6*, 56-83, DOI: 10.1080/10705519909540119.

Heene, M., Hilbert, S., Draxler, C., Ziegler, M., & Bühner, M. (2011). Masking Misfit in Confirmatory Factor Analysis by Increasing Unique Variances: A Cautionary Note on the Usefulness of Cutoff Values of Fit Indices. *Psychological Methods, 16*, 319-336, DOI: 10.1037/a0024917.

Heene, M., Hilbert, S., Freudenthaler, H. H., & Bühner, M. (2012). Sensitivity of SEM Fit Indexes With Respect to Violations of Uncorrelated Errors. *Structural Equation Modeling: A multidisciplinary Journal, 19:1*, 36-50, DOI: 10.1080/10705511.2012.634710.

Hooper, D., Coughlan, J., & Mullen, M. (2008). Structural Equation Modelling: Guidelines for Determining Model Fit. *Electronic Journal of Business Research Methods, 6*, 53-60.

Kenny, D. A., & McCoach, D. B. (2003). Effects of the Number of Variables on Measures of Fit in Structural Equation Modeling. *Structural Equation Modeling: A Multidisciplinary Journal ,10*, 333-351, DOI: 10.1207/S15328007SEM1003_1.

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, DOI: 10.1016/j.lindif.2005.08.001.

Muthén, L. K., & Muthén, B. O. (2002). How to use a Monte Carlo study to decide on sample size and determine power. *Structural Equation Modeling, 4*, 599-620.

Peterson, R. A. (2000). A meta analysis of variance accounted for and factor loadings in exploratory factor analysis. *Marketing Letters, 11*, 261-275.

R Development Core Team. (2012). *R: A language and environment for statistical computing.* Vienna: R Foundation for Statistical Computing http://www.R-project.org/.

Rosseel, Y. (2012). lavaan: An R Package for Structural Equation. *Journal of Statistical Software , 48* (2).

Schermelleh-Engel, K., Moosbrugger, H., & Müller, H. (2003). Evaluating the Fit of Structural Equation Models. *MPR-online, 8*, 23-74.

Steiger, J. H., & Lind, J. M. (1980). Statistically based tests for the number of common factors. *Paper presented at the meeting of the Psychometric Society, Iowa City, IA.*

Appendix

```
#Population model with low factor loadings and uncorrelated
#factors

pop.model.low <- '
f1 =~ 0.50*x1 + 0.38*x2 + 0.32*x3 + 0.31*x4 + 0.35*x5 + 0.46*x6 +
     0.37*x7 + 0.49*x8
f2 =~ 0.33*x9 + 0.39*x10 + 0.33*x11 + 0.35*x12 + 0.45*x13 + 0.32*x14 +
     0.39*x15 + 0.32*x16
f3 =~ 0.41*x17 + 0.30*x18 + 0.50*x19 + 0.36*x20 + 0.43*x21 + 0.36*x22
     + 0.50*x23 + 0.48*x24
f1 ~~ 0*f2
f1 ~~ 0*f3
f2 ~~ 0*f3'


#Models to be tested

#2-Factor-model (misspecified, one factor too less)
Model_2factor <- '
f1 =~ x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8 + x17 + x18 + x19 + x20 +
     x21 + x22 + x23 + x24
f2 =~ x9 + x10 + x11 + x12 + x13 + x14 + x15 + x16
f1 ~~ 0*f2'

#3-Factor-model (correct specified)
Model_3factor <- '
f1 =~ x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8
f2 =~ x9 + x10 + x11 + x12 + x13 + x14 + x15 + x16
f3 =~ x17 + x18 + x19 + x20 + x21 + x22 + x23 + x24
f1 ~~ 0*f2
f1 ~~ 0*f3
f2 ~~ 0*f3'

#4-Factor-model (misspecified, one factor too many)
Model_4factor <- '
f1 =~ x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8
f2 =~ x9 + x10 + x11 + x12 + x13 + x14 + x15 + x16
f3 =~ x17 + x18 + x19 + x20
f4 =~ x21 + x22 + x23 + x24
f1 ~~ 0*f2
f1 ~~ 0*f3
f1 ~~ 0*f4
f2 ~~ 0*f3
f2 ~~ 0*f4
f3 ~~ 0*f4'
```

```
#Data generation and fitting of the models for sample size
#n = 150

library("lavaan")  #loading library lavaan
n <- 150           #sample size

#Data generation
#Variances of the latent variables is set to 1, therefore "std.lv = T"
#Factor loadings (population model) are completely standardized,
#therefore "standardized = T"
myData <- simulateData(model = pop.model.low, model.type = "cfa",
    std.lv = T, standardized = T, sample.nobs = n)

#Fitting of the models
fit_2factor <- cfa(Model_2factor, data = myData, std.lv = TRUE)
fit_3factor <- cfa(Model_3factor, data = myData, std.lv = TRUE)
fit_4factor <- cfa(Model_4factor, data = myData, std.lv = TRUE)

#Chi-square statistic and Fit-Indices are saved
index_2factor <- fitMeasures(fit_2factor, c("chisq", "pvalue", "cfi",
    "rmsea", "srmr"))
index_3factor <- fitMeasures(fit_3factor, c("chisq", "pvalue", "cfi",
    "rmsea", "srmr"))
index_4factor <- fitMeasures(fit_4factor, c("chisq", "pvalue", "cfi",
    "rmsea", "srmr"))
```