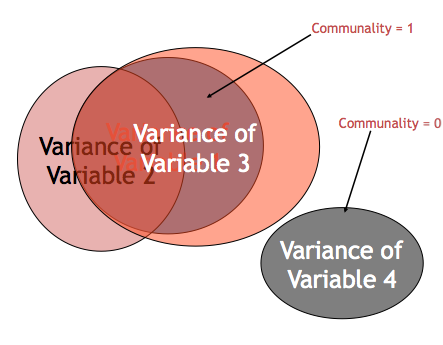
PCA and EFA

**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, therefore the *total* variance is used.
* 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, the left over variance is considered error.

**Definitions/Abbreviations:**

* Variance Types
  + Common variance = overlapping variance between items (systematic variance)
  + Unique variance = variance only related to that item (error variance)
  + Communality – the common variance for the item
    - You can think of it as R2for that item
  + EFA = describes the common variance
  + PCA = describes common variance + unique variance



* Correlation Matrices
  + 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.
* Matrices in output
  + 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 an 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 – A mathematical representation of the variance accounted for by that grouping of items
  + Confusing part: You will see the number of eigenvalues as you have items because they are calculated before extraction
  + Only a few should be large

**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.
* Rules of thumb:
  + 10-15 participants per item
  + <100 is not acceptable (believe me, I know this).
  + 300 is generally agreed upon as the best; however, most people see it as the gold criteria and are ok with less.
* Ways to test sampling adequacy:
  + Kaiser-Meyer-Olkin (KMO) test – scores closer to one are better

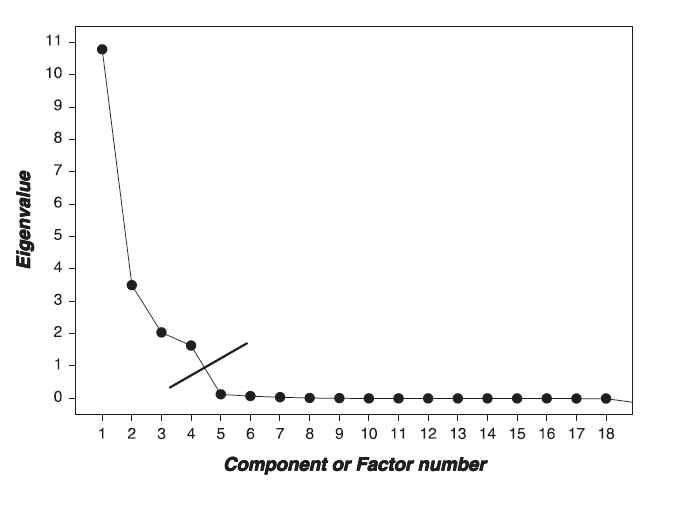
**Assumptions:**

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. Types of variables – without running special types of analysis, interval or ratio data are recommended for EFA/PCA.
3. Sample size – see above under power.
4. Missing data – neither analysis will allow missing data. Replace the data or eliminate the participants.
5. 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.
6. 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.
7. Normality – the variables should be normally distributed, but some violations are ok because of the large sample size.
8. Linearity – EFA/PCA are forms of regression, so linearity is assumed.
9. Homogeneity – you want each of the variables to be normally distributed.
10. Homoscedasticity – you want the error terms to be normally distributed as well (equal spread of each variable down the other variable).

**Rules/Questions to ask yourself:**

**How many factors/components do I have?**

1. **How many factors/components do I have?**
   1. Theory – 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).
   2. 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.
  2. Kaiser criterion – this method is an older rule of thumb that is not well supported anymore. You would look at the number of eigenvalues that are greater than 1 (or .70 in new literature). This rule tends to overestimate the number of factors/components needed.

1. **Can I achieve *simple structure*?**
   1. Simple structure is the final solution of a factor analysis that has the simplest solution.
   2. **How can I get to simple structure (set up the analysis)?**
      1. Factor rotation – process by which the solution is made “better” (smaller residuals) without changing the mathematical properties.
         1. Oblique – Oblimin is the most common rotation. Factors are allowed to be correlation when they are rotated.
         2. Orthogonal – Varimax is the most common rotation. Orthogonal rotation holds factors completely uncorrelated.
      2. Fitting estimation = MATH that is used to determine factor loadings.
         1. For EFA – maximum likelihood is the most common fitting estimation.
         2. For PCA – principle components is the most common fitting estimation.
         3. Half-half – principle axis factoring is the type of fitting estimation that’s sort of both analyses.



