PCA and EFA

**Description:** Both PCA and EFA are descriptive analyses used to understand the underlying pattern in the data. They both group variables together based on the high correlations between patterns of answers on those variables. They are used for data reduction.

* PCA: Principle Components Analysis
  + Components are combinations of correlated variables, and the variables are thought to *cause* the components.1
  + Components are produced, which is the term to use when writing up.
  + All the variance in the variables in analyzed.
* EFA: Exploratory Factor Analysis – describes the data and summarizes “factors”, often used as a first step on a scale or data.
  + Factors are thought to *cause* variables, the underlying construct is what creates the scores on each variable.
  + Factors are produced, which is the term to use when writing up.
  + Only shared variables and unique variance is analyzed.

**Definitions/Abbreviations:**

* Observed correlation matrix – the correlations between all of the variables (very similar to doing a bivariate correlation chart).
* Reproduced correlation matrix – correlation matrix created from the factors created.
* Residual correlation matrix – the difference between original and reduced correlation matrix. This matrix will be very small if you had a good fit for your model.
* Factor rotation – process by which the solution is made “better” (smaller residuals) without changing the mathematical properties.
  + Oblique – Oblimin is the most common rotation. Factors are allowed to be correlation when they are rotated.
  + Orthogonal – Varimax is the most common rotation. Orthogonal rotation holds factors completely uncorrelated.
* Loading matrix – correlations between the variables and the factors.
* Factor correlation matrix – correlations between the factors. If these are correlated at all, you should use a oblique rotation.
* Structure Matrix – correlations between the factors and the variables.
* Pattern matrix – unique correlation between each factor and variables. This matrix is the information you will report and interpret. These correlations are similar to pr in regression.
* Eigenvalues – eigenvalues are combinations of variance – usually people look at how many values are over 1 to indicate how many factors are predicting a useful amount of variance.

**Research Questions:**

* Number of underlying patterns (factors/components): How many best fit the data?
  + Does this match the expected theory?
* Scale development: building a new measure, does it match your expected theory? Does it measure what you are expecting it to measure?
  + What are the underlying pieces? How do the questions group together?
  + What questions can we eliminate as not being important?

**Power:**

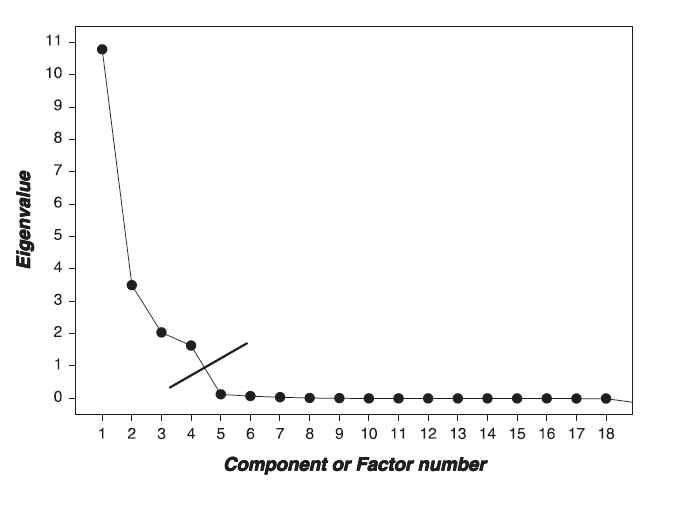
* Generally power is a problem of sample size, since PCA/EFA are testing model fit. Either your model is going to fit in an expected way or not. If power is a concern, the solution is to test more participants.
* Large sample sizes are needed for either analysis, and usually scales are tested several times. If you have a large dataset, people will often randomly split them to get two tests of the model as well.

**Assumptions:**

1. Number of variables – EFA/PCA groups information, so only using 5 variables doesn’t allow you to create groups easily. At least 10 variables is recommended.
2. Sample Size – both analyses require a LOT of people. Here are the rules of thumb.
   1. Very minimum = 100 people.
   2. Safe = 200 people.
   3. Best bet 300 or more.
3. Missing data – neither analysis will allow missing data. Replace the data or eliminate the participants.
4. Outliers – multivariate outliers are combinations of answers on the questions that are strange. When you are trying to group questions, you do not want people in the study who have strange combinations of answers on those questions.
5. Multicollinearity – variables that are correlated over .95 tends to be a problem mathematically. Best test – run it, if it causes problems, try eliminating or combining variables.
6. Normality – the variables should be normally distributed, but some violations are ok because of the large sample size.
7. Linearity – EFA/PCA are forms of regression, so linearity is assumed.

**Rules:**

1. Number of factors/components – you might have a theory on how many factors you expect from the scale. If you do have a theory, people generally run that many factors and then two more (1 more factor, 1 less factor). If you do not have a theory on how many you are expecting, then you can use the following information:
   1. Scree Plots – scree plots are a visual depiction of the eigenvalues. You will look for the large drop off to figure out how many to use:



* 1. Parallel Analysis – (only in Factor) – this analysis gives you how many factors are greater than chance, which you can use in combination with a scree plot to look at the number of factors.

1. Variable Loading - variables “load” on a factor when they have a value over >.300. You want variables to load onto only one factor.
   1. Split variables – you want to get rid of variables that load onto two or more factors.
   2. Non-Loading Variables – you want to get rid of variables that don’t load on any factor.

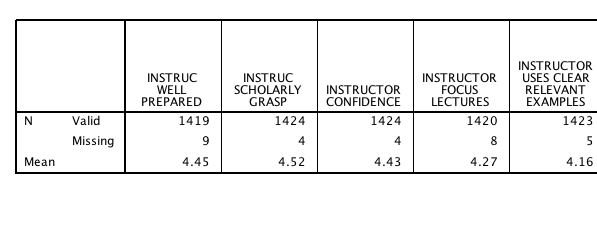
# Complete Example PCA (Factor and SPSS)

**Research Question:** Is there an underlying pattern to the way that students answer questions on their teaching evaluations?

