Explorative Factor Analysis

Basics

Goals of EFA

Researchers in the social sciences use questionnaires to measure psychological and social constructs. The goal of an explorative factor analysis (EFA) is to study the **internal structure** of a set of questionnaire items. More specifically, using EFA you want to investigate whether a limited number of factors can explain the associations between the items. These factors represent hypothetical constructs, also referred as latent variables. The factors describe characteristics on which people differ, but which cannot be directly observed. Examples include personality factors (the big Five), attitudes, aptitude, preferences, values, etc.

From a practical point of view, EFA amounts to answering two main substantive questions:

- De the items measure one single factor or multiple factors?
- How does the factor structure look like? Which items load on which factors?

Exploratory versus Confirmatory Factor Analysis

A distinction is made between **confirmative** and **explorative** factor analysis. Researchers use confirmative factor analysis to test a-priori hypotheses about the internal structure of the questionnaire. Hypotheses may be theoretically based or based on extant factor analyses in other studies or populations. Hence, the starting point are articulated hypotheses about the number of factors, and which items load on which factors. In exploratory factor analysis, the researcher starts without any a-priori hypotheses and just sees what comes out of the analysis. Hence, in EFA the researcher lets the data speak.

Principal Component Analysis (PCA) versus Factor Analysis (FA)

In the literature, and research papers, FA and PCA are often confused. PCA is a data reduction tool. The goal of a PCA is to find patterns in the data and to see if we can summarize the data by means of components (= weighted sums of the items) without losing information. For example, if we have 10 items that correlate highly, it means that the items describe similar differences, and we may use the summed score instead of the single item scores in the analyses without losing too much information. In general, one can say that the goal of PCA is to explain the item variances with a few components. PCA is not so much concerned in explaining the correlations between the items.

The situation in EFA is different. EFA is used to see if we can find underlying factors that may explain the *associations between* the items. Thus, EFA focuses on what the items *have in common*.

Requirements for Factor Analysis

Prior to an EFA, one should evaluate whether EFA is meaningful at all. The goal of EFA is to explain the associations between the items. To accomplish this goal EFA uses the correlation matrix. There are a few issues involved:

The first issue is the linearity assumption. The correlation matrix describes the linear association between item pairs. Hence, EFA assumes that the relationships between items is linear. Moreover, it assumes that the scores are bivariate normally distributed, which means that the scatter plots of the scores on item pairs should be ellipsoids (cigars). Fortunately, EFA is quite robust against violations. This means that we can trust the results even though the assumptions linearity and normality may be somewhat violated.

A second issue concerns the question whether EFA is meaningful at all. To find factors, the items must have enough in common to be factorized. For example, if all items only correlate .10 or less, then the items have almost nothing in common, and it doesn't make sense to look for underlying common factors. Therefore, as a first step one may evaluate the **factorability** based on the inter-item correlation matrix. The assumption of factorability is usually evaluated the KMO index. If the KMO is large enough, we may be sure that items have enough in common to extract factors (more about this issue later).

The third issue is the sample size. To get stable results from the EFA, one needs samples that are large enough. It is hard to come up with precise guidelines but in general one need at least 150 to 200 persons, but 350 to 500 would be a safer choice.

The Common Factors Model

The common factors model assumes that the scores on the items are the (weighted) sum of the underlying factors and a residual part. For example, if we have six items, and we assume two factors, then for each item we describe the item score as a weighted sum of the factors and an error part. That is, we have a two-factor model given by…

$$\begin{aligned} X_1 &= a_{11}F_1 + a_{12}F_2 + e \\ X_2 &= a_{21}F_1 + a_{22}F_2 + e \\ X_3 &= a_{31}F_1 + a_{32}F_2 + e \\ X_4 &= a_{41}F_1 + a_{42}F_2 + e \\ X_5 &= a_{51}F_1 + a_{52}F_2 + e \\ X_6 &= a_{61}F_1 + a_{62}F_2 + e \end{aligned}$$

where, F_1 and F_2 represent the common factors. You may conceive the factors as latent variables describing differences in latent constructs. For example, F_1 may represent extraversion. Some persons score high on this factor, other persons score low. The weights a_{jk} (j is the item index, k is the factor index; e.g., a_{32}, reads as, the factor loading of item 3 on factor 2) how much each factor loads on the respective items. You can conceive the loadings as regression coefficients. They describe differences in the item score for a one unit difference in the factor. The higher the loadings, the higher the effect of F_1 on the item scores. Thus, if the factor loading for an item is high on a particular factor, we have an item that is a strong indicator of that factor. Finally, we have the residuals (e), which describe the variance in the item that cannot be explained by the factors. In other words, the residuals describe the **unicity** of the items. The complement is the **communality**. Hence, the communality is that part of the item variance that can be explained by the common factors.