* 1. **How do I tell if my set up achieved simple structure?**
     1. Variable Loading - variables “load” on a factor when they have a value over >.300.
     2. You want variables to load onto only one factor.
     3. Split variables – you want to get rid of variables that load onto two or more factors.
     4. Non-Loading Variables – you want to get rid of variables that don’t load on any factor.
     5. Factors/Components with only one/two items loading onto it are considered *unique.* Three to four items are suggested for each factor/component.
     6. What to do if bad items?
        1. In this step you might run several rounds of analyses. Find the bad items, run the EFA/PCA again without them.

1. **How adequate is my model?**
   1. So is that simple structure any good? This step is akin to checking *p*-values and effect sizes to determine if that structure is appropriate (i.e. your means might be different in ANOVA, but you can’t say that unless it’s significant or has a large effect size).
   2. Fit indices – a measure of how well the rotated matrix matches the original matrix
      1. Goodness of fit statistics – want large values, compares reproduced correlation matrix to real correlation matrix
      2. Residual statistics – want small values, look at the residual matrix (i.e. reproduced – real correlation table)

Goodness of fit statistics:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Fit** | **Name** | **Good** | **Acceptable** | **Poor** |
| NNFI/TLI | Non-normed fit index, Tucker-Lewis index | >.95 | >.90 | <.90 |
| CFI | Comparative fix index | >.95 | >.90 | <.90 |
| NFI | Normed fit index | >.95 | >.90 | <.90 |

Residual fit statistics:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Fit** | **Name** | **Good** | **Acceptable** | **Poor** |
| RMSEA | Root mean square error of approximation | <.06 | .06-.08 | >.10 |
| RMSR | Root mean square of the residual | <.06 | .06-.08 | >.10 |

* 1. Reliability – an estimate of how much your items “hang together” and might replicate
     1. Cronbach’s alpha most common
     2. .70 or .80 is acceptable
  2. Theory –Do the item loadings make any sense? Can you label the factor/component?

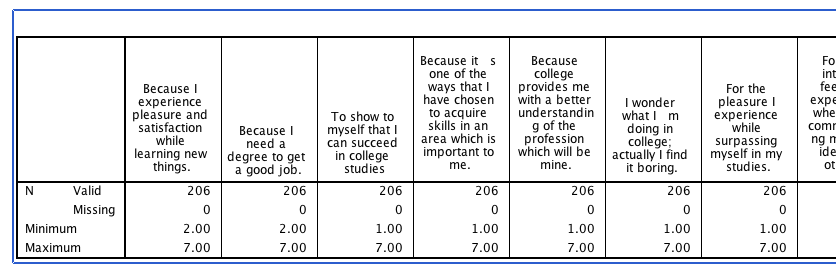
# Complete Example EFA (Factor and SPSS)

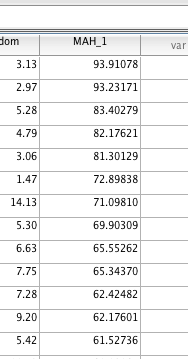
**Research Question:** Many people are interested in self-determination theory – a theory of motivation that “is concerned with supporting our natural or intrinsic tendencies to behave in effective and healthy ways”. In line with this idea, a scale for assessing why students are in college was developed. Here we will test if those questions fit a factor structure.

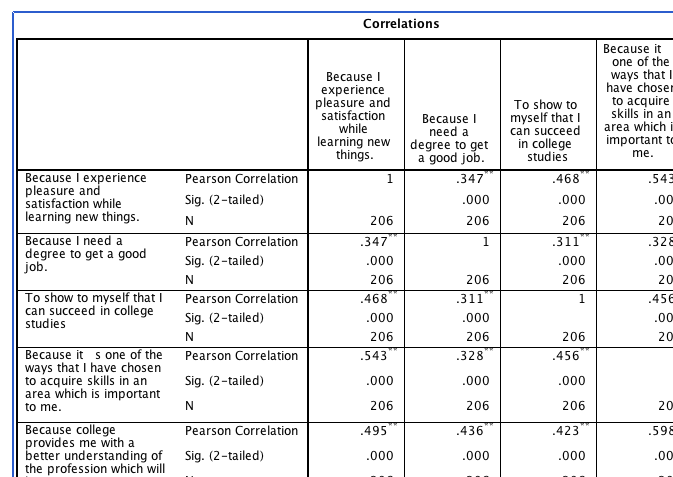
**Assumption Checks:**

Note: you should see previous notes for data screening how to steps.

