Testing the number of factors in EFA and CFA

Tihomir Asparouhov and Bengt Muthén

Mplus Web Note No.25
Version 1

December 14, 2024

1 Introduction

In this note we discuss the commonly occurring situation where the number of factors in the model is unknown and needs to be determined. Usually this problem is solved by estimating a sequence of consecutive models with an increasing number of factors. Suppose that H_m is the estimated model with m factors, where $m = 1, ..., M$. Suppose that H_0 is the unrestricted variance covariance model. Model H_m is nested within H_0 for any m. The models H_m, however, are not necessarily nested within each other. Here there is a difference between the EFA and CFA modeling frameworks. If H_m is the EFA model with m factors for every m, then the factor models are nested within each other. In the CFA modeling framework, however, the models are not always nested. One situation where H_m is nested within H_{m+1} in the CFA framework is the case where the indicators of the $m-$th factor in model H_m are the indicators for the $m-$th and the $m+1$-th factors in the H_{m+1} model. The nesting of the models is important as it determines which techniques are available to use in determining m. For example, the likelihood ratio test (LRT) is available only when the models are nested.

There are four general paths to determine the value of m.

- P1: Comparing H_m against H_0. Select m as the smallest m yielding a favorable comparison. This method does not require that the models are nested.
- P2: Comparing H_m against H_{m+1}. This method requires nested models. Select m as the smallest m not rejected in favor of the model with one more factor
- P3: Comparing the entirety of H_m models with information criteria such as BIC. This method does not require that the models are nested.
- P4: Select m based on the eigenvalues of the sample correlation matrix. This applies only to EFA.

Each of these paths has some variations. For example, the comparison in P2 can be done with LRT, Wald test, robust LRT testing (Satorra and Bentler (2001)), strictly positive robust LRT testing (Satorra and Bentler (2010)). Similarly, P1 can be accomplished by chi-square test of fit or various fit indices. For Method P3, AIC or BIC can be used. Method P4 can be done with Kaiser's rule or with Parallel Analysis.

In addition to these methodological alternatives, consideration must be given to the fact that H_m might not be estimable. Consider for example the EFA model. Typically, there is a limit to the number of factors that can be estimated. That is, the EFA model can be estimated with $1, 2, ..., m_0$ but any EFA model with more than m_0 factors fails to converge. The convergence issue can also be of various types: the likelihood optimization algorithm fails, the information matrix is singular, the rotation optimization fails, the information matrix for the rotated solution is singular, or the estimated solution is inadmissible because of negative residual variance of the indicators or non-positive definite matrix for the factors. Clearly the convergence issue itself is setting up a selection mechanism: using the highest number of factors for which there is an admissible converging solution. Furthermore the convergence of the estimation may depend on the estimator that is used (ML, MLR or MLF) or the algorithm used for the estimation. For example, the EFA model in Mplus can be estimated in three different ways: using type=EFA classic algorithm, using an ESEM model, and using a PSEM model with penalty set to a rotation criteria for the loadings. The convergence limit m_0 for the different estimators can be different.

In what follows we illustrate these methods and consider their effectiveness in a variety of situations.

2 Determining the number of factors in EFA

Enumerating the number of factors in EFA analysis is a classic problem which has been discussed in many articles, see Schmitt (2011) and Hayashi et al. (2007) for an overview.

Here we illustrate some practical aspects of factor enumeration in EFA. We use a simulated example based on 10 normally distributed indicators measuring 3 factors. Each factor has 3 pure indicators (indicators with non-zero loading on one factor only) and one of the indicators loads on all 3 factors. We use a sample with 500 individuals and we generate and analyze 100 data sets. We use ESEM to analyze the data and to estimate the EFA model with 1,2,3 and 4 factors. For each of the analyses we use 50 random starting values to ensure full maximization of the likelihood. Figure 1 shows the input file for the simulation study for EFA with 3 factors. The results of the factor enumeration for all methods are given in Table 1.

2.1 Convergence

The models with $m = 1, 2, 3$ had no convergence issues. The model with $m = 4$ has 54% convergence problems of the 4 types listed above (point estimates and standard errors for unrotated and rotated solution). Those are completely thrown out automatically by the Mplus Montecarlo algorithm. This however does not include admissibility of solution. If we add admissibility of solution to the criteria for a valid model, the rate of non-convergence is 65%. In fact 64% of the models with 4 factors had inadmissible solutions. Thus we can conclude that at the core of convergence issues arise from inadmissible solutions.

Convergence rates increase as the sample size increases. At $N = 1000$, the 4 factor model improves convergence and admissibility by another 6% and it drops by another than 10% with $N = 200$. At $N = 200$, the 3 factor model also yields inadmissible solutions 16% of the time.

2.2 Chi-square test of fit

To select the number of factors m we consider the test of fit and the corresponding p-value for all converging models. The number m is the smallest m for which the chi-square test of fit does not reject the H_m model. If the model is rejected for all H_m that converge, we select the converging model with the smallest p-value (typically this is the largest m for which H_m converges).