In EFA, we use empirical data to estimate the factor model. This means that we estimate the loadings (e.g., using *principal axis factoring*; to be discussed). Based on the pattern of high and low loadings we may understand how the factors explain the item scores. Using the estimated model we can also evaluate how well the factor model explains the structure in the data including item variances and inter-item correlations.

Goodness-of-Fit: The residual Correlations

The goal of EFA is to explain the relationship between the items. An important question is: how well does the factor model explain the correlations? To evaluate this question, we may compare the expected correlations between the under the postulated factor model with the observed correlations. If the differences are small, we have a model that fits the data well. The expected correlations are usually referred to as the **reproduced correlations** or model implied correlations. The difference between the observed correlations and reduced correlations are **residual correlations**. You may use the following guidelines for evaluating model fit: absolute residuals < 0.05 => *good fit*; most absolute residuals < 0.05, the others < 0.10 => *acceptable fit*. Notice that "absolute residuals" means that we ignore the signs; thus, -0.04 and 0.03 represent good fit.

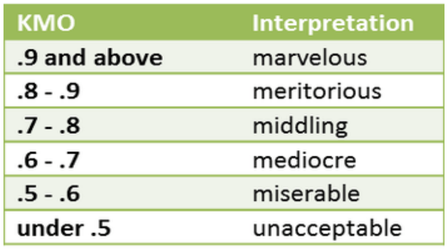
Practical EFA

A typical EFA takes four steps:

1. Evaluate Whether EFA is Meaningful

To evaluate whether EFA is meaningful you may first inspect the correlation matrix to see how well the items correlate. It is hard to draw any firm conclusions, but it may give you a first impression of the strength and direction of the associations and a possible structure. For example, you may already see that some items cluster together because they correlate high with each other, but not with other items.

Second, you may use the KMO index to evaluate the factorability. High factorability means that correlations between (subsets of) items are high enough to extract meaningful factors. The KMO has the following interpretation:



KMO	Interpretation
.9 and above	marvelous
.8 - .9	meritorious
.7 - .8	middling
.6 - .7	mediocre
.5 - .6	miserable
under .5	unacceptable

Hence, the higher the KMO value the better, with a bare medium of .5.

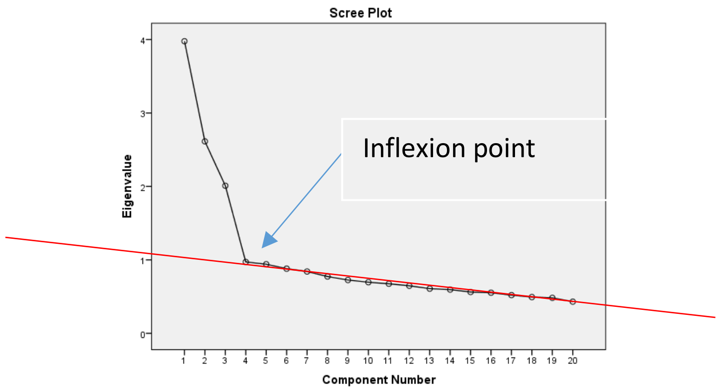
2. Decide about the Number of Factors to Extract

There are different procedures to decide about the number of factors, but the most popular procedures is the propose by Cattell's, which uses the scree plot.

Scree plot. Below you find an example of a scree plot. The scree plot is a graphical display of a component analysis (PCA). On the *x*-axis we see the components. Components are weighted sums of the item scores. There can be as many components as there are items. On the *y*-axis are the eigenvalues. The eigenvalue is the variance of the component scores (= weighted item sums). The basic idea behind the plot is the following. The scree plot assumes standardized items (i.e., the variance of each item is 1). Now, suppose we have seven items, the sum of the variances of the standardized items sum up to 7. The first eigenvalue is the variance of the linear combination of the items (i.e., weighed sum) that describes as much as possible of the items variances. For example, if the first eigenvalue is 3.5, it means that the first component accounts for 50% of the items variances. The second eigenvalue is the linear combination that describes as much as possible from the remaining variances, and so forth.

