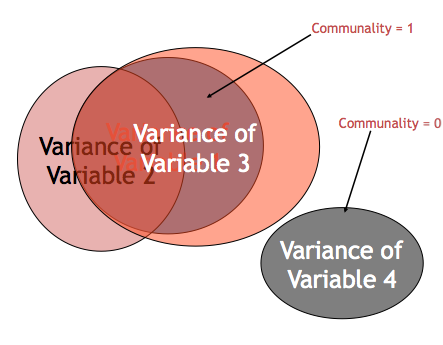
PCA and EFA

**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.
  + 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* the variables, and 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 R2 for 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 reproduced correlation matrix. This matrix will be very small if you had a good fit for your model. The residual matrices are used to calculate how well the analysis went for you.
* 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:**

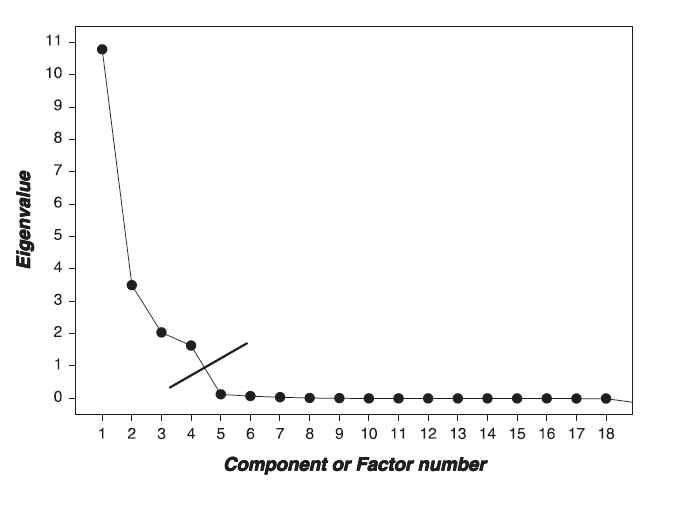
* Generally power is a problem of sample size, since PCA/EFA are testing model fit. Either your model is going to fit in an expected way or not. If power is a concern, the solution is to test more participants.
* Large sample sizes are needed for either analysis, and usually scales are tested several 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 from experience).
  + 300 is generally agreed upon as the best; however, most people see it as the gold criteria and are ok with less.

**Assumptions:**

1. Variables:
   1. Number of variables – EFA/PCA group variables into factors/components, so only using 5 variables does not allow you to create those clusters easily. At least 10 variables are recommended.
   2. Types of variables – without running special types of analysis, interval or ratio data are recommended for EFA/PCA.
2. Sample size – see above under power.
3. Normal data screening:
   1. Screen: Accuracy, missing, outliers
   2. Assumptions: Additivity\*, Linearity, Normality, Homogeneity, and Homoscedasticity
   3. \* Additivity is checked to make sure that variables aren’t perfectly correlated so the analysis will run. You expect variables to be highly correlated, as that is the point of the analysis.

**Rules/Questions to ask yourself:**

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 – this analysis tells 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 (how to 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 (and only one; hence the simple solution name).
     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 cannot say that unless it is 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** | **Excellent** | **Acceptable** | **Poor** |
| NNFI/TLI | Non-normed fit index, Tucker-Lewis index | >.95 | >.90 | <.90 |

Residual fit statistics:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Fit** | **Name** | **Excellent** | **Acceptable** | **Poor** |
| RMSEA | Root mean square error of approximation | <.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