In our simulation, H_1 is rejected in all 100 replications, H_2 is rejected in 96 replications and thus the selection is $m = 2$ for 4 replications. Model H_3 is rejected in 3 replications only but in those 3 replications the model with 4 factors does not converge (and therefore the model with 3 factors must be selected).

One important observation is that underestimation in the number of factors is a matter of power. With $N = 500$ in 4% of the replications we failed to establish the need for 3 factors. If we use $N = 1000$ that percentage is 0%, i.e., with larger sample size, the power to detect the three factors increases to 100%. With $N = 200$, however, the model with 2 factors is not rejected in 38% of the replications, i.e., the power to detect all 3 factors decreases.

It should be noted also that the test of fit has an expected type I error of 5%. That is, with this method, we expect that the 3-factor model would be rejected 5% of the time in favor of a higher number of factors. In our simulation that type I error is 3%. However, the 4-factor model is not converging

for those replications and thus, the resulting type I error is actually 0%.

The test of fit method is implemented in Mplus when using EFA estimation via the ANALYSIS option TYPE=EFA 1 m; which estimates a sequence of models with 1 factor, 2 factors, ..., m factors.

2.3 Sequential LRT

The sequential LRT approach proceeds as follows. For each replication we estimate H_m for $m = 1, .., 4$, and we test H_m against H_{m+1} if both models converge to admissible solutions. We select m to be the largest value not rejected by the model with $m + 1$ factors. If the model with $m + 1$ factors does not converge, that is interpreted as acceptance of the model with m factors.

In our simulation, the LRT test of 1 factor vs 2 factors rejects the 1 factor model 100% of the time. The LRT test of 2 factor vs 3 factors rejects the 2 factor model 100% of the time. The 4 factor model converged to an admissible solution in 35% of the replications. Of those 35% of the replications, the 3-factor model was rejected in favor of the 4-factor model 4% of the replications. We can interpret the 4% error as the expected 5% type I error.

Hayashi et al. (2007) illustrate that the type I error for sequential LRT testing in EFA is actually higher than expected resulting in overfactoring. This is due to the fact that the LRT is applied to a hypothesis which is on the border of admissible space and therefore there is a deviation in the LRT test statistics from the expected chi-square distribution. To understand why the test is at the border of the admissible space we must understand how an m-factor EFA model is nested within the $m + 1$-factor EFA model. More specifically, what values of the parameters in the $m + 1$-factor model makes it equivalent to an m-factor EFA model. For $m = 1$, it is very simple. The 2-factor model becomes equivalent to a 1-factor model when the factor correlation is ± 1. More generally, for an arbitrary value of m, this happens precisely when the correlation matrix of the $m + 1$-factor model is singular (determinant of 0) or equivalently there is a linear relationship between the factors (one of the factors is a linear combination of the others). That is, the $m + 1$-factor EFA model becomes equivalent to an m-factor EFA model precisely when the determinant of the correlation matrix is 0. Thus, the LRT test for m v.s. $m + 1$ factors is essentially a test for the determinant of the factor correlation matrix to be positive. This determinant value is at the border of the admissible space since an admissible factor correlation matrix

is positive definite and therefore has a determinant greater than 0.

Simulations reported in Hayashi et al. (2007) lead to type I error of around 15%. What we see here, however, is that overfactoring is not much of a problem once the inadmissibility/convergence of the solutions is taken into account. Since only a third of the replications were admissible, essentially the type I error is divided by 3. If we only consider the log-likelihood value for the 4-factor model and we ignore the additional analysis produced by Mplus where rotation, information matrix condition number, and admissibility of the parameters, we would expect the type I error to be higher as in Hayashi et al. (2007).

The main conclusions of this simulation however is that both the test of fit and the sequential LRT work about equally well. The power of the sequential LRT is somewhat bigger than that of the test of fit.

The sequential method is implemented in Mplus when using EFA estimation via the option TYPE=EFA 1 m.

2.4 Fit indices

In this section we explore factor enumeration based on the fit indices CFI, TLI, RMSEA, and SRMR. We use the cutoff values suggested by Hu and Bentler (1999): CFI > 0.95, TLI > 0.95, RMSEA < 0.06 and SRMR < 0.08. As with the chi-square test of fit, the number of factors is the smallest m for which the fit index yields an acceptable model, and if such does not exist we use the largest m with converging H_m. All of these fit indices do not have enough power to detect all 3 factors, and select $m = 2$ most often. Larger sample sizes will not improve the situation. Asymptotically, all the fit indices converge to a particular value that measures the distance between the modeled correlation matrix and the sample correlation matrix. The fact that these fit indices yield $m = 2$ means that the best 2-factor EFA model yields a decently close approximation to what the 3-factor EFA model produces.

2.5 Information criteria

In this section we explore factor enumeration based on the information criteria AIC and BIC. The number of factors m is selected based on the smallest information criteria among the converging and admissible H_m models. In this simulation study, AIC appears to work better (selects $m = 3$ 96% of the time and $m = 4$ 4% of the time) than BIC (selects $m = 2$ 52% of the time

and $m = 3$ 48% of the time) but asymptotically, only BIC is guaranteed to select the correct model. It is important to note here that BIC has lower power than LRT: both test of fit and the sequential testing.