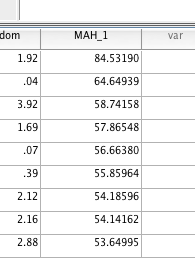
* What components influence their answering?

Assumption Checks:

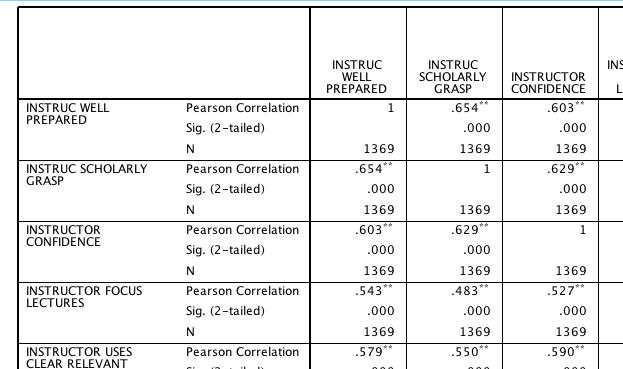
1. Number of variables – There are 12 questions on the end of year evaluations.
2. Sample Size – we have a large sample = N = 1428
3. Missing data – we do have lots of missing data.
   1. You can mean replace or use a different replacement method.



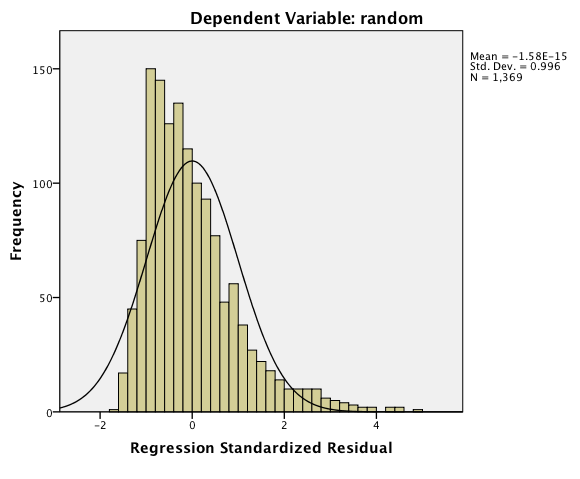
1. Outliers – in this case I will have to create a random variable to figure out the outliers.
   1. Transform > compute.
   2. Label the variable something you will remember (random).
   3. Use a random number generator (like RV ChiSq) – don’t forget you need a number to fill in the question mark.
   4. Hit ok.
   5. Run a fake regression to get your Mahalanobis values.
      1. Analyze > Regression > Linear.
      2. Random variable in the DV.
      3. All your variables in the IV.
      4. Save > Mahalanobis.
   6. We have 59 people over our cut off of 32.91 for 12 variables p<.001.
   7. Sometimes outliers don’t do a thing to the fit of the model, sometimes they do. If you have a small sample size and don’t want to eliminate that many outliers, you can test with and without.
   8. With my very large sample size, I’m going to exclude them.



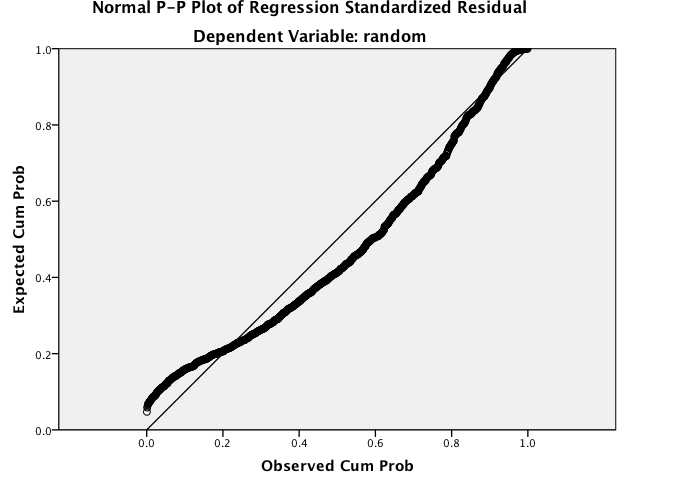
1. Multicollinearity
   1. Analyze > correlate > bivariate.
   2. Here’s the thing – you **want** them to be correlated – but not so much that mathematically it explodes.
   3. Check for r values over .95 or .99.



1. Normality – rerun your fake regression and ask for the residual plots (you especially have to rerun if you excluded outliers).
   1. Analyze > regression > linear.
   2. Plots > ZPRED in Y, ZRESID in X, normality pp, histogram.
   3. Not 100% normal, but we have a lot of participants, so it should be ok.

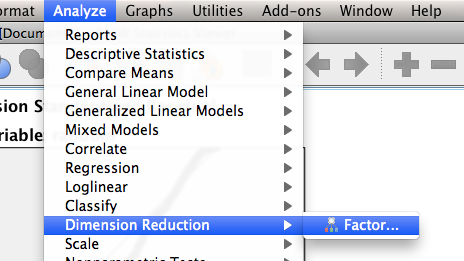


1. Linearity

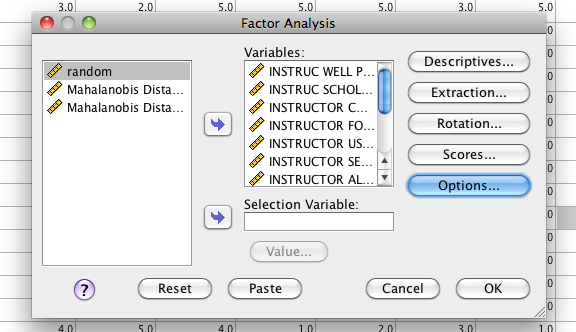


**How to Run Analysis SPSS:**

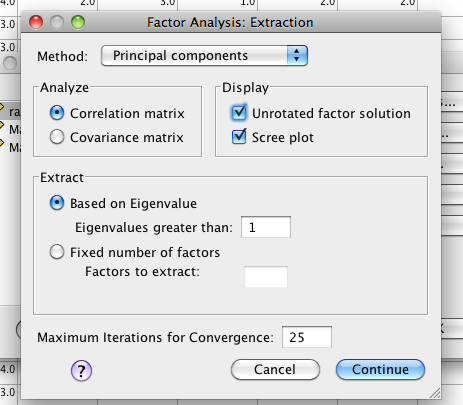
1. Analyze > Dimension Reduction > Factor.