To fully understand what eigenvalues are, we need matrix algebra, which is beyond the scope of this overview. What is important for applied EFA researchers is the *pattern* of eigenvalues. The eigenvalues are by definition ordered, such that the eigenvalue of the first component is always the highest, followed by the next one. The scree plot shows the eigenvalues against components and is *always* a decreasing function. The suggested number of factors is the number of dots above the "inflexion point".

Example of scree plot



In the example, we would extract three factors.

Important:

- To extract the factors, you need at least three items per factor. Thus the number of factors cannot be larger than #items/3.
- It may not always be immediately clear from the scree plot how many factors you should extract. Sometimes the scree plot may have two inflexion points. Usually this means that you have a few dominant factors (i.e., the number of factors above the first inflexion points), and a fewer smaller

factors, but which may not very meaningful. In such cases, you may evaluate the factor structure for different number of factors and then choose as the final solution the one that is best interpretable from a substantive point of view.

3. Estimate the Loadings given the Number of Hypothesized Factors

Once a decision is made about the number of factors to retain, we have to estimate the final factor model given the postulated number of factors in such a way that it gives interpretable results. In practice, this amounts to running SPSS again, but now we specify the method for factor extraction (i.e., the procedure by which SPSS estimates the factor loadings), the number of factors we want to extract, and whether or not we allow a non-zero correlation between the factors. The latter decision amounts to choosing the appropriate **rotation method**.

Method of Factor Extraction. There are different methods to find the loadings of the factor solution. The methods differ in how they tackle the problem statistically. The exact details are beyond the scope of this summary, so we stick to the method that is recommended in practice, which is *principal axis factoring*.

Rotation. To ensure that we have an interpretable solution, we use so-called "rotation". By using rotation we may see more easily if the items satisfy **a simple structure.** As explained above, a simple structure means that for each factor we have a few items that load high on that factor and low on the other factors. If the items satisfy a simple structure, it means that we have identifiable factors.

A distinction is made between **orthogonal** and **non-orthogonal** rotation. If we use **orthogonal rotation** we look for an interpretable solution assuming that the factor are independent (i.e., uncorrelated). A popular method for orthogonal rotation is *Varimax* rotation. However, the assumption of independent factors is a rather restrictive assumption. Fortunately, there is an alternative called **non-orthogonal rotation**, also known as **oblique rotations**, which allows a non-zero correlation between the factors. In general, you are advised to use non-orthogonal rotation unless you have strong reasons to be believe that the factors are indeed independent. Also for non-orthogonal rotation methods there are different procedures, of which the *Direct Oblimin* is most popular.

Simple Structure. In practice, researchers strive for factor structures that satisfy a **simple structure**. A simple structure is satisfied if for each factor there are a few items (say three or more) that substantially load only on that factor and not (or only very weakly) on the other factors. If this characteristic is satisfied we can define distinguishable factors. As a general rule, you may consider (absolute) loadings of .3 or more as “substantial”. Also notice that if we look at the size of the loadings, we ignore the plus or minus sign. For example, loadings of -.4 and .4 or both considered “substantial”. However, if you use the factor model to divide the questionnaire items into subscales, then you should only cluster items for which the loadings have the same sign, otherwise you have subscales that have both items that are indicative and contra-indicative for the construct.

4. Evaluate the Final Factor Solution

Finally, we need to evaluate the final factor model. First, we may evaluate the fit. Good fit means that the correlations predicted by the model are close to what we actually see in the sample. If the common factors cannot explain the correlations well, we may question the validity of the model. To accomplish this goal we look at the residual correlations. In general, we consider the fit of the model as good if the residual correlations are less than .05, and acceptable if most of the residual correlations are less .05 and the others less than .10.

Second we may see how much of the item variances are explained by the factor, and what part is unique. This information is given by the **communalities**. Because EFA uses standardized scores, the communalities can be conceived as the variance in the item score that can be explained from the factors. For example, if the communality for an item is .53, it means that the factors explain 53% of the item variance. The remaining part (47%) is the variance in the item that is unique. This part is called **unicity**. Notice that low communalities does not mean that the factor model is invalid, it just shows that the underlying factors have a small effect on the items.