# Exploratory Factor Analysis

**Research Question:**

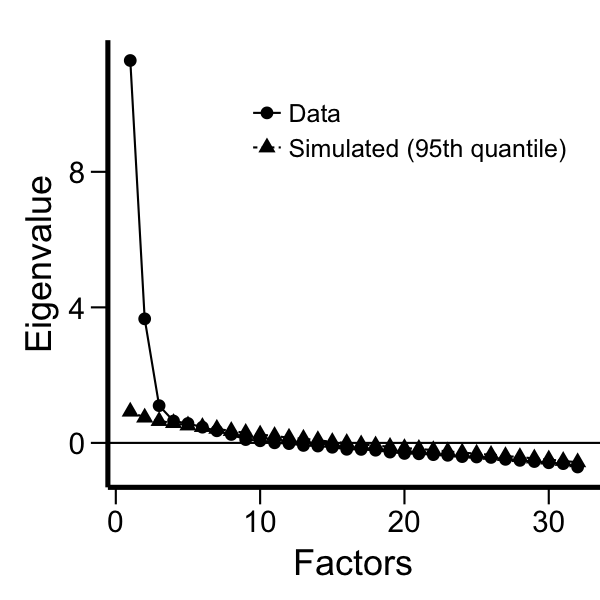
Many people are interested in self-determination theory – a theory of motivation that “is concerned with supporting our natural or intrinsic tendencies to behave in effective and healthy ways”. In line with this idea, a scale for assessing why students are in college was developed. Here we will test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

**How many factors?**

**Parallel analysis suggests 6 factors.**

**Theory two to three factors.**

### Scree Plot



Scree says three factors.

Kaiser says three factors over 1.

**Can I achieve simple structure?**

**Removed questions 4 and 15 for double loading, then simple structure was achieved.**

| **Factor Loadings** | | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | | **Factor 1** | | **Factor 2** | | **Factor 3** | | **Uniqueness** | |
| q1 |  | 0.491 |  | 0.168 |  | 0.221 |  | 0.547 |  |
| q10 |  | 0.528 |  | -0.070 |  | 0.032 |  | 0.716 |  |
| q11 |  | 0.530 |  | 0.241 |  | 0.151 |  | 0.512 |  |
| q12 |  | 0.572 |  | 0.291 |  | 0.173 |  | 0.395 |  |
| q13 |  | 0.009 |  | -0.155 |  | 0.768 |  | 0.492 |  |
| q14 |  | 0.712 |  | -0.004 |  | 0.169 |  | 0.381 |  |
| q16 |  | 0.788 |  | -0.179 |  | -0.179 |  | 0.424 |  |
| q17R |  | 0.029 |  | 0.871 |  | -0.085 |  | 0.298 |  |
| q18 |  | 0.655 |  | 0.053 |  | 0.175 |  | 0.438 |  |
| q19 |  | -0.015 |  | -0.140 |  | 0.739 |  | 0.538 |  |
| q2 |  | -0.088 |  | 0.186 |  | 0.509 |  | 0.647 |  |
| q20 |  | 0.728 |  | 0.052 |  | 0.152 |  | 0.347 |  |
| q21 |  | 0.819 |  | -0.169 |  | -0.152 |  | 0.381 |  |
| q22 |  | 0.736 |  | 0.056 |  | -0.065 |  | 0.477 |  |
| q23 |  | 0.566 |  | 0.166 |  | 0.222 |  | 0.453 |  |
| q24 |  | 0.475 |  | 0.267 |  | 0.157 |  | 0.551 |  |
| q25 |  | 0.641 |  | 0.171 |  | 0.251 |  | 0.312 |  |
| q26 |  | 0.103 |  | 0.232 |  | 0.633 |  | 0.345 |  |
| q27R |  | -0.108 |  | 0.744 |  | 0.033 |  | 0.436 |  |
| q28 |  | -0.018 |  | -0.073 |  | 0.549 |  | 0.737 |  |
| q29 |  | 0.620 |  | 0.038 |  | 0.138 |  | 0.524 |  |
| q3 |  | 0.548 |  | 0.030 |  | 0.089 |  | 0.650 |  |
| q30 |  | 0.077 |  | 0.050 |  | 0.730 |  | 0.384 |  |
| q31 |  | 0.491 |  | 0.109 |  | 0.136 |  | 0.653 |  |
| q32 |  | 0.755 |  | -0.157 |  | -0.155 |  | 0.476 |  |
| q5 |  | 0.196 |  | 0.231 |  | 0.387 |  | 0.609 |  |
| q6R |  | -0.001 |  | 0.780 |  | 0.012 |  | 0.383 |  |
| q7 |  | 0.819 |  | -0.036 |  | -0.046 |  | 0.358 |  |
| q8 |  | 0.775 |  | -0.004 |  | -0.105 |  | 0.446 |  |
| q9R |  | -0.005 |  | 0.923 |  | -0.024 |  | 0.169 |  |
|  | | | | | | | | | |