1. Put all your variables in the variables box.



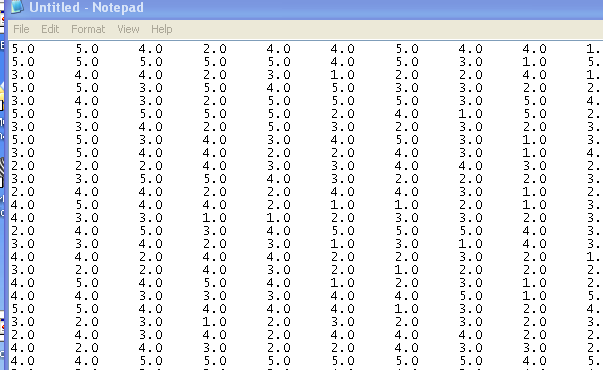
1. Hit extraction.
   1. For PCA – make sure the method says Principal components.
   2. Check for a scree plot – to help you decide on the number of factors.
   3. If you do not know how many factors you want to run, just let it decide based on Eigenvalues. You can change it later.
   4. If you do know the number of factors you want, select the second option.



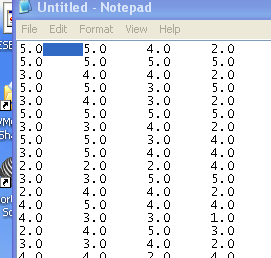
1. PCA does not rotate, so you do not need the rotation options.

**How to run in Factor:**

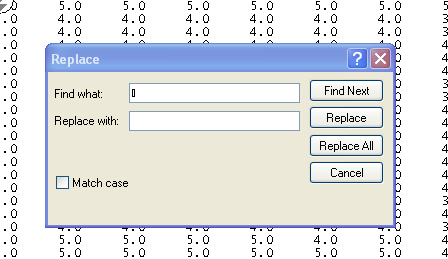
1. First you need to save your data as “space delimited”.
2. You can do File > Save as > Tab Delimited, but that only gets you half way there.
3. Copy all the data.
4. Paste the data into Notepad or some other simple text editing software (TextWrangler is cool stuff for Macs).



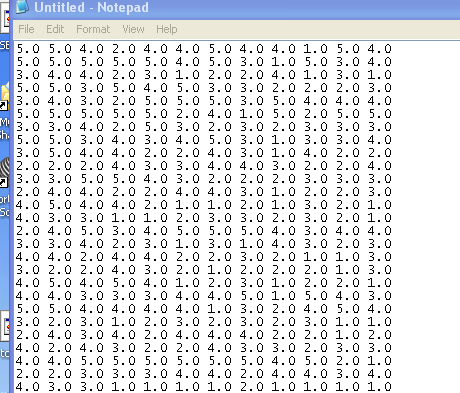
1. Delete all the tabs. Here’s the easy way (for novices). Highlight one of the tabs and right click (or ctrl c) copy.



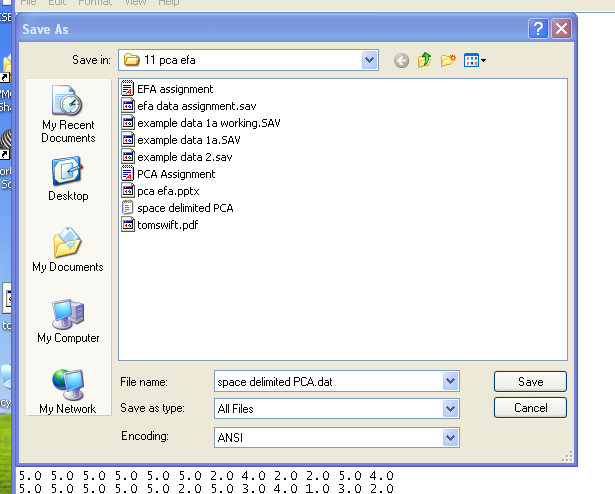
1. Then hit Edit > Replace.
2. Paste the “tab” into the find box. If you’ve done it right, you’ll get a small box in the find box.



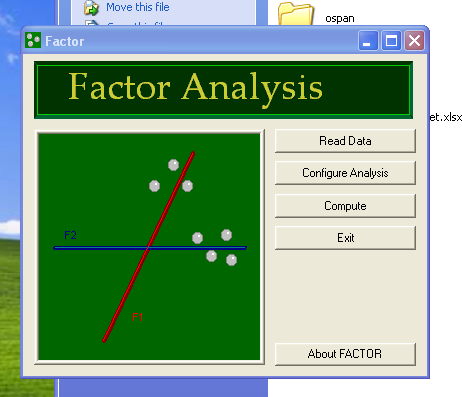
1. Hit the space bar in the replace box (important!! Or you’ll have a bunch of crazy numbers).
2. Hit replace all.