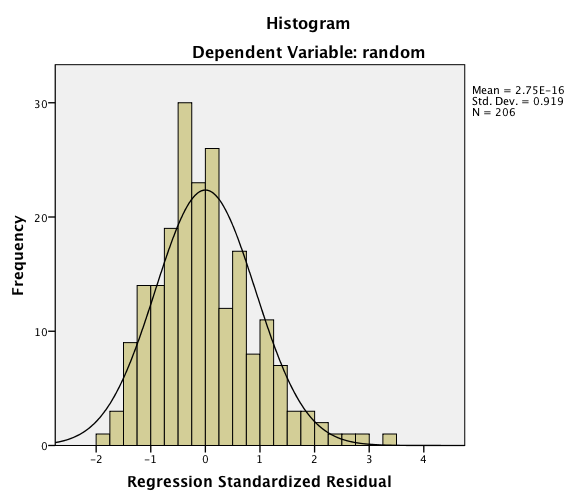
1. Number of variables – 32 questions, so we are good.
2. Types of variables – 1 to 7 Likert scales, which are at least interval.
3. Sample Size – 206 – also good.
4. Accuracy – data should be accurate, so it looks we are good with all questions going from 1-7.



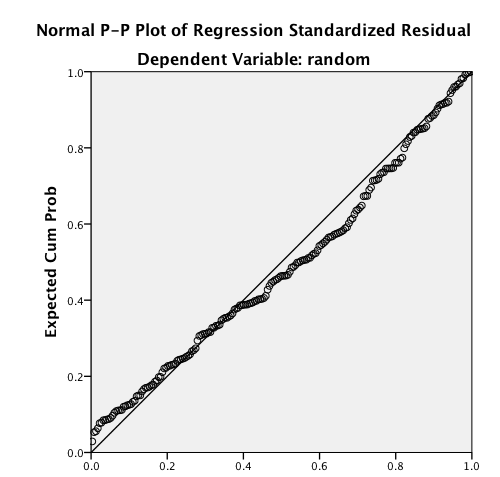
1. Missing data – none as shown above.
2. Outliers – you will run these steps just like a fake regression analysis described in the data screening guide.
   1. Mahalanobis cut off with 32 questions = 62.49.
   2. Got a bunch of outliers:
   3. 
   4. It’s 10 people. Here’s the deal … with EFA sometimes it makes a difference, sometimes it doesn’t at all. Let’s leave them in and see what happens!
3. Multicollinearity – some of them are high, but none appear to be over .95 or so.



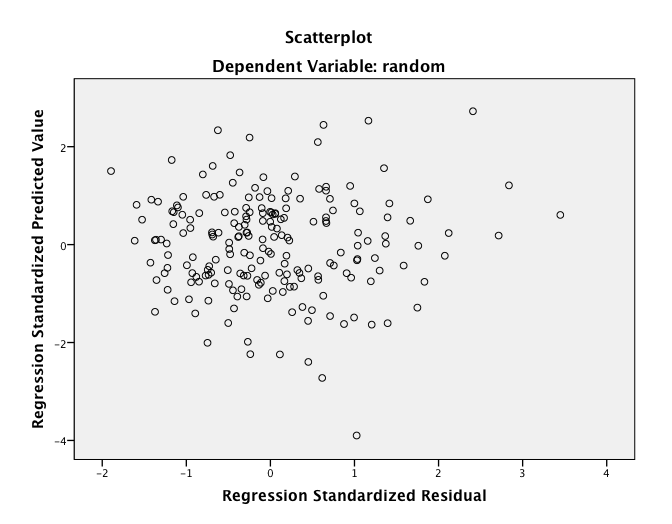
1. Normality



1. Linearity

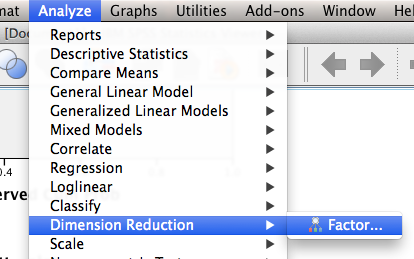


1. Homogeneity/homoscedasticity: the people on the edges are probably my outliers. I could run this analysis with and without to determine if it’s going to make a big difference. Otherwise, the majority of the data appear to be homogeneic / homoscedastic.