Tables 2 and 3 also show the average AIC and BIC values for different sample sizes where the lowest in each row is bolded. These values confirm the known behavior of the criteria. BIC will pick the correct model with sufficiently large sample size while the AIC is susceptible to overfactoring and increasing the sample size will not resolve that problem. At $N = 1000$, BIC has substantially lower value at $m = 3$ than at $m = 2$ or $m = 4$, while AIC has about the same value for $m = 3$ and $m = 4$. If we are to plot these values as a function of m, the BIC curve will look parabolic with a pronounced low point at $m = 3$, while the AIC curve will look like a flat line around $m = 3$ and $m = 4$. This emphasizes the ability of BIC to pick the correct number of factors more easily in asymptotic settings.

At $N = 1000$, the BIC picks the correct number of factors $m = 3$ in 99% of the replications and in only 1% of the replications it picks $m = 2$. For that sample size, the AIC picks $m = 3$ in 94% of the replications and $m = 4$ in 6% of the replications.

It should be noted here that AIC and BIC have a minor incompatibility with the EFA model, particularly the BIC. The AIC and BIC incorporate penalties for every model parameter. The BIC has a stronger parameter penalty. In this simulation study at $N = 500$, the BIC parameter penalty is more than 3 times larger than the parameter penalty of the AIC. In the EFA settings, there are many parameters (cross-loadings) that are essentially zero. These parameters do not improve the model fit but only contribute to that penalty. As a result, for the EFA model, AIC and BIC have unfair penalties. In non-exploratory models, a parameter is included in the model because it improves the model fit and the model fit improvement is counterbalanced by the penalty so that we do not overfit the data. That is, in non-exploratory models, a parameter is included in the model only if it provides a model fit improvement that is bigger than the penalty. In the EFA framework, this logic is broken and many parameters are included that do not improve the fit but only contribute to the penalty. As a result, for moderate sample sizes, the BIC criteria will have the tendency to under-factor in the EFA settings. Nevertheless, the BIC will overcome this disadvantage asymptotically.

2.6 Methods based on the sample correlation eigenvalues

There are three different methods in Mplus for determining the number of factors using the sample correlation matrix's eigenvalues. The first method is the Kaiser's rule which simply counts the number of eigenvalues greater than 1.

The second method is the Parallel Analysis (PA) method which counts the number of eigenvalues greater than the mean eigenvalues of randomly generated matrices. That is, the sample correlation eigenvalues are computed and ordered. Several random independent variable samples are also generated. The sample correlation is computed for each random sample. The eigenvalues for each random sample correlation matrix are computed and ordered. The random sample eigenvalues are averaged across samples. We then simply count the number of sample correlation eigenvalues greater than the corresponding average random sample eigenvalues. The comparison is done as follows: the largest sample correlation eigenvalue is compared with the average of the largest random sample eigenvalue. Then the second largest eigenvalues are compared, etc. There are a number of variations to this method. More information can be found in Timmerman and Lorenzo-Seva (2011).

The third method available in Mplus is the Parallel Analysis which uses the 95 percentile of the distribution of the random matrix eigenvalues instead of the average. We denote this method by PA_{95}.

In our simulation study, all three methods conclude that the number of factors is 2 in 100% of the replications. Increasing the sample size does not resolve the problem and the third eigenvalue never increases beyond 0.8, while the eigenvalues of the random matrices converge to 1. It appears that the eigenvalues method can detect major factors only but the method is likely not suitable for situations where larger correlations between the factors can be found. In this simulation study we used a correlation of 0.7 between the first two factors. Nevertheless, because the method does not improve asymptotically, it is difficult to justify it. At least three methods already mentioned above are guaranteed to work fairly well asymptotically: BIC, sequential LRT testing, and the test of fit. Thus, it is difficult to justify the eigenvalue methods beyond using it as an approximation with the known assumption that the method is likely under-counting the number of factors.

2.7 Methods based on Wald test or Z-test

Testing hypotheses can be done with the log-likelihood (test of fit and sequential LRT) or with the standard errors of the estimated parameters. Here we explore the possibility for factor enumeration using testing based on the standard errors of the model parameters. If we use a single parameter hypothesis, we essentially are using a Z-score test. If we use multiple parameter hypotheses, we are using a Wald test (Model test in Mplus).

If we want to test a 1-factor EFA model v.s. a 2-factor EFA model, we can simply consider the hypothesis that the factor correlation c between the two factors is ± 1. That is, using the confidence interval for the factor correlation $(c - 1.96SE(c), c + 1.96SE(c))$, we reject the hypothesis of a 1-factor model if the confidence interval does not include ± 1.

Next we consider the general case of the EFA model with m factors. Determining the dimensions of an EFA model can be done with orthogonal or oblique rotation. Regardless of which type of rotation we want to use for the final model, for the purpose of determining the number of factors we can use orthogonal rotation. With such rotation, testing for the significance of the factor is the same as testing that the loadings of that factor are significant. Suppose that the loading matrix in the EFA model is Λ. Testing the significance of the $j-$th column in Λ is the same as testing the existence of the $j-$th factor. We can use the Wald test to simultaneously test all the loadings in the $j-$th column for significance. However, a formal test for m v.s. $m-1$ factors is translated as a hypothesis as follows: one of the m columns in the loading matrix is zero, without specifying which column that is. The null hypothesis of $m-1$ factors is that column 1 = 0 or column 2 = 0 or column 3 = 0 ... or column m =0. We need this hypothesis rejected to establish the existence of the all m factors. This situation is described as multiple hypothesis testing and unfortunately the "or" operand can not be easily handled by the Wald test.

