Factor Analysis: A Short Introduction, Part 1 - The Analysis Factor

by Maike Rahn, PhD

Why use factor analysis?

Factor analysis is a useful tool for investigating variable relationships for complex concepts such as socioeconomic status, dietary patterns, or psychological scales.

It allows researchers to investigate concepts that are not easily measured directly by collapsing a large number of variables into a few interpretable underlying factors.

What is a factor?

The key concept of factor analysis is that multiple observed variables have similar patterns of responses because they are all associated with a latent (i.e. not directly measured) variable.

For example, people may respond similarly to questions about income, education, and occupation, which are all associated with the latent variable socioeconomic status.

In every factor analysis, there are the same number of factors as there are variables. Each factor captures a certain amount of the overall variance in the observed variables, and the factors are always listed in order of how much variation they explain.

The eigenvalue is a measure of how much of the variance of the observed variables a factor explains. Any factor with an eigenvalue ≥1 explains more variance than a single observed variable.

So if the factor for socioeconomic status had an eigenvalue of 2.3 it would explain as much variance as 2.3 of the three variables. This factor, which captures most of the variance in those three variables, could then be used in other analyses.

The factors that explain the least amount of variance are generally discarded. Deciding how many factors are useful to retain will be the subject of another post.

What are factor loadings?

The relationship of each variable to the underlying factor is expressed by the so-called factor loading. Here is an example of the output of a simple factor analysis looking at indicators of wealth, with just six variables and two resulting factors.

Variables	**Factor 1**	**Factor 2**
Income	0.65	0.11
Education	0.59	0.25

Occupation	0.48	0.19
House value	0.38	0.60
Number of public parks in neighborhood	0.13	0.57
Number of violent crimes per year in neighborhood	0.23	0.55

The variable with the strongest association to the underlying latent variable. Factor 1, is income, with a factor loading of 0.65.

Since factor loadings can be interpreted like standardized regression coefficients, one could also say that the variable income has a correlation of 0.65 with Factor 1. This would be considered a strong association for a factor analysis in most research fields.

Two other variables, education and occupation, are also associated with Factor 1. Based on the variables loading highly onto Factor 1, we could call it “Individual socioeconomic status.”

House value, number of public parks, and number of violent crimes per year, however, have high factor loadings on the other factor, Factor 2. They seem to indicate the overall wealth within the neighborhood, so we may want to call Factor 2 “Neighborhood socioeconomic status.”

Notice that the variable house value also is marginally important in Factor 1 (loading = 0.38). This makes sense, since the value of a person’s house should be associated with his or her income.

About the Author: *Maike Rahn is a health scientist with a strong background in data analysis. Maike has a Ph.D. in Nutrition from Cornell University.*

Factor Analysis: A Short Introduction, Part 2-Rotations - The Analysis Factor

by Maike Rahn, PhD

Rotations

An important feature of factor analysis is that the axes of the factors can be rotated within the multidimensional variable space. What does that mean?

Here is, in simple terms, what a factor analysis program does while determining the best fit between the variables and the latent factors: Imagine you have 10 variables that go into a factor analysis.

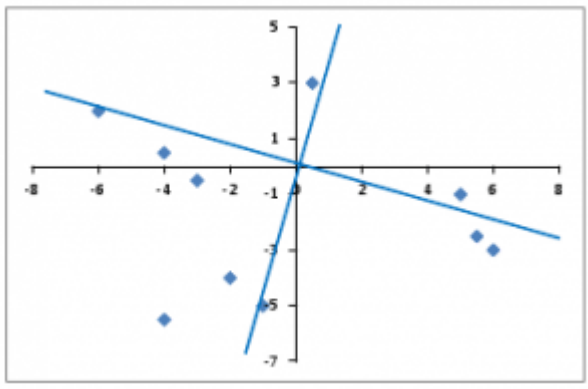
The program looks first for the strongest correlations between variables and the latent factor, and makes that Factor 1. Visually, one can think of it as an axis (Axis 1).

The factor analysis program then looks for the second set of correlations and calls it Factor 2, and so on.