Third, we can evaluate the **factor loadings** (which are shown in the **pattern matrix** if non-orthogonal rotation is used). The factor loadings represent the direct effects of the factors on the items. If the value is 0, it means that the factor does not explain the item. If the loading is close to 0, it means that factor has a small effect. The larger the loading, the larger the effect of the factor. In practice, usually we want the loadings to be at least .3. So, when inspecting the matrix, we see on which factors the item load more than .3 and see if we find clusters of items that pertain to a particular factor.

SPSS also produces a **structure matrix**, which shows for each item the correlation with the factors. The structure matrix may also be consulted to uncover the structure, particularly if the correlation between the factors is not too high, but in general it is better to use the pattern matrix to draw conclusions about the structure.

Finally, you may look at the factor correlations. If the correlation is low, it means that we have clearly distinguishable factors. As the correlation becomes larger, the factors more and more describe overlapping variance, and cannot be well distinguished.

A worked out example

For the example, we will use data on six items on coping with malodor. The questions were:

Keep Windows Closed
Seek diversion
Try to adapt to situation
Think of something else
File complaint with producer
Try to find a solution
Call housing agency

Step 1: Factorability?

Correlation Matrix

		Keep Windows Closed	Seek diversion	Try to adapt to situation	Think of something else	File complaint with producer	Try to find a solution	Call housing agency
Correlation	Keep Windows Closed	1.000	.403	.251	.306	.094	.072	.079
	Seek diversion	.403	1.000	.375	.381	.134	.094	.063
	Try to adapt to situation	.251	.375	1.000	.358	.101	.038	.044
	Think of something else	.306	.381	.358	1.000	.053	.044	.036
	File complaint with producer	.094	.134	.101	.053	1.000	.258	.334
	Try to find a solution	.072	.094	.038	.044	.258	1.000	.167
	Call housing agency	.079	.063	.044	.036	.334	.167	1.000

Correlations range from .036 to .403. We see clusters of items that have higher correlations among each other than with other items. For example, the first four items form such a cluster, and the last three items.

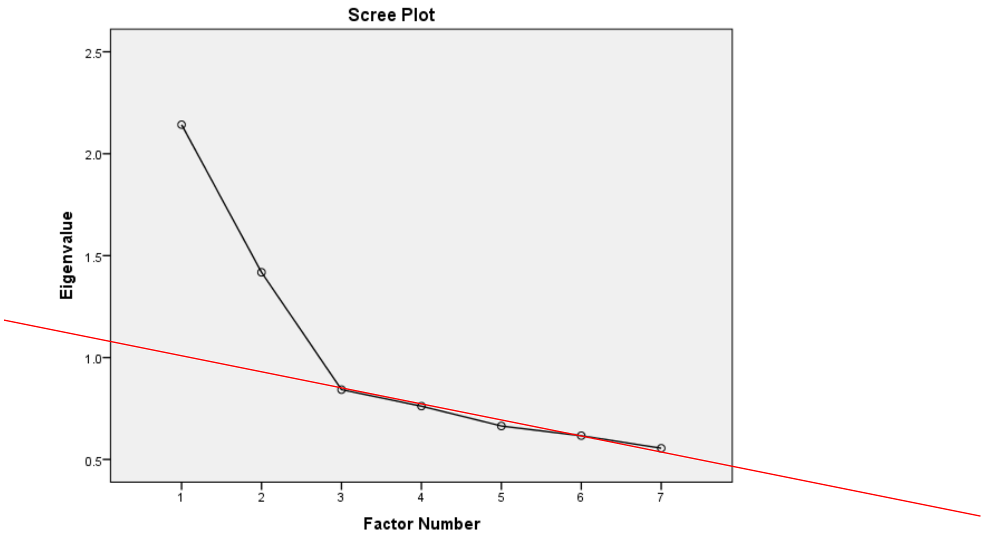
S

KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.702
Bartlett's Test of Sphericity	Approx. Chi-Square	648.337
	Df	21
	Sig.	.000

The KMO index suggests that factorability is "middling", so it's not perfect but good enough. That the factorability is not that impressive is not a surprise given that the inter-item correlation were in general low (highest correlation was .403), and if the relations are weak it becomes harder to find factors.