One alternative is to use the weakest column of loadings and test just the one column hypothesis. This method however does not work well. The estimates of the loading parameters are highly correlated typically and separating one loading column from the rest does not yield the needed power and accuracy (simulations not reported here).

Here we describe a different alternative to the Wald test. Testing that at

least one loading in the $j-$th column is not zero is equivalent to testing that

$$W_j = \sum_{i=1}^{P} \lambda_{ij}^2 \tag{1}$$

is significantly different from 0. Here P is the number of indicators in the model. To test that all columns have statistically significant values, we can test the product of the above quantities is significantly different from 0. That is, we compute

$$W = \prod_{j=1}^{m} W_j = \prod_{j=1}^{m} \sum_{i=1}^{P} \lambda_{ij}^2 \tag{2}$$

and evaluate its statistical significance. Statistical significance of $W > 0$ implies that for every j, $W_j > 0$, which implies that for every j there is a statistically significant loading λ_{ij}, which in turn implies that all m factors are statistically significant. The test W can be constructed as a new parameter in Mplus with MODEL CONSTRAINT. This is illustrated in Figure 2 for the EFA model with 3 factors. The null hypothesis of $m-1$ factors is then rejected if the estimates of the m-factor model yield a Z-score $Z = W/SE(W)$ greater than 1.96. This method can then be used for factor enumeration as follows. The number of factors is the largest m for which the model H_m converges to an admissible model and W is significantly different from zero.

Unfortunately, this method has the drawback that it tests a value on the border of admissible space. This is because we are testing $W = 0$ and the admissible space is $W \geq 0$. Because of that, the asymptotic theory which we are using for this test is invalid. Just as the LRT sequential test, using W as is will suffer slightly from overfactoring. However, there is an easily available remedy to this issue. Let's call this method W_+. Here we again compute W as above and $Z = W/SE(W)$ but we consider W to be significant only if Z is greater than 3. This alternative method is guaranteed asymptotically to not under-factor. This is because Z for the correct number of factors converges to infinity when the sample size increases. In addition, the method is guaranteed to not over-factor asymptotically. With the higher threshold, we can take advantage of the fact that the asymptotic confidence interval construction for any internal point in the admissible space is valid. Thus, if $Z > 3$, the estimate of W can be viewed as an internal point, and the 95% confidence interval will be valid and not include zero 95% of the time. Essentially we are replacing the hypothesis $W = 0$ with the hypothesis

$W \leq SE(W)$, which is no longer at the border of admissible space. This method of dealing with borderline tests is also used for testing the variance of random effects. Clearly, because of the higher threshold of significance the power of W_+ is slightly smaller than that of W. This is the price to pay for preventing overfactoring. In that way, the relationship between W and W_+ is the same as that of the sequential LRT method and the test of fit. The cutoff value of 3 is not precise and backed by a theoretic result. The actual distribution of W under the null hypothesis of $W = 0$ for an infinitely large sample is unknown and likely depends on the exact EFA model. However, our experience is that the approach works well in finite sample size situations and that this level of cutoff increase is sufficient to account for the error in the normal distribution approximation.

Note also that in the simple case of 1-factor v.s. 2-factor EFA, when we use the confidence interval for the factor correlation $(c - 1.96SE(c), c + 1.96SE(c))$ to test that the correlation is significantly different from ± 1, we are also testing at the boundary of admissible space. A safer approach is to conclude that 2 factors are needed when the Z-test of the correlation being ± 1 has an absolute value of at least 3.

In our simulation, both W and W_+ perform well. As expected, W leads to a slight over-factoring (8% of the time picked $m = 4$) and W_+ leads to minor loss of power (4% of the time picked $m = 2$).

2.8 Summary

The test of fit, the sequential LRT, BIC, W and W_+ are the five reliable methods. Our simulation study indicates that BIC has somewhat of a lower power than the other four methods and therefore it can also be dismissed in practical settings when the sample size is not large. Simply using the convergence selection method appears to narrow down the selection of models to consider. Typically, only the largest two or three m values with converging H_m models are in contention.

In practical settings, there are more complications than our simulations allow. Correlation uniqueness (residual correlations) between the indicators is a common occurrence. If these are not modeled properly, additional factors are likely to be suggested only to accommodate these residual correlations. If an EFA factor has just two significant EFA loadings, it should be replaced by a residual correlation in the absence of a substantive argument against doing so.