Sometimes, the initial solution results in strong correlations of a variable with several factors or in a variable that has no strong correlations with any of the factors.

In order to make the location of the axes fit the actual data points better, the program can rotate the axes. Ideally, the rotation will make the factors more easily interpretable.

Here is a visual of what happens during a rotation when you only have two dimensions (x- and y-axis):



The original x- and y-axes are in black. During the rotation, the axes move to a position that encompasses the actual data points better overall.

Programs offer many different types of rotations. An important difference between them is that they can create factors that are correlated or uncorrelated with each other.

Rotations that allow for correlation are called **oblique rotations**; rotations that assume the factors are not correlated are called **orthogonal rotations**. Our graph shows an orthogonal rotation.

Once again, let's explore indicators of wealth.

Let's imagine the orthogonal rotation did not work out as well as previously shown. Instead, we get this result:

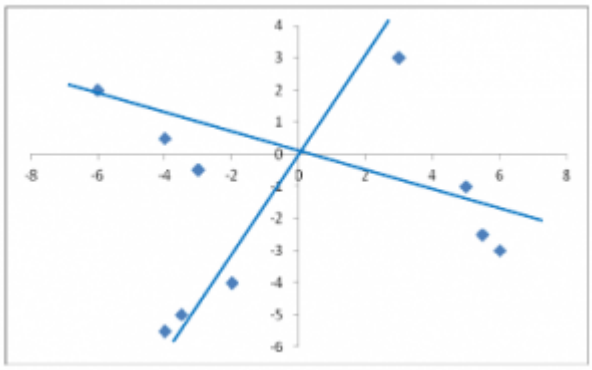
Variables	**Factor 1**	**Factor 2**
Income	0.63	0.14
Education	0.47	0.24
Occupation	0.45	0.22
House value	0.39	0.25
Number of public parks in neighborhood	0.12	0.20
Number of violent crimes per year	0.21	0.18

Clearly, no variable is loading highly onto Factor 2. What happened?

Since our first attempt was an orthogonal rotation, we specified that Factor 1 and 2 are not correlated.

But it makes sense to assume that a person with a high "Individual socioeconomic status" (Factor 1) lives also in an area that has a high "Neighborhood socioeconomic status" (Factor 2). That means the factors *should* be correlated.

Consequently, the two axes of the two factors are probably closer together than an orthogonal rotation can make them. Here is a display of the oblique rotation of the axes for our new example, in which the factors are correlated with each other:



Clearly, the angle between the two factors is now smaller than 90 degrees, meaning the factors are now correlated. In this example, an oblique rotation accommodates the data better than an orthogonal rotation.

Factor Analysis: A Short Introduction, Part 3-The Difference Between Confirmatory and Exploratory Factor Analysis

by Maike Rahn, PhD

An important question that the consultants at The Analysis Factor are frequently asked is:

What is the difference between a confirmatory and an exploratory factor analysis?

A **confirmatory factor analysis** assumes that you enter the factor analysis with a firm idea about the number of factors you will encounter, and about which variables will most likely load onto each factor.

Your expectations are usually based on published findings of a factor analysis.

An example is a fatigue scale that has previously been validated. You would like to make sure that the variables in your sample load onto the factors the same way they did in the original research.

In other words, you have very clear expectations about what you will find in your own sample. This means that you know the number of factors that you will encounter and which variables will load onto the factors.

The criteria for variable inclusion are much more stringent in a confirmatory factor analysis than in an exploratory factor analysis. A rule of thumb is that variables that have factor loadings <|0.7| are dropped.

If you would like to include hypothesis testing such as goodness-of-fit tests in your confirmatory factor analysis, you also may want to consider running it in structural equation modeling software, like AMOS, MPlus or LISREL.

An **exploratory factor analysis** aims at exploring the relationships among the variables and does not have an a priori fixed number of factors. You may have a general idea about what you think you will find, but you have not yet settled on a specific hypothesis.

Or you may have formulated a research question based on your theoretical understanding, and are now testing it.

Of course, in an exploratory factor analysis, the final number of factors is determined by your data and your interpretation of the factors. Cut-offs of factor loadings can be much lower for exploratory factor analyses.