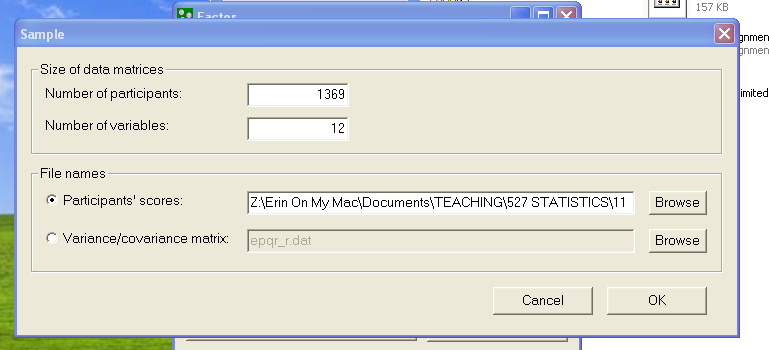
1. Save your file. Hit File > Save As. Change the type to ALL FILES. Be sure to type “.dat” at the end. You are creating a .dat or data file. It reads the same as a text file and will open with Notepad, word etc.



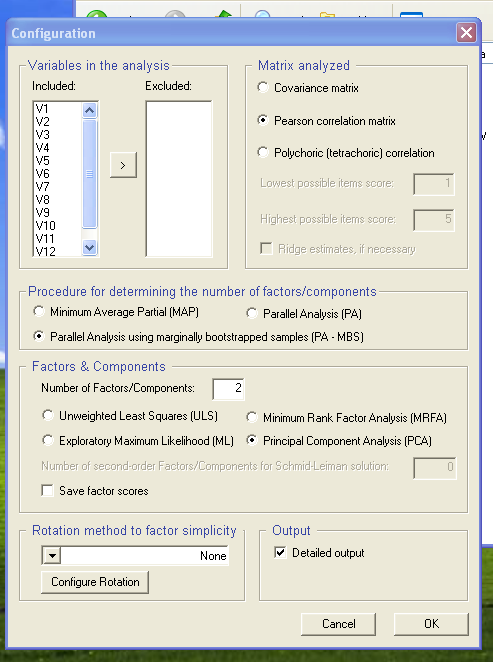
1. Open Factor.



1. First, hit read data.
2. Enter the number of ROWS in the dataset for participants. How many people did you have?
3. Enter the number of COLUMNS for variables. How many questions did you have?

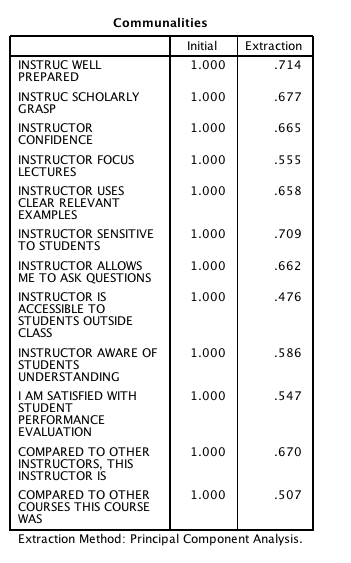


1. Hit ok. If you’ve done this right – the front screen of Factor should say “Ready!”.
2. Hit configure analysis.
3. Variable screen:
   1. Variables – the most confusing part to people is this variable side. They are all included automatically. If you move them to the right, you will leave them out. This set up is backwards from SPSS.
   2. Procedures for factors components: Leave on Parallel Analysis Bootstrapped. This option will tell you how many components are greater than chance.
   3. Under Factors/Components be sure to pick Principal Components Analysis!
   4. Under rotation – pick none so you don’t confuse yourself when looking at the output.
   5. Hit Ok.
4. Hit compute – you’ll get an open text file when done.

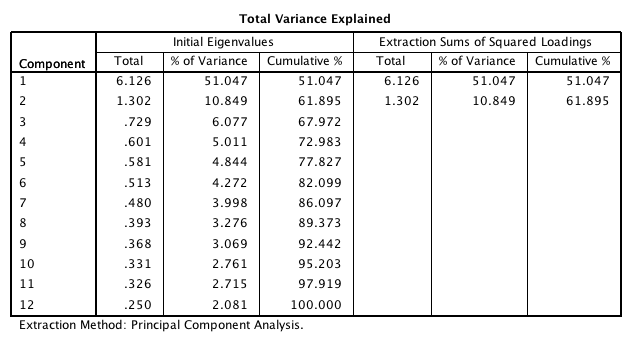
****

**Reading the Output SPSS:**

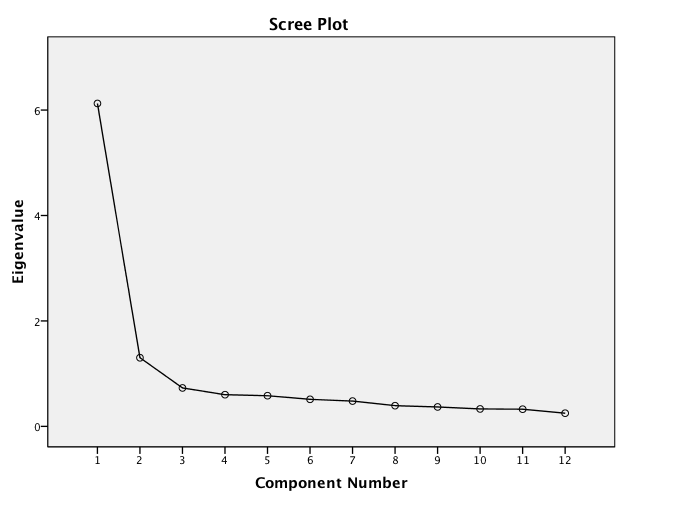
1. Communalities – communality is the amount of variance accounted for in that question alone. We got 71% of the variance covered in our well prepared variable.



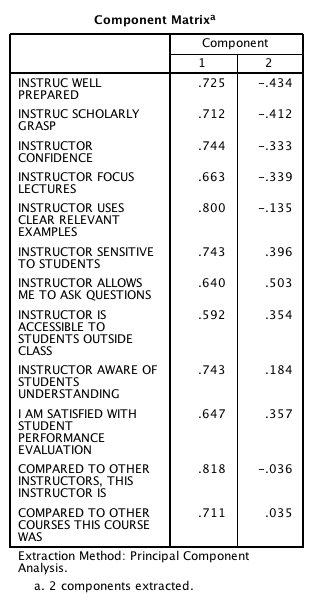
1. Total variance explained – This box gives you the eigenvalues – you well get as many eigenvalues as you have questions.
   1. The “total” column is the eigenvalue – meaning the grouping of the variance that is the highest.
   2. The % variance is the amount of variance taken up by just that component.
   3. Cumulative is total variance of those eigenvalues added up in order.