Table 1: Selecting the number of factors in EFA

Method	$m = 1$	$m = 2$	$m = 3$	$m = 4$
Convergence	0	0	65	35
Test of Fit	0	4	96	0
Sequential LRT	0	0	96	4
CFI	0	99	1	0
TLI	0	61	39	0
RMSEA	0	59	41	0
SRMR	0	100	0	0
AIC	0	0	96	4
BIC	0	52	48	0
Kaiser	0	100	0	0
PA	0	100	0	0
PA_{95}	0	100	0	0
W	0	1	91	8
W_+	0	4	96	0

Non-normality in the indicators or cluster sampling should be taken into account by using the MLR estimator instead of the ML estimator. Likelihood based tests should be replaced by the robust likelihood ratio tests. All four approaches: the test of fit, the sequential LRT, W, and W_+, can accommodate robust methods.

Note also that our conclusions are based on this specific simulation study and the general asymptotic theory. In practice, sample size is always moderate or small, asymptotic theory might be irrelevant, and the power of the different methods may differ from what we observed here.

Figure 1: Simulation study for EFA with 3 factors

```
MONTECARLO:
NAMES ARE y1-y10;
NOBSERVATIONS = 500;
NREPS = 100;
results=res3.dat;

analysis:  iter=10000; starts=50;

MODEL POPULATION:
f1 BY y1-y3*1 y10*0.5;
f2 BY y4-y6*1 y10*0.5;
f3 BY y7-y9*1 y10*0.5;
f1-f3@1;
f1 WITH f2*.7;
f1 WITH f3*.5;
f3 WITH f2*.1;
y1-y10*1;

MODEL:
f1 BY y1-y3*1 y4-y9*0 y10*0.5(*1);
f2 BY y1-y3*0 y4-y6*1 y7-y9*0 y10*0.5 (*1);
f3 BY y1-y6*0 y7-y9*1 y10*0.5 (*1);
f1-f3@1;
f1 WITH f2*.7;
f1 WITH f3*.5;
f3 WITH f2*.1;
y1-y10*1;

output:tech9;
```

Figure 2: Computing W for EFA with 3 factors

```
MONTECARLO:
NAMES ARE y1-y10;
NOBSERVATIONS = 500;
NREPS = 100;
results=res3.dat;

analysis: iter=10000; starts=50;
rotation=geomin(orthogonal);

MODEL POPULATION:
f1 BY y1-y3*1 y10*0.5;
f2 BY y4-y6*1 y10*0.5;
f3 BY y7-y9*1 y10*0.5;
f1-f3@1;
f1 WITH f2*.7;
f1 WITH f3*.5;
f3 WITH f2*.1;
y1-y10*1;

MODEL:
f1 BY y1-y3*1 y4-y9*0 y10*0.5(*1);
f2 BY y1-y3*0 y4-y6*1 y7-y9*0 y10*0.5 (*1);
f3 BY y1-y6*0 y7-y9*1 y10*0.5 (*1);
f1-f3@1;
y1-y10*1;
f1 BY y1-y3*1 y4-y9*0 y10*0.5(L1-L10);
f2 BY y1-y3*0 y4-y6*1 y7-y9*0 y10*0.5 (L11-L20);
f3 BY y1-y6*0 y7-y9*1 y10*0.5 (L21-L30);

model constraints:
new(W1-W3);
W1=L1^2+L2^2+L3^2+L4^2+L5^2+L6^2+L7^2+L8^2+L9^2+L10^2;
W2=L11^2+L12^2+L13^2+L14^2+L15^2+L16^2+L17^2+L18^2+L19^2+L20^2;
W3=L21^2+L22^2+L23^2+L24^2+L25^2+L26^2+L27^2+L28^2+L29^2+L30^2;
new(W); W=W1*W2*W3;
```

Table 2: Average BIC

Number of Factors	$m = 1$	$m = 2$	$m = 3$	$m = 4$
N=200	6752	**6634**	6650	6672
N=500	16716	16371	**16370**	16402
N=1000	33292	32558	**32519**	32554

Table 3: Average AIC

Number of Factors	$m = 1$	$m = 2$	$m = 3$	$m = 4$
N=200	6653	6506	6495	**6494**
N=500	16589	16206	**16172**	16175
N=1000	33145	32366	**32289**	**32289**

3 Determining the number of factors in CFA

In this section we consider the problem of determining the number of factors in CFA. All of the methods described for EFA can be used for CFA as well with the exception of the eigenvalues based models. Model nesting with different numbers of factors is not always easily determined. For the purpose of this discussion, we pick the CFA models to be nested and with specific purpose. We use the same data generation (10 indicator 3 factor model) as in the previous section and will again consider models with 1,2,3, and 4 factors. The CFA model with 3 factors is the same as the data generating model. Each of the three factors has 3 pure indicators and the 10-indicator loads on all three factors.