When you are developing scales, you can use an exploratory factor analysis to test a new scale, and then move on to confirmatory factor analysis to validate the factor structure in a new sample.

For example, a depression scale with the underlying concepts of depressed mood, fatigue and exhaustion, and social dysfunction can first be developed with a sample of rural US women using an exploratory factor analysis.

If you would like to next use that scale in a sample of urban US women, you would use a confirmatory factor analysis to validate the depression scale in your new sample.

Factor Analysis: A Short Introduction, Part 4-How many factors should I find? - The Analysis Factor

Factor Analysis: A Short Introduction, Part 4–How many factors should I find?

by Maike Rahn, PhD

One of the hardest things to determine when conducting a factor analysis is how many factors to settle on. Statistical programs provide a number of criteria to help with the selection.

Eigenvalue > 1

Programs usually have a default cut-off for the number of generated factors, such as all factors with an eigenvalue of ≥1.

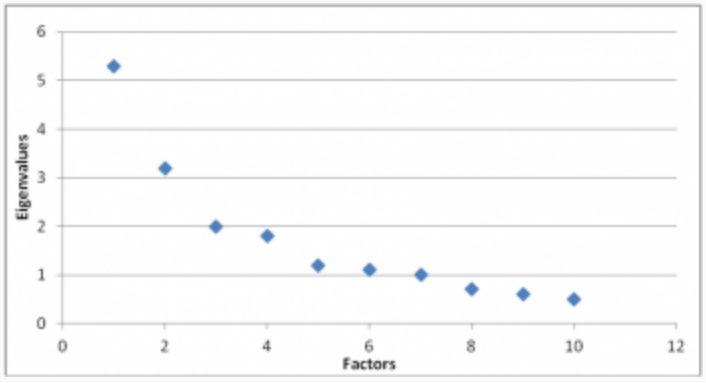
This is because a factor with an eigenvalue of 1 accounts for as much variance as a single variable, and the logic is that only factors that explain at least the same amount of variance as a single variable is worth keeping.

But often a cut-off of 1 results in more factors than the user bargained for or leaving out a theoretically important factor whose eigenvalue is just below 1. So use this criterion only with extreme caution.

Scree Plot

Another option is the scree plot. A scree plot shows the eigenvalues on the y-axis and the number of factors on the x-axis. It always displays a downward curve.

The point where the slope of the curve is clearly leveling off (the "elbow) indicates the number of factors that should be generated by the analysis.



Unfortunately, both criteria sometimes yield an unreasonably high number of factors. In the above example, a cut-off of an eigenvalue ≥1 would give you seven factors. And the scree plot suggests either three or five factors due to the way the slope levels off twice.

It is important to keep in mind that one of the reasons for running a factor analysis is to *reduce* the large number of variables that describe a complex concept such as socioeconomic status to a few interpretable latent variables (=factor). In other words, we would like to find a smaller number of interpretable factors that explain the maximum amount variability in the data.

Total Percent Variance Explained

Therefore, another important metric to keep in mind is the total amount of variability of the original variables explained by each factor solution.

Remember that every factor analysis has the same number of factors as it does variables, and those factors are listed in the order of the variance they explain. You'll always be able to explore more total variance by keeping more factors in the solution, but later factors explain so little variation, they don't add much.

If the first three factors together explain most of the variability in the original 10 variables, then those factors are clearly a good, simpler substitute for all 10 variables. You can drop the rest without losing much of the original variability.

But if it takes 7 factors to explain most of the variance in those 10 variables, you might as well just use the original 10.

Meaningful Factors

It is also important that the rotated factors make theoretical sense to the researcher.

Do the variables that are loading on the same factor make sense together? If you can name the concept they represent, that's indicative that the factor solution is a reasonable one.

Likewise, do the variables that are loading on different factors measure something different? If you've created a scale with two items that are just different wordings of the same underlying question, a factor solution that puts them on different factors doesn't make a lot of sense.

Keep in mind that each of the identified factors should have *at least* three variables with high factor loadings, and that each variable should load highly on only one factor.