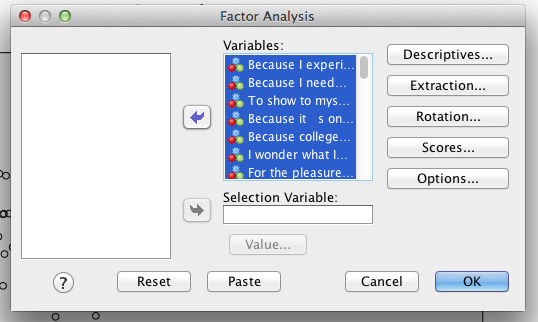


**How to Run Analysis SPSS:**

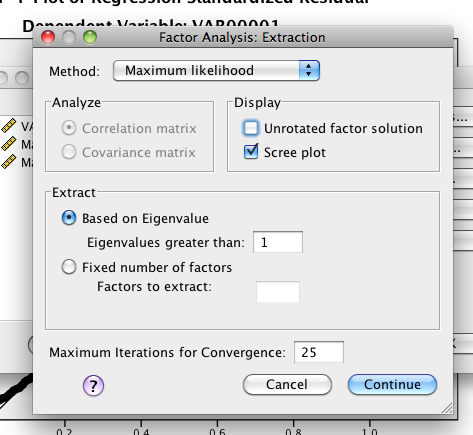
1. Analyze > Dimension Reduction > Factor.



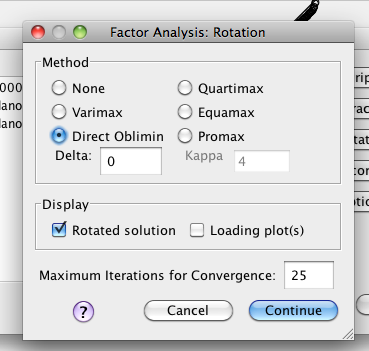
1. Put all your variables in the variable box.



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).

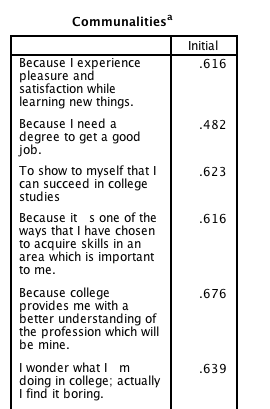


1. Hit continue and ok. You can use the other options to get the factor coefficients (the regression equations used to create the factor analysis.

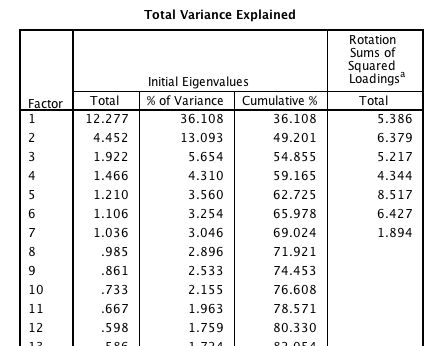
**Reading the Output SPSS:**

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

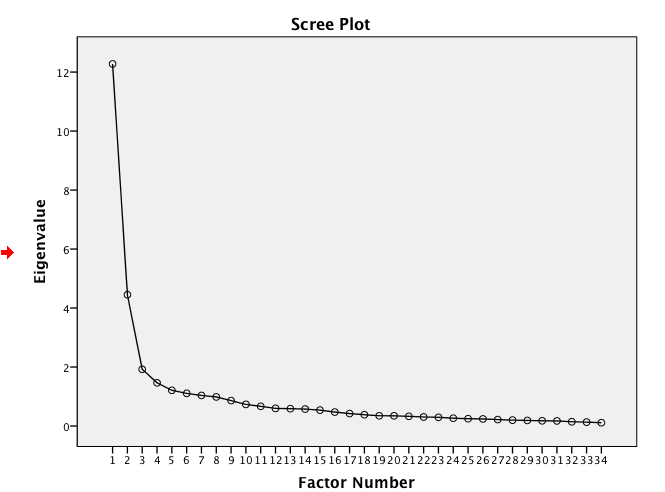
1. Communalities – communality is the total amount of variance accounted for by that question in the whole model (here across 7 factors – clearly you don’t want this many but that was the estimated amount for the Kaiser criterion SPSS does automatically).



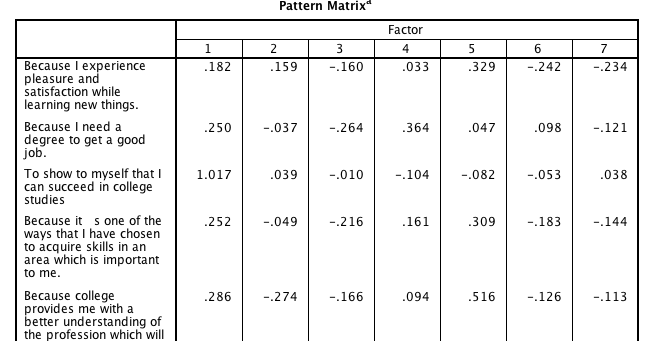
1. Eigenvalues – eigenvalues are the total variance explained by each factor. The first factor obviously has a lot of variance (36%).