Step 2: How many factors to choose?



The scree plot suggest two factors. Notice that with 7 items we cannot extract more than two factors anyway.

Step 3: How does the solution look like?

Communalities

	Initial	Extraction
Keep Windows Closed	.198	.282
Seek diversion	.287	.484
Try to adapt to situation	.202	.297
Think of something else	.219	.349
File complaint with producer	.166	.503
Try to find a solution	.078	.132
Call housing agency	.121	.220

Extraction Method: Principal Axis Factoring.

The communalities under extraction (last column) describe how much of the item variance is explained by the factors. Because the items are standardized we may interpret the values as proportions. For example, the two-factor model explain 28.2% of the variance in Keep Windows closed. The communalities are not very impressive. It means that the variances in the items are for a large part unique (i.e., specific to the item). The only exception is try to find a solution. This makes sense because try to find a solution may involve any type of coping.

Reproduced Correlations

		Keep Windows Closed	Seek diversion	Try to adapt to situation	Think of something else	File complaint with producer	Try to find a solution	Call housing agency
Reproduced Correlation	Keep Windows Closed	.282[a]	.370	.289	.310	.108	.067	.064
	Seek diversion	.370	.484[a]	.379	.407	.133	.084	.078
	Try to adapt to situation	.289	.379	.297[a]	.321	.084	.056	.048
	Think of something else	.310	.407	.321	.349[a]	.057	.043	.029
	File complaint with producer	.108	.133	.084	.057	.503[a]	.257	.333
	Try to find a solution	.067	.084	.056	.043	.257	.132[a]	.170
	Call housing agency	.064	.078	.048	.029	.333	.170	.220[a]
Residual[b]	Keep Windows Closed		.033	-.037	-.004	-.014	.004	.015
	Seek diversion	.033		-.004	-.026	.001	.010	-.014
	Try to adapt to situation	-.037	-.004		.037	.016	-.018	-.004
	Think of something else	-.004	-.026	.037		-.004	.001	.007
	File complaint with producer	-.014	.001	.016	-.004		.001	.001
	Try to find a solution	.004	.010	-.018	.001	.001		-.002
	Call housing agency	.015	-.014	-.004	.007	.001	-.002	

Extraction Method: Principal Axis Factoring.
a. Reproduced communalities
b. Residuals are computed between observed and reproduced correlations. There are 0 (0.0%) nonredundant residuals with absolute values greater than 0.05.

Here, we see that none of the residuals is larger than 0.05. This means that the factor model fits well!

Pattern Matrix[a]

	Factor	
	1	2
Keep Windows Closed	.522	.036
Seek diversion	.688	.033
Try to adapt to situation	.546	-.004
Think of something else	.601	-.055
File complaint with producer	-.006	.711
Try to find a solution	.021	.358
Call housing agency	-.019	.473

Extraction Method: Principal Axis Factoring.
Rotation Method: Oblimin with Kaiser Normalization.
a. Rotation converged in 3 iterations.

The pattern matrix shows the loadings; that is, the direct effect of each factor on the item score. The loadings that are larger than .3 are highlighted. We see that the items tend to load on one factors. This means that first four items form a factor, and the last three items form a factor. Because all items load on one dimension only the criteria of a simple structure are satisfied.

Structure Matrix

	Factor	
	1	2
Keep Windows Closed	.530	.157
Seek diversion	.695	.193
Try to adapt to situation	.545	.123
Think of something else	.588	.085
File complaint with producer	.159	.709
Try to find a solution	.105	.363
Call housing agency	.091	.469

Extraction Method: Principal Axis Factoring.
Rotation Method: Oblimin with Kaiser Normalization.

The structure matrix shows the correlations of the items with the factors. The structure matrix may also be consulted to uncover the structure if the correlation between the factors is not too high, but most researchers prefer the pattern matrix to draw conclusions about the structure.

Factor Correlation Matrix

Factor	1	2
1	1.000	.233
2	.233	1.000

Extraction Method: Principal Axis Factoring.
Rotation Method: Oblimin with Kaiser Normalization.

The correlation between the factors is .233. Thus, there is a weak correlation.