****

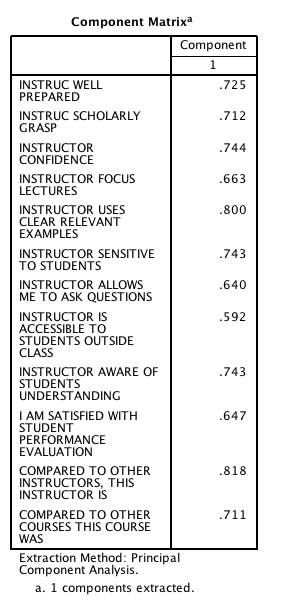
1. The scree plot is a visual depiction of the eigenvalues. You use this plot to figure out how many components to run. For example, there’s a very large drop off after 1 eigenvalue, so I’d probably use one component.

****

1. Component Matrix – is the important box! You will look here for loadings. Here are the rules again:
   1. You want things to load over >.300.
   2. You want things to not split components or have two loadings over >.300 (eek!).
   3. You want things to load on at least one component.
   4. Our problem here is that everything loads on the first component, so when we include the second one, they start to split. This pattern is a strong indication that there’s only one component to answering teaching eval questions.

****

1. Unless everything goes perfectly, you’ll probably have to run at least one or two more models. Here I’m going to run the 1 component model due to the scree plot and the matrix loadings. (in both programs). I’d only report the final model however.



**How to read the Output Factor:**

1. A lot of this output will be redundant / the same. The first thing you want to find is the parallel analysis.

PARALLEL ANALYSIS (PA - MBS)

Lattin, Carroll, & Green (2003)

Variable Real-data Mean of random 95 percentile of random

eigenvalues eigenvalues eigenvalues

1 6.12504\* 1.15439 1.19081

2 1.30208\* 1.11437 1.14266

3 0.72941 1.08521 1.10791

4 0.60141 1.05790 1.07796

5 0.58121 1.03247 1.05028

6 0.51270 1.00921 1.02636

7 0.47983 0.98701 1.00526

8 0.39305 0.96336 0.98055

9 0.36828 0.94004 0.95903

10 0.33142 0.91439 0.93370

11 0.32589 0.88777 0.91106

12 0.24969 0.85388 0.88118

\* Advised number of dimensions: 2

1. You are looking here for the advised number of dimensions. What a parallel analysis tests is a fake sample to see if the eigenvalues are greater than chance. Here it actually suggests two, but as seen in the SPSS output above, that was kind of a mess.
2. Next look for Fit Indices (root mean squared residual – sometimes listed as RMSR mostly listed as RMR in reporting). It’s at the bottom.

Root Mean Square of Residuals (RMSR) = 0.0602

* 1. RMR <.10 is Ok fit.
  2. RMR <.05 is good fit.
  3. We’re close, but the loading matrix was still bad with two factors.

1. Rerun the analysis and change to 1 component.
2. Look at the fit indices then – the Parallel analysis will not change.
   1. Root Mean Square of Residuals (RMSR) = 0.0877
   2. RMR almost always goes DOWN with more components (because you are accounting for more variance), so the decrease didn’t put me above .10, which I would say is good business.

**Example Write Up:**

Results

A Principal Components Analysis was used to analyze end of term teaching evaluations with SPSS and FACTOR. We examined these questions to understand the pattern of answers given to teaching evaluations and to see if those patterns produced distinct components. Data were screened for multivariate assumptions, and 59 participants were excluded as multivariate outliers (Mahalanobis X2(12) > 32.91) leaving 1,369 participants in the analysis.

A parallel analysis indicated that two components were greater than chance, while a scree plot examination indicated that only one component might best it the data. Both one-component and two-component models were run. Table 1 shows the component loadings for the one-component model. All questions loaded highly on the first component, and when a second component was added, questions began to split loadings. The RMR was better for the two-component model (0.06) than the one-component model (0.09), but there was not a significant loss of fit by eliminating the component. The one-component model indicated that participants used the same cognitive processes to make teaching evaluation ratings across all questions and did not have multiple patterns to answering questions.

Table 1. *Component Loadings for One-Component Model*

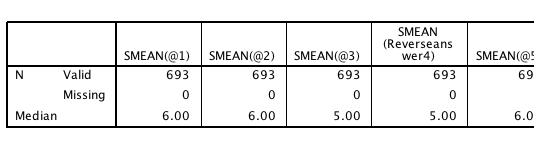
|  |  |
| --- | --- |
| Question | Loading |
| Instructor Well Prepared | 0.725 |
| Scholarly Grasp | 0.712 |
| Confidence | 0.744 |
| Focus Lectures | 0.663 |
| Clear Relevant Examples | 0.800 |
| Sensitive To Students | 0.743 |
| Allows Me to Ask Questions | 0.640 |
| Accessible Outside Class | 0.592 |
| Aware of Student Understanding | 0.743 |
| Satisfied with Student Performance Evaluations | 0.647 |
| Compared to Other Instructors | 0.818 |
| Compared to Other Courses | 0.711 |
|  |  |

# Complete Example EFA (Factor and SPSS)

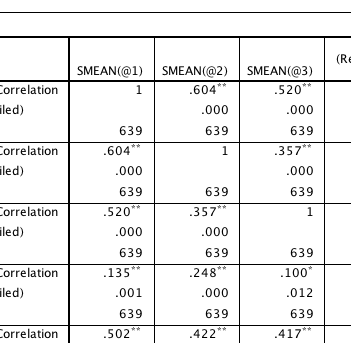
**Research Question:** We are trying to develop of scale of understanding how people feel about therapy. However, it’s currently 55 questions long. Are all of those questions necessary? Can we get rid of some? And after we get rid of some – what are the underlying factors in the dataset?