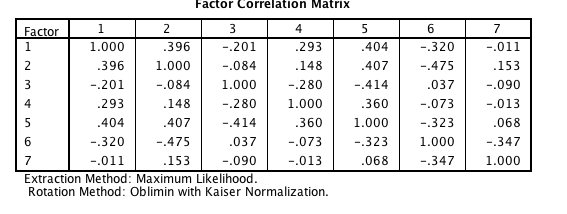
1. Scree plot – plot of the eigenvalues, used to figure out the number of factors.



1. Factor matrix – ignore this output.
2. 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 need to use an oblique rotation.

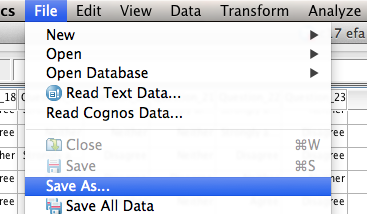


*Note:* This output is only for the first round of the factor analysis – as you will see in the write up, we ran through several tests of the model using parallel analyses and eliminating questions as we went. I switched to FACTOR because it gives me all the information I need – except the Scree plot.

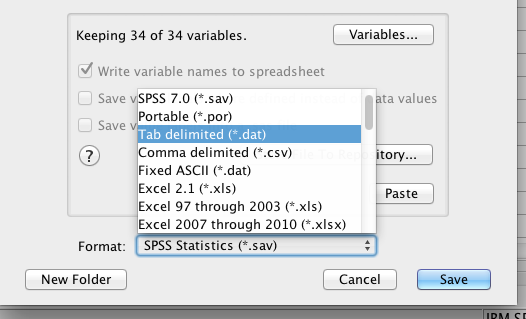
**How to Run Analysis Factor:**

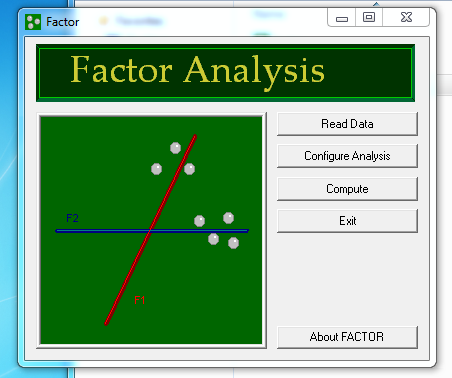
**Link to factor:** [**http://psico.fcep.urv.es/utilitats/factor/**](http://psico.fcep.urv.es/utilitats/factor/)

1. Preparing the data:
   1. You will need a .dat file to start.
   2. File > save as in SPSS.

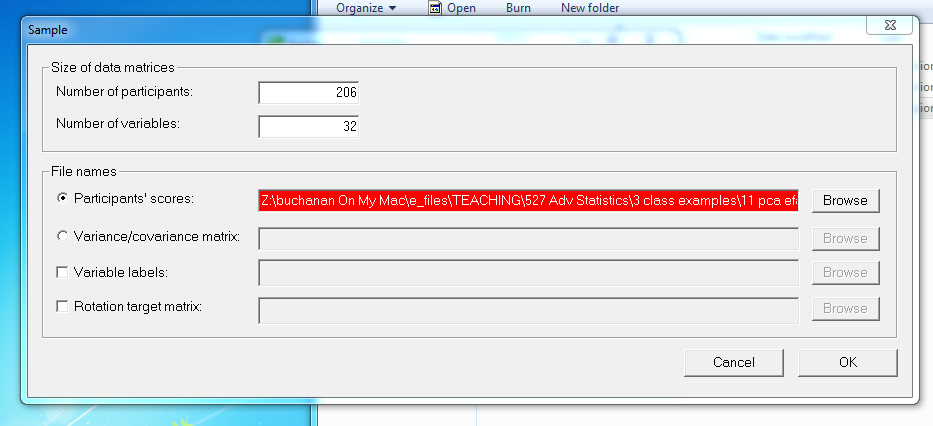


* 1. Change the drop down to Tab delimited .dat.
  2. Save the file somewhere you can find it on the computer.
  3. Make sure you note
     1. Number of participants (rows)
     2. Number of variables (columns)