The 4-factor CFA model is an extension of the 3-factor model where the first 3 factors retain the same measurement structure as in the 3-factor model while the fourth factor loads on most indicators. This construction is along the lines of the EFA model with target rotation. First note that the fourth factor can not load on all the indicators because such a model is not an identified model. If we replace the factor covariance of the first 3 factors with one second order factor, the second order factor or any portion of it can play the role of the fourth factor and thus the model with an additional factor

which loads on all indicators is not identified. In target rotation, we select at least $m-1$ targets for each factor. Here, the existing first three CFA factors have such targets (which are essentially loadings fixed to 0). For the fourth factor we fix to zero 3 loadings (for each of the first 3 factors we fix one loading on one of their pure indicators). The remaining seven loadings are estimated as free parameters. Together with the 3 additional factor correlations the added fourth dimension gives 10 more parameters (as many as the number of indicators). This is the same count as an additional dimension in EFA. The construction of the 4-factor model is meant to mimic the situation where we have no prior substantive vision on how an additional dimension can be measured. The construction is also based on the exploratory idea of target rotation and is meant to obtain an identified model which can cover any potential measurement pattern.

As constructed, the 4-factor CFA model preserves the measurement definition of the first 3 factors and aims to have as general as possible measurement structure for the 4-th factor. This is done so no alternative 4-factor measurement structure can yield a better fit and it can be missed in the factor numeration process. In practice, if we have a CFA model and we are uncertain if the number of factors is sufficient, an additional dimension that loads on most indicators is a good exploratory approach which does not require correct measurement factor specification.

The CFA with 2-factors is constructed as a downgrade of the 3-factor model. In this model we combine the two factors with the highest correlation into one. Thus in the 2-factor CFA, one factor has 6 pure indicators, the second factor has 3 pure indicators and the 10-th indicator loads on both factors. This model is nested within the 3-factor CFA and is a competitive 2-factor model given that the true correlation between the first two factors is 0.7. Finally, the 1-factor model is just a factor model where all indicators load on the factor. This model is also nested within the 2-factor CFA. Thus, we have a complete sequence of 4 CFA models nested within each other. The sequence of models also resembles a sequence of models that could be explored with real data. The following are the loading matrices for the 2,3,and 4, factor analysis models where "*" represents the free loading parameters that are

estimated

$$\Lambda_2 = \begin{pmatrix} * & 0 \\ * & 0 \\ * & 0 \\ * & 0 \\ * & 0 \\ * & 0 \\ 0 & * \\ 0 & * \\ 0 & * \\ * & * \end{pmatrix} \tag{3}$$

$$\Lambda_3 = \begin{pmatrix} * & 0 & 0 \\ * & 0 & 0 \\ * & 0 & 0 \\ 0 & * & 0 \\ 0 & * & 0 \\ 0 & * & 0 \\ 0 & 0 & * \\ 0 & 0 & * \\ 0 & 0 & * \\ * & * & * \end{pmatrix} \tag{4}$$

$$\Lambda_4 = \begin{pmatrix} * & 0 & 0 & 0 \\ * & 0 & 0 & * \\ * & 0 & 0 & * \\ 0 & * & 0 & 0 \\ 0 & * & 0 & * \\ 0 & * & 0 & * \\ 0 & 0 & * & 0 \\ 0 & 0 & * & * \\ 0 & 0 & * & * \\ * & * & * & * \end{pmatrix}. \tag{5}$$

The factor enumeration results for the various methods are given in Table 4.

3.1 Convergence

As in the EFA case, only the 4-factor model had convergence problems. Out of 100 replications, 39 did not produce results. In almost all of these cases the condition number is deemed 0 and the model not identified. Additionally, some of the solutions that were obtained were inadmissible due to negative residual variance for the indicators (8 cases) or non-positive definite factor covariance (13 cases). Only 40 of the replications yield a competitive model. Only those replications are used in the subsequent comparison.

3.2 Test of fit

All of the 1-factor and 2-factor models are rejected with the test of fit. In the 3-factor model, only 6 of the replications were rejected and only 2 of these replications lead to an admissible 4-factor model.

3.3 Sequential LRT

The sequential LRT, like in the EFA case, is susceptible to over-factoring. This is again due to the fact that the sequential hypothesis testing involves testing at the border of admissible space and the actual likelihood ratio test statistic does not have an exact chi-square distribution. The LRT test for 1-factor model vs the 2-factor model rejects the 1-factor model in all replications. The LRT test for 2-factor model vs the 3-factor model rejects the 2-factor model also in all replications. The LRT test for 3-factor model vs the 4-factor model rejects the 3-factor model in 11 replications. The average value of this LRT test statistic is 15 and with 10 degrees of freedom this yields an inflated rejection rate of 11%. This result is similar to what is reported for the EFA case in Hayashi et al. (2007). In the CFA case, the reduction due to non-convergence, even though it is similar to the EFA case, did not lead to reduction in the inflated rejection rate.

3.4 Fit indices

The fit indices appear to have sufficient power for this sequence of CFA models. This is likely because the CFA model is not as flexible as the EFA model with fewer avenues to compensate for the difference between the 2-factor and

the 3-factor models. The SRMR appears to still have insufficient power, but CFI, TLI, and RMSEA picked $m = 3$ in almost all of the replications.

3.5 Information criteria

Both AIC and BIC perform well and pick the 3-factor model in almost all replications. Here again, the CFA model rigidity forces better separation between the models which in turn results in better factor enumeration. In addition, in CFA, every parameter contributes to the model fit. Unlike in the EFA framework, here there are no parameters that contribute only to the parameter penalty part of the information criteria. This also explains the improvement in the performance of the information criteria in the CFA settings as compared to the EFA settings.