**Assumption Checks:**

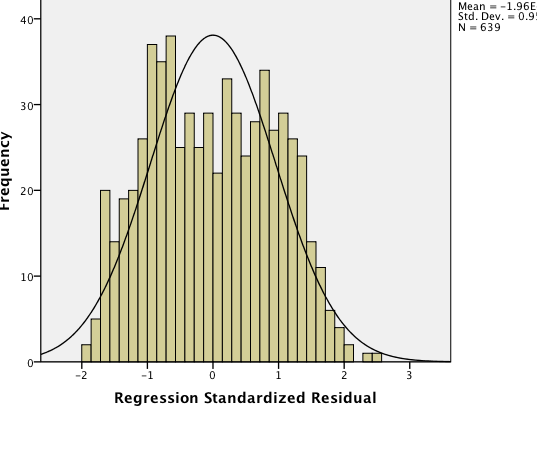
1. Number of variables – 55 questions, so we are good.
2. Sample Size – 693 – also good.
3. Missing data – none (this data set had already been mean replaced).



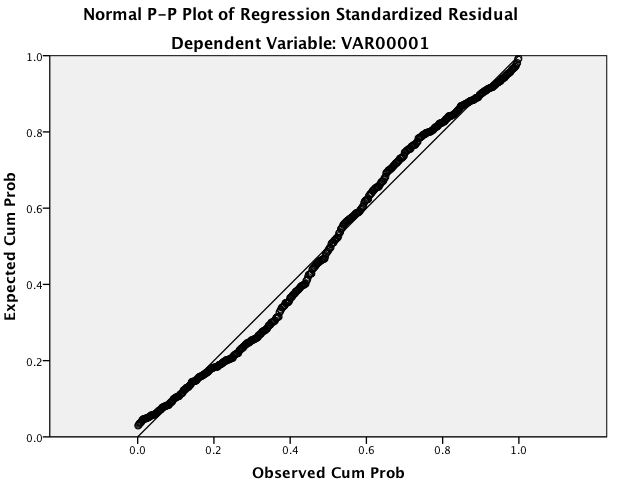
1. Outliers – see above how to run outliers analysis. Mahalanobis cut off with 55 questions = 93.17. Eliminated 54 outliers that were over the cut off score (which left us with 639 people – still a lot).
2. Multicollinearity – some of them are high, but none appear to be over .95 or so.



1. Normality

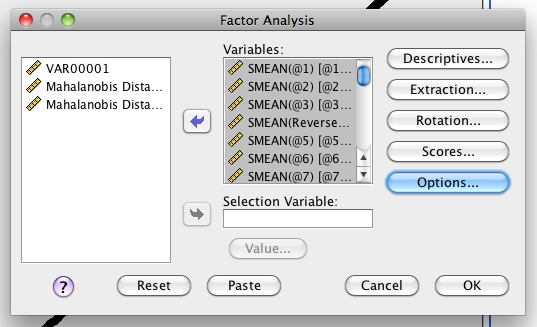


1. Linearity

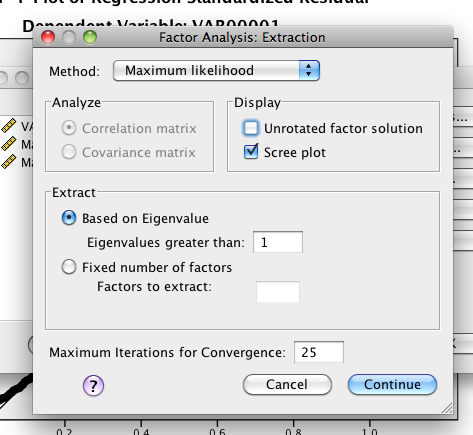


**How to Run Analysis SPSS:**

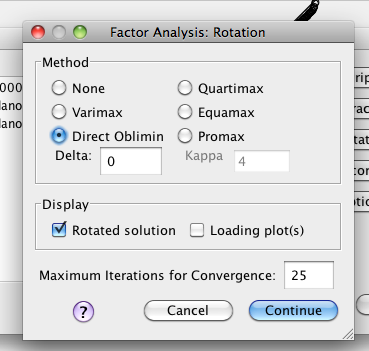
1. Analyze > Dimension Reduction > Factor.
2. Put all your variables in the variable box.



1. Hit extraction.
   1. Change to Max Likelihood. If this crashes and burns, Unweighted Least Squares is also an option. You can also up the max iterations to help with analyses that don’t want to converge (i.e. give you output).
   2. Here I turn off the unrotated solution, so I don’t look at the wrong box.



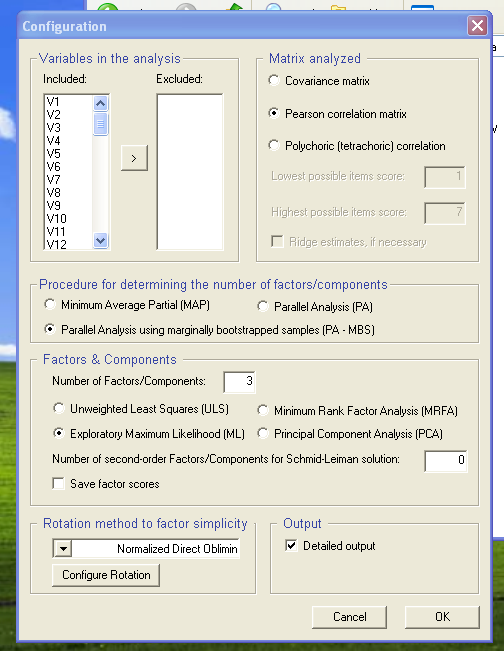
1. Hit rotation > direct oblimin (remember orthogonal rotations are dumb).