After looking at the scree plot as a guide, I often wind up forcing my analysis to run between one and five factors, and then develop the five models separately.

Usually it quickly becomes clear when to drop a factor solution, especially when one factor has only two important variables and therefore does not explain much of the overall variability, or if it is not very convincing based on my theoretical expectations.

Factor Analysis: A Short Introduction, Part 5-Dropping unimportant variables from your analysis - The Analysis Factor

by Maike Rahn, PhD

When are factor loadings not strong enough?

Once you run a factor analysis and think you have some usable results, it's time to eliminate variables that are not "strong" enough. They are usually the ones with low factor loadings, although additional criteria should be considered before taking out a variable.

As a rule of thumb, your variable should have a rotated factor loading of at least |0.4| (meaning ≥ +.4 or ≤ −.4) onto one of the factors in order to be considered important.

Some researchers use much more stringent criteria such as a cut-off of |0.7|. In some instances, this may not be realistic: for example, when the highest loading a researcher finds in her analysis is |0.5|.

Other researchers relax the criteria to the point where they include variables with factor loadings of |0.2|. Which cut-offs to use depends on whether you are running a confirmatory or exploratory factor analysis, and on what is usually considered an acceptable cut-off in your field. In addition, a variable should ideally only load cleanly onto one factor.

How many variables and observations?

Another question often asked is how many variables a researcher should use for analysis. Generally, each factor should have at least three variables with high loadings.

It is also important to have a sufficient number of observations to support your factor analysis: per variable you should ideally have about 20 observations in the data set to ensure stable results. A common minimum is the lesser of 10 observations per variable and 100 observations. However, some statisticians would go as low as five observations per variable .

Factor Analysis: A Short Introduction, Part 6-Common Problems - The Analysis Factor

by Maike Rahn, PhD

In the previous blogs I wrote about the basics of running a factor analysis. Real-life factor analysis can become complicated. Here are some of the more common problems researchers encounter and some possible solutions:

- **The factor loadings in your confirmatory factor analysis are only |0.5| or less.**

Solution: lower the cut-offs of your factor loadings, provided that lower factor loadings are expected and accepted in your field.

- **Your confirmatory factor analysis does not show the hypothesized number of factors.**

Solution 1: you were not able to validate the factor structure in your sample; your analysis with this sample did not work out.

Solution 2: your factor analysis has just become exploratory. Something is going on with your sample that is different from the samples used in other studies. Find out what it is.

- **A few key variables in your confirmatory factor analysis do not behave as expected and/or are correlated with the wrong factor.**

Solution: the good news is that you found the hypothesized factors. The bad news is that something is different about your sample compared with previous analyses. Find out what it is. You may be able to add valuable information to your field.

- **Your program indicates eight factors, but you think you really may have fewer factors in your data.**

Solution: force your factor analysis to show you other solutions, say from two to seven factors. See whether having fewer factors improves the interpretability of your results. Pursue the solution that gives you the most conclusive results based on theory.

- **You have a factor containing only variables with factor loadings of |0.3| or less.**

Solution: you may have too many factors. Force the program to reduce the number of factors and rerun the analysis. When you understand your data better, drop the variables with factor loadings below your cut-off.

- **You have a factor with only two variables.**

Solution: you may have too many factors. Force your program to reduce the number of factors and check whether your variables get incorporated into another factor. On the other hand, if they still stay with their previous factor, this factor may be very stable, and you may want to keep it separate. Of course, your results need to remain interpretable.

Reading material:

Finally, I would like to suggest some reading material. I like the factor analysis booklets from Sage's Quantitative Applications in the Social Sciences. They are extensive and detailed, yet allow a novice to start at a beginner's level and work her way up.

Kim, Jae-On and Mueller, Charles W (1978) Introduction to factor analysis. Series: Quantitative Applications in the Social Sciences. Sage Publications: Beverly Hills, CA.

Kim, Jae-On and Mueller, Charles W (1978) Factor analysis. Statistical methods and practical issues. Series: Quantitative Applications in the Social Sciences. Sage Publications: Beverly Hills, CA.