3.6 Methods based on Wald test or Z-test

Here again we explore the possibility for factor enumeration based on the standard errors of the parameter estimates. Just as in the EFA case, to test a 1-factor CFA model v.s. a 2-factor CFA model, we can simply consider the hypothesis that the factor correlation c between the two factors is ± 1. We can utilize the W confidence interval $(c - 1.96SE(c), c + 1.96SE(c))$ or the W_+ confidence interval $(c - 3SE(c), c + 3SE(c))$ to check that ± 1 is not in the interval, in which case we reject the 1-factor CFA model. The W_+ confidence interval has the advantage that it protects against violations due to testing at the border of admissible space and should be proffered. The W_+ method has the disadvantage that it has lower power than the W method.

Next we consider the more general case of a CFA model with m factors. We want to test if the m-th factor is providing a statistically significant improvement of fit. This can be accomplished as follows. For the m-th factor, we first change the metric so that the variance of the factor is free, i.e., we set the metric of the factor by fixing one of the loadings to 1. Next, instead of estimating a factor covariance matrix for the m factors, we estimate the following equivalent model. Factor covariance is estimated for the first $m-1$ factors, and the m-th factor is regressed on the first $m - 1$ factors, instead of being correlated with those. The statistical significance of the m-th factor is then equivalent to the statistical significance of its residual variance. Note that if the residual variance is not statistically significant, we can then fix it

to 0. Then, the $m-$th factor is a linear combination of the first $m-1$ factors

$$\eta_m = \beta_1\eta_1 + ... + \beta_{m-1}\eta_{m-1} \tag{6}$$

and the factor model becomes

$$Y_p = \nu_p + \lambda_{p1}\eta_1 + ... + \lambda_{m-1,p}\eta_{m-1} + \lambda_{mp}\eta_m = \tag{7}$$

$$\nu_p + (\lambda_{p1} + \beta_1\lambda_{mp})\eta_1 + ... + (\lambda_{m-1,p} + \beta_{m-1}\lambda_{mp})\eta_{m-1}. \tag{8}$$

The above model is an $m-1$ factor model which has a fit that is not significantly worse than the fit of the m factor model, given that the residual variance of the $m-$th factor is not statistically significant. Note here that the measurement model for the $m-1$ factors changes after the $m-$th factor is substituted. Thus, the test here is that there is a need for the $m-$th factor and not necessarily that the measurement model for the first $m-1$ factors is correct.

This procedure can be applied for factor enumeration and we will call this method W. Because here again we are testing a variance parameter at the border of admissible space we can expect some deviation from the asymptotic expectation for a normally distributed Z-score. Using W_+ which requires a Z-score above 3 for the significance of the residual variance parameter can be used to resolve the issue just as we did for the EFA model. In this simulation study, both methods work equally well and identify the correct m in 99 out of 100 replications. Figure 3 shows the input file that can be used to test the significance of the fourth factor in the 4-factor CFA model.

3.7 Conclusions

Almost all of the methods performed well for our simulation study. The only exception is SRMR. Notably, CFI, TLI, RMSEA and BIC performed better in the CFA settings than in the EFA settings. One reason the fit indices perform better in CFA settings is because the distance between CFA models is larger. The more flexible EFA models have more avenues (alternatives to adding a factor) to compensate for deficiencies in the fit. This flexibility is not present in the CFA settings. The improvement in the BIC performance is likely due to the fact that CFA is more parsimonious and does not have penalties associated with parameters that do not improve data fit.

Figure 3: Simulation study for testing the fourth factor in CFA using W

```
MONTECARLO:
NAMES ARE y1-y10;
NOBSERVATIONS = 500;
NREPS = 100;
results=res4.dat;

analysis:  iter=10000; starts=50;

MODEL POPULATION:
f1 BY y1-y3*1 y10*0.5;
f2 BY y4-y6*1 y10*0.5;
f3 BY y7-y9*1 y10*0.5;
f1-f3@1;
f1 WITH f2*.7;
f1 WITH f3*.5;
f3 WITH f2*.1;
y1-y10*1;

MODEL:
f1 BY y1-y3*1 y10*0.5;
f2 BY y4-y6*1 y10*0.5;
f3 BY y7-y9*1 y10*0.5;
f1-f3@1;
f1 WITH f2*.7;
f1 WITH f3*.5;
f3 WITH f2*.1;
y1-y10*1;

f4 by y2@1 y3*1 y5-y6*1 y8-y10*1;
f4*1; f4 on f1-f3;

output:tech9;
```

Table 4: Selecting the number of factors in CFA

Method	$m = 1$	$m = 2$	$m = 3$	$m = 4$
Convergence	0	0	58	42
Test of Fit	0	0	98	2
Sequential LRT	0	0	89	11
CFI	0	1	99	0
TLI	0	0	100	0
RMSEA	0	0	100	0
SRMR	0	97	3	0
AIC	0	0	95	5
BIC	0	0	100	0
W	0	0	99	1
W_+	0	0	99	1

4 Factor enumeration based on robust methods

In this section we illustrate how robust methods can be used for factor enumeration. Such methods are typically used because the data is not normally distributed or because the sampling in the data is complex sampling which may involve weights, cluster sampling or stratification.