1. Hit continue and ok. You can use the other options to get the factor coefficients (the regression equations used to create the factor analysis.
2. Again, to run in Factor – you will need to save the data as space delimited. See above in PCA to get a space delimited version.

**How to Run Analysis Factor:**

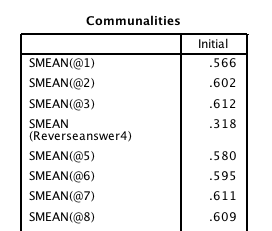
1. See above on how to read the data. Hit read data and enter the number of row (participants) and columns (questions).
2. Hit configure analysis.
3. Use parallel analysis like it’s listed.
4. Use Maximum Likelihood.
5. Rotation – you have a lot of options. If you want your numbers to exactly match SPSS use Normalized Direct Oblimin.
6. Hit ok.



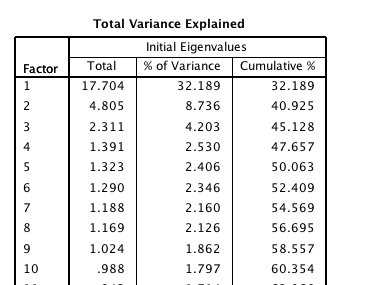
**Reading the Output SPSS:**

(in an effort to save space, I’ve only copied part of the boxes).

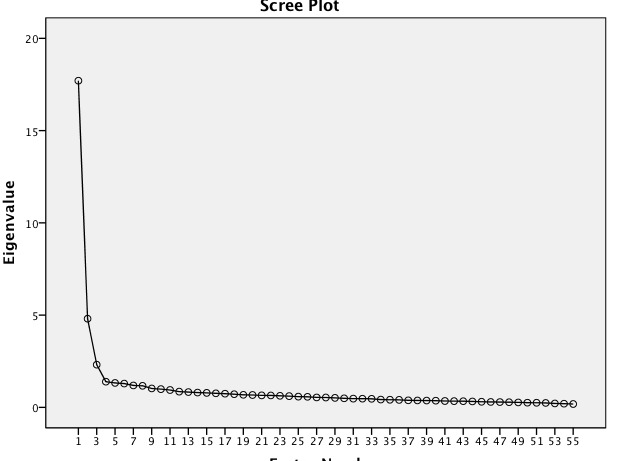
1. Communalities – communality is the total amount of variance accounted for by that question in the whole model (here across 3 factors).



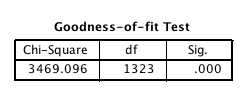
1. Eigenvalues – eigenvalues are the total variance explained by each factor. The first factor obviously has a lot of variance (32%).



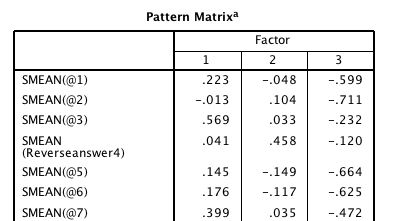
1. Scree plot – plot of the eigenvalues, used to figure out the number of factors.



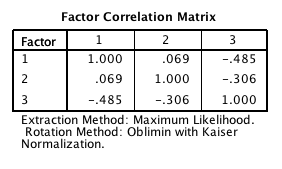
1. Goodness of Fit – you actually want this to be non-significant. The problem is that with large df, almost anything is significant. This is why the Factor program is useful. You can report more goodness of fit statistics.



1. Pattern Matrix – these are the unique relationships of each question with each factor. Think about this like pr in regression. You want them to load over .3 and only on one factor.



1. Structure Matrix – ignore this output.
2. Correlations Between Factors – if this number is greater than zero, you will use a oblique rotation.



*Note:* This output is only for the first round of the factor analysis – as you will see in the write up, we ran through several tests of the model using parallel analyses and eliminating questions as we went.

**Reading the Output Factor:** (note this is the final model, see below).

1. Gives you a correlation table automatically.
2. Gives you the same eigenvalue table as presented in SPSS.
3. Go down to parallel analysis:
   1. Here it suggests 3 factors are greater than chance (same as PCA parallel analysis).

PARALLEL ANALYSIS (PA - MBS)

Lattin, Carroll, & Green (2003)

Variable Real-data Mean of random 95 percentile of random

eigenvalues eigenvalues eigenvalues

1 17.70073\* 1.61890 1.67678

2 4.80537\* 1.56250 1.60735

3 2.31058\* 1.51955 1.55656

4 1.39071 1.48406 1.51788

5 1.32324 1.45147 1.48037

6 1.29099 1.42110 1.44887

7 1.18706 1.39297 1.42014

8 1.16938 1.36640 1.39152

1. Fit indices IMPORTANT. – These values tell you how “good” your model is:
   1. RMSEA ok at < .10.
   2. RMSEA good/great/yay at <.05
   3. Chi-square you want to be small, but it is biased by large sample sizes (which are required! Doh!), so people generally ignore them.
   4. NNFI, CFI, GFI, AGFI “ok” at >.80.
   5. NNFI, CFI, GFI, AGFI “good” at >.90.
   6. NNFI, CFI, GFI, AGFI “great” at >.95.
   7. RMR ok at < .10
   8. RMR good at <.05

GOODNESS OF FIT STATISTICS

Root Mean Square Error of Approximation (RMSEA) = 0.05

Estimated Non-Centrality Parameter (NCP) = 1646.04

Degrees of Freedom = 1032

Test of Approximate Fit

H0 : RMSEA < 0.05; P = 0.481

Chi-Square with 1032 degrees of freedom = 2681.625 (P = 0.000010)

Chi-Square for independence model with 1176 degrees of freedom = 16056.053

Non-Normed Fit Index (NNFI; Tucker & Lewis) = 0.87

Comparative Fit Index (CFI) = 0.89

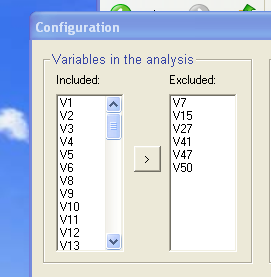
Goodness of Fit Index (GFI) = 0.99

Adjusted Goodness of Fit Index (AGFI) = 0.99

1. Gives you unrotated solution – most people ignore this.
2. Rotated factor solution – will be the same as the pattern matrix in SPSS if you use the normalized direct oblimin solution.
3. Rotated solution with holes – you NEVER report this sort of thing, everyone wants to see all the loadings. However, this sort of chart will allow you to see simply where things don’t load (blank) and things split load (loadings on two factors).