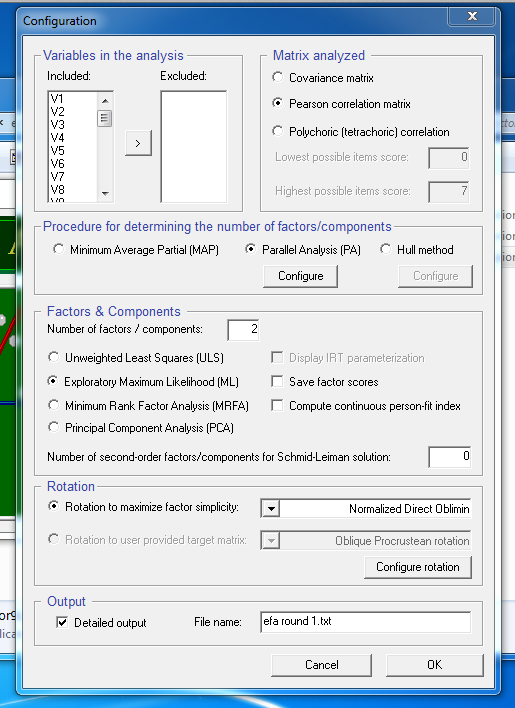




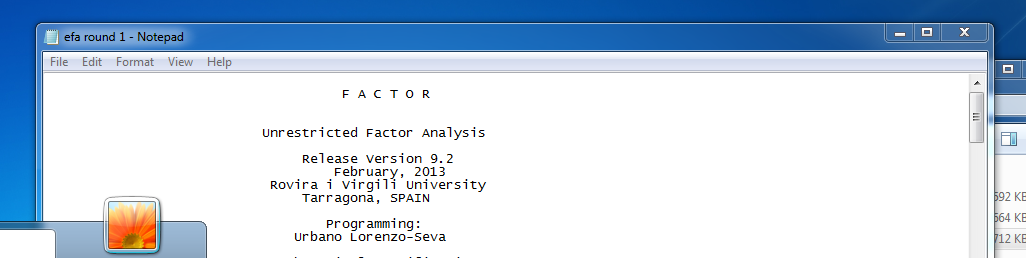
1. Read in the data:
   1. Click read data.
   2. Enter the number of participants (rows).
   3. Enter the number of variables (columns).
   4. Find the .dat file you saved earlier.
   5. If you entered something too large for the variables/participants, it will beep at you. If you did too small, no warning.



1. Hit configure analysis.
   1. Variables you are using are on the left – this is backwards from SPSS, don’t exclude them!
   2. Use parallel analysis like it’s listed.
   3. Pick a number of factors to start (use that scree plot to determine).
   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. Name the file something useful (round number?).
   7. Hit ok.

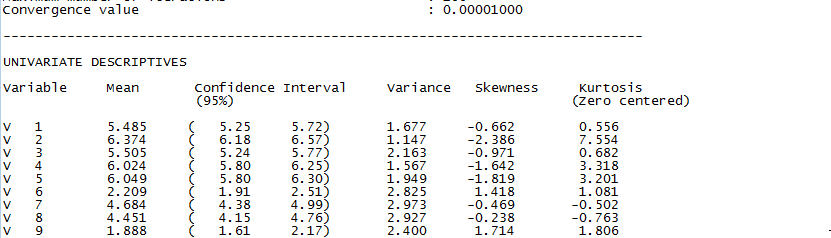


1. Hit compute to run the analysis.
   1. Factor is a cranky old person. Do not do anything else while running it. Seriously, do not click anything or it might crash.
   2. The progress bar means nothing. It gives up rendering after a minute.
   3. A notepad window will open when it’s done.

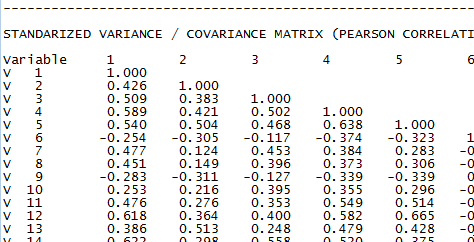


**Reading the Output Factor:**

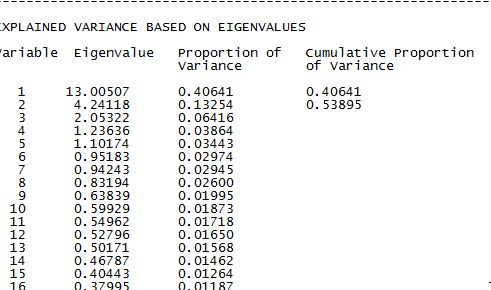
1. Shows you the means and standard deviations for each question.