We use one data set for this illustration based on the generation method used in the previous sections. To introduce some non-normality in the distribution of the variables, we square the first two indicators. Using estimator=MLR, we estimate the four models but only the 1-factor, 2-factor and 3-factor CFA models converge. It should be noted here that when we square the variables, not only are the variables non-normally distributed, but also the factor analysis model is no longer correct and the data technically is not generated from a 3-factor analysis model, although the deviation from that concerns just two of the variables.

The results of the estimations are given in Table 5. We have only included methods here that are based on the robust MLR methodology. Excluded are the information criteria which are identical to the ML information criteria as well as SRMR which also does not use robustification.

The chi-square test of fit rejects all 3 models and thus we select the 3-factor model as the most well fitting model. The sequential likelihood ratio test requires the following computation of the correction scale

$$c_{m+1,m} = \frac{p_{m+1}c_{m+1} - p_m c_m}{p_{m+1} - p_m}. \tag{9}$$

The robust sequential LRT test statistic is then

$$\frac{2(LL_{m+1} - LL_m)}{c_{m+1,m}} \tag{10}$$

which is compared against the chi-square distribution with $p_{m+1} - p_m$ degrees of freedom. When we compare the 1-factor model v.s. the 2-factor model, we obtain a negative correction factor $c_{2,1} = -0.5862$. This invalidates the test completely. It is possible to use the strictly-positive Satorra and Bentler (2010) version of the robust test. This test is implemented in Mplus and described in Asparouhov and Muthén (2013). The test, however, also fails because the information matrix is not positive definite for the M_{10} model and the alternative correction scale can not be computed. This is not an unusual outcome. The failure tends to happen the most when the models are close in terms of number of parameters but with drastically different fit, which can be seen here by computing the non-robust LRT value: 2000 with 3 degrees of freedom. In those cases, all other alternative methods will yield the conclusion that the model with the smaller number of factors should be rejected.

When the models are quite distant in terms of the number of parameters, such as in the test of fit statistic that compares the factor models with the unrestricted variance covariance model, the robust test works well.

When we compare the 2-factor model v.s. the 3-factor model, we obtain $c_{32} = 1.155$ and a robust test statistics of 127.7 which with 3 degrees of freedom results in a rejection of the 2-factor model. This implies that with this method we also select the 3-factor model.

The three fit indices: CFI, TLI and RMSEA reject the 1-factor and 2-factor models while the 3-factor model is deemed acceptable and thus we conclude that the fit indices also work well. Using the W method based on the robust standard errors, we obtain a Z-score of 18.648 for the 2-factor model vs. the 1-factor model and a Z-score of 4.884 for the 3-factor model v.s. the 2-factor model. Both W and W_+ conclude that the number of factors is 3.

Table 5: Robust LRT

Model	$m = 1$	$m = 2$	$m = 3$
Number of parameters p_m	30	32	35
Scaling factor c_m	1.5610	1.4268	1.4035
Log-likelihood LL_m	-55245.510	-54197.405	-54049.910
Degrees of Freedom	35	33	30
Test of Fit	2780	464	176
Sequential LRT	NA	failed	127.7
CFI	0.595	0.936	0.978
TLI	0.479	0.913	0.968
RMSEA	0.162	0.066	0.040
W method Z-score	NA	18.648	4.884

5 Conclusions

In this note we illustrate a variety of methods that can be used for the purpose of determining the number of factors in CFA and EFA models. Several of the methods work very well for both models: test of fit, sequential LRT, and the W methods. These are also theoretically well justified, although the sequential LRT, and the W method suffer from small over-factoring due to testing at the border of admissible space. Fit indices as well as the information criteria work better for CFA than for EFA. The W methods we discussed here might be considered novel and may need more extensive simulations to determine power and applicability. Overall, we conclude that the arsenal of methods we described is sufficient for factor enumeration purposes in a variety of practical applications.

References

[1] Asparouhov, T. & Muthén, B. (2013) Computing the Strictly Positive Satorra Bentler Chi-Square Test in Mplus. Mplus Web Notes: No 12. https://www.statmodel.com/examples/webnotes/SB5.pdf.

[2] Hayashi, K., Bentler, P. M., & Yuan, K. H. (2007). On the likelihood ratio test for the number of factors in exploratory factor analysis. Structural Equation Modeling, 14, 505-526

[3] 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.

[4] Satorra, A., & Bentler, P.M. (2001). A scaled difference chi-square test statistic for moment structure analysis. Psychometrika, 66, 507-514.

[5] Satorra, A. & Bentler, P.M. (2010). Ensuring positiveness of the scaled difference chi-square test statistic. Psychometrika, 75, 243-248.

[6] Schmitt, T. A. (2011). Current methodological considerations in exploratory and confirmatory factor analysis. Journal of Psychoeducational Assessment, 29, 304-321.

[7] Timmerman, M. E., & Lorenzo-Seva, U. (2011). Dimensionality assessment of ordered polytomous items with parallel analysis. Psychological Methods, 16, 209-220.