**Running this particular analysis:**

1. I started with 55 questions.
   1. Scree plot suggested 3 factors.
   2. Parallel Analysis suggested 3 factors.
   3. Since I ran three factors, I looked at the loadings.
      1. Split: 7 15 27 47 50
      2. No loading: 41
   4. Eliminated those questions.
2. Now running as 3 factors with 49 questions.



1. Now everything loads on one factor only, and no non-loadings.
   1. Parallel Analysis still suggested 3 factors.
   2. I want to look at the fit indices here to see how well I did.
   3. *Note.* I pasted the final model above on how to read the output.
2. You can get *more* restrictive (i.e. factors must load at least .4 or higher) if you want to get rid of questions. You cannot get *less* restrictive.

**Example Write Up:**

Results

An exploratory factor analysis (EFA) was used to analyze the underlying factors in the Measure of Disseminatability Questionnaire using the FACTOR program and SPSS. Data were screened for multivariate assumptions (normality, linearity, etc.), and all assumptions were met. 54 multivariate outliers were detected using Mahalanobis distance (X2(55) = 93.17), and these participants were eliminated from further analyses. No missing data were present. The following EFA analyses were conducted using guidelines outlined in Preacher and MacCallum (2004).

A parallel analysis and scree plot examination suggested 3 overall factors, therefore a 3-factor model was tested. The model had adequate fit with several problems. 5 questions split across several factors, and 1 question did not load onto any factor (loadings > .300). These questions were eliminated from further analyses. Another 3-factor model was tested, and the factor loadings are presented in Table 1. This model had good fit with all questions loading onto only one factor. The RMSEA indicated excellent fit at .05, while the CFI (.89), GFI (.99), and AGFI (.99) indicated a mix of good and excellent fit indices.

Factor 1 contained the most questions, with 27 questions. See Appendix A for the questionnaire. An analysis of these questions indicated that Factor 1 taps into ratings of effectiveness and an overall “how good is the treatment” factor. Factor 2 had 11 questions, and seemed to measure the negative feelings or effects someone might have when choosing a treatment. Finally, Factor 3 contained 11 questions and appears to measure the cost and therapist knowledge of a potential therapy.

Table 1. *3-Factor Model Loadings.*

|  |  |  |  |
| --- | --- | --- | --- |
| Question | Factor 1 | Factor 2 | Factor 3 |
| 3 | **0.593** | 0.039 | 0.183 |
| 8 | **0.679** | 0.000 | 0.089 |
| 9 | **0.677** | 0.029 | 0.175 |
| 10 | **0.456** | 0.016 | 0.263 |
| 11 | **0.519** | 0.069 | -0.056 |
| 12 | **0.592** | 0.014 | 0.190 |
| 18 | **0.661** | 0.049 | 0.199 |
| 20 | **0.674** | -0.028 | 0.064 |
| 21 | **0.777** | 0.021 | 0.156 |
| 24 | **0.756** | 0.065 | -0.088 |
| 25 | **0.652** | 0.085 | 0.108 |
| 26 | **0.532** | 0.107 | 0.171 |
| 29 | **0.521** | 0.054 | 0.258 |
| 33 | **0.687** | 0.010 | 0.017 |
| 34 | **0.779** | -0.024 | -0.161 |
| 35 | **0.646** | 0.062 | 0.033 |
| 38 | **0.448** | -0.204 | -0.109 |
| 40 | **0.662** | 0.003 | 0.032 |
| 42 | **0.389** | 0.095 | 0.140 |
| 43 | **0.552** | 0.083 | 0.092 |
| 44 | **0.631** | 0.069 | 0.119 |
| 45 | **0.770** | 0.028 | 0.098 |
| 46 | **0.656** | 0.116 | 0.102 |
| 49 | **0.489** | -0.059 | 0.073 |
| 53 | **0.681** | 0.052 | 0.025 |
| 54 | **0.609** | -0.007 | 0.083 |
| 55 | **0.553** | 0.049 | 0.033 |
| 4 | 0.036 | **0.458** | 0.137 |
| 13 | 0.098 | **0.638** | -0.156 |
| 14 | -0.060 | **0.586** | 0.124 |
| 19 | 0.003 | **0.569** | -0.086 |
| 23 | -0.220 | **0.483** | -0.059 |
| 30 | 0.148 | **0.303** | 0.151 |
| 31 | 0.098 | **0.659** | -0.101 |
| 36 | 0.166 | **0.564** | 0.092 |
| 48 | -0.242 | **0.471** | 0.273 |
| 51 | 0.256 | **0.591** | -0.141 |
| 52 | -0.077 | **0.463** | 0.125 |
| 1 | 0.219 | -0.052 | **0.604** |
| 2 | -0.027 | 0.098 | **0.738** |
| 5 | 0.147 | -0.142 | **0.651** |
| 6 | 0.182 | -0.111 | **0.605** |
| 16 | 0.255 | 0.209 | **0.451** |
| 17 | 0.139 | -0.017 | **0.400** |
| 22 | 0.043 | 0.189 | **0.628** |
| 28 | -0.304 | 0.099 | **0.084** |
| 32 | 0.168 | -0.070 | **0.484** |
| 37 | 0.113 | -0.057 | **0.407** |
| 39 | 0.054 | 0.197 | **0.580** |

*Note*. Factor loadings have been sorted and bolded for ease of reading.