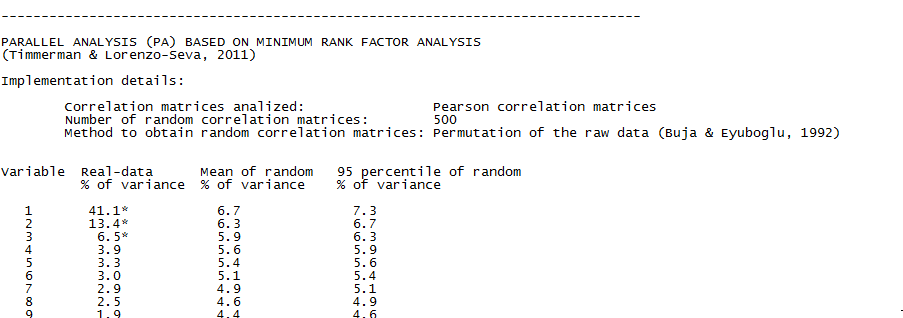
1. Gives you a correlation table of all the questions automatically.



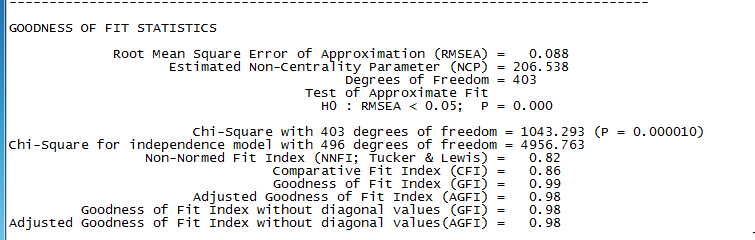
1. Gives you the same eigenvalue table as presented in SPSS.



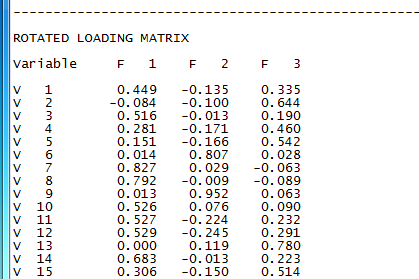
1. Go down to parallel analysis:
   1. Here it suggests 3 factors are greater than chance.



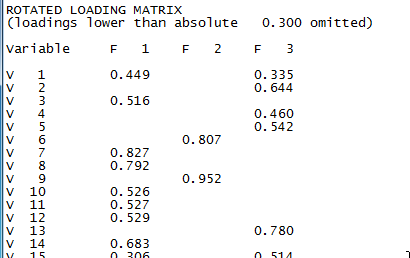
1. Fit indices – this section will help answer the “is my solution adequate?” question.



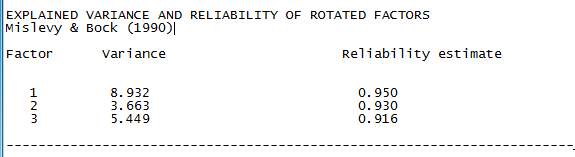
1. Gives you unrotated solution – most people ignore this.
2. Rotated factor solution – will be the same as the pattern matrix in SPSS if you use the normalized direct oblimin solution.



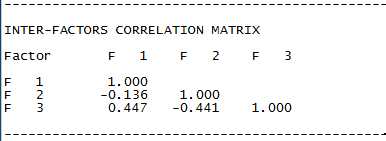
1. Rotated solution with holes – you NEVER report this sort of thing, everyone wants to see all the loadings. However, this sort of chart will allow you to see simply where things don’t load (blank) and things split load (loadings on two factors).

****

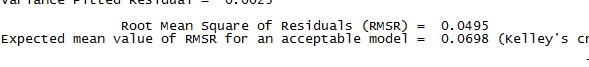
1. Reliability numbers – Cronbach’s alpha.



1. Factor correlation matrix.

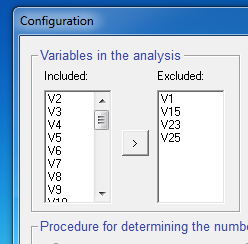


1. Structure matrix – ignore this output.
2. Look for the residuals section – SRMR is listed in that section. (you can also call it RMSR).



**Running this particular analysis:**

1. **How many factors?**
   1. Scree plot suggested 2 factors.
   2. Parallel Analysis suggested 3 factors.
   3. Kaiser criterion suggested 5 factors.
   4. Theory: 2-3 factors were expected.
   5. I went with three factors.
2. **Simple structure:**
   1. **Math:**
      1. Maximum Likelihood.
      2. Normalized direct oblimin because I knew the factors would be correlated.
   2. **Loadings:**
      1. Round 1:
         1. Split: 1, 15, 23, 25
         2. No loading: None
         3. Eliminated those questions.
         4. 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.
      2. Round 2 without bad questions:
         1. Under configure analysis, move those questions over to the right.
         2. Note: factor is odd, control shift does not do what you think it does. Best to move them over one at a time.



* + - 1. Note: be sure to change the name of the output, or it will over-write your first round.
      2. In this round, everything loads *cleanly*: on one and only one factor. Each factor has 3 or more questions. Simple structure = yes.

1. **Adequacy of the solution:**
   1. Goodness of fit:
      1. NNFI: .83
      2. CFI: .87
   2. Residuals
      1. RMSEA: .09
      2. SRMR: .05
   3. Reliability:
      1. Factor 1: .93
      2. Factor 2: .91
      3. Factor 3: .95
   4. What do the factors mean? Look at question themes and give the factors a label.
      1. Factor 1: Doubts about college
         1. Boring, wasting my time, don’t know what doing, should I keep going
      2. Factor 2: Career goals
         1. Job, skills for area, profession skills, degree, job, career, etc.
      3. Factor 3: Intrinsic motivation/love of learning
         1. Success, pleasure of studies, intense feeling, learn lots, satisfaction, pleasure of reading/learning new things, etc.

**Example Write Up:**

**Results**

An exploratory factor analysis (EFA) was used to analyze the underlying factors in the self-determination motivation for college questionnaire using the FACTOR program and SPSS. Data were screened for multivariate assumptions (normality, linearity, etc.), and all assumptions were met. Ten multivariate outliers were detected using Mahalanobis distance (*X2*(32) = 62.49); however, these participants were left in for 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 two-three overall factors, and a 3-factor model was tested based on theory. Maximum likelihood estimation was used with normalized direct oblimin rotation because of expected factor correlation. After testing all 32 questions, four items split across several factors (1, 15, 23, 25) using the criterion that loadings must be greater than .300. These items were eliminated from further analyses. Another 3-factor model was tested, and the factor loadings are presented in Table 1. This model achieved simple structure with each item loading on one and only one factor. This model had moderate fit: the RMSEA indicated moderate fit at .09 and SRMR with excellent fit (.05), while the CFI (.87) and NNFI (.83) indicated room for improvement.

Factor 1 included four questions that appeared to assess a student’s doubt about motivation for college studies with questions like “I wonder what I am doing in college, I actually found it boring”. See Appendix A for the questionnaire. Factor 2 included eight items that assessed career goals for a student, including “Because I need a degree to get a good job” and “Because it was the only way to be considered for the career I want”. Finally, Factor 3 included 16 items that measured the intrinsic motivation of attending college with questions such as “For the pleasure that I experience when I feel completely absorbed by what certain authors have written” and “7. For the pleasure I experience while surpassing myself in my studies”. The reliability of all three factors was very high with .93, .91, and .95 for Factors 1, 2, and 3 respectively.

Table 1. *3-Factor Model Loadings.*

|  |  |  |  |
| --- | --- | --- | --- |
| Item | Factor 1 | Factor 2 | Factor 3 |
| 6 | **0.801** | 0.012 | 0.018 |
| 9 | **0.946** | 0.049 | 0.014 |
| 17 | **0.886** | 0.101 | -0.030 |
| 27 | **0.754** | -0.007 | 0.121 |
| 2 | -0.101 | **0.658** | -0.076 |
| 4 | -0.182 | **0.456** | 0.286 |
| 5 | -0.181 | **0.525** | 0.164 |
| 13 | 0.120 | **0.804** | 0.005 |
| 19 | 0.103 | **0.755** | -0.004 |
| 26 | -0.184 | **0.714** | 0.065 |
| 28 | 0.053 | **0.563** | 0.004 |
| 30 | -0.011 | **0.805** | 0.055 |
| 3 | -0.021 | 0.199 | **0.509** |
| 7 | 0.018 | -0.046 | **0.820** |
| 8 | -0.020 | -0.075 | **0.797** |
| 10 | 0.068 | 0.099 | **0.528** |
| 11 | -0.239 | 0.226 | **0.529** |
| 12 | -0.261 | 0.277 | **0.532** |
| 14 | -0.028 | 0.218 | **0.681** |
| 16 | 0.139 | -0.178 | **0.828** |
| 18 | -0.069 | 0.171 | **0.631** |
| 20 | -0.078 | 0.163 | **0.709** |
| 21 | 0.132 | -0.164 | **0.848** |
| 22 | -0.073 | -0.031 | **0.736** |
| 24 | -0.235 | 0.240 | **0.439** |
| 29 | -0.056 | 0.156 | **0.619** |
| 31 | -0.105 | 0.171 | **0.469** |
| 32 | 0.132 | -0.170 | **0.791** |

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