**Is the structure any good?**

**Eh? RMSEA says it’s ok**

**TLI says not good**

**Reliability**

| **Factor Loadings** | | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | | **Factor 1** | | **Factor 2** | | **Factor 3** | | **Uniqueness** | |
| q1 |  | 0.491 |  | . |  | . |  | 0.547 |  |
| q10 |  | 0.528 |  | . |  | . |  | 0.716 |  |
| q11 |  | 0.530 |  | . |  | . |  | 0.512 |  |
| q12 |  | 0.572 |  | . |  | . |  | 0.395 |  |
| q13 |  | . |  | . |  | 0.768 |  | 0.492 |  |
| q14 |  | 0.712 |  | . |  | . |  | 0.381 |  |
| q16 |  | 0.788 |  | . |  | . |  | 0.424 |  |
| q17R |  | . |  | 0.871 |  | . |  | 0.298 |  |
| q18 |  | 0.655 |  | . |  | . |  | 0.438 |  |
| q19 |  | . |  | . |  | 0.739 |  | 0.538 |  |
| q2 |  | . |  | . |  | 0.509 |  | 0.647 |  |
| q20 |  | 0.728 |  | . |  | . |  | 0.347 |  |
| q21 |  | 0.819 |  | . |  | . |  | 0.381 |  |
| q22 |  | 0.736 |  | . |  | . |  | 0.477 |  |
| q23 |  | 0.566 |  | . |  | . |  | 0.453 |  |
| q24 |  | 0.475 |  | . |  | . |  | 0.551 |  |
| q25 |  | 0.641 |  | . |  | . |  | 0.312 |  |
| q26 |  | . |  | . |  | 0.633 |  | 0.345 |  |
| q27R |  | . |  | 0.744 |  | . |  | 0.436 |  |
| q28 |  | . |  | . |  | 0.549 |  | 0.737 |  |
| q29 |  | 0.620 |  | . |  | . |  | 0.524 |  |
| q3 |  | 0.548 |  | . |  | . |  | 0.650 |  |
| q30 |  | . |  | . |  | 0.730 |  | 0.384 |  |
| q31 |  | 0.491 |  | . |  | . |  | 0.653 |  |
| q32 |  | 0.755 |  | . |  | . |  | 0.476 |  |
| q5 |  | . |  | . |  | 0.387 |  | 0.609 |  |
| q6R |  | . |  | 0.780 |  | . |  | 0.383 |  |
| q7 |  | 0.819 |  | . |  | . |  | 0.358 |  |
| q8 |  | 0.775 |  | . |  | . |  | 0.446 |  |
| q9R |  | . |  | 0.923 |  | . |  | 0.169 |  |
|  | | | | | | | | | |

**1, 3, 7, 8, 10, 11, 12, 14, 16, 18, 20, 21, 22, 23, 24, 25, 29, 31, 32**

**6, 9, 17, 27**

**2, 5, 13, 19, 26, 28, 30**

| **Scale Reliability Statistics** | | | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | | | **mean** | | | | **sd** | | **Cronbach's α** | |
| scale | |  | 5.094 | | |  | 1.076 |  | 0.941 |  |
|  | | | | | | | | | | |
| Note.  Scale consists of items q1, q3, q7, q8, q10, q11, q12, q14, q16, q18, q20, q21, q22, q23, q24, q25, q29, q31, q32 | | | | | | | | | | |
| *Scale Reliability Statistics* | | | | | | | | | | |
|  | | **mean** | | **sd** | | **Cronbach's α** | |
| scale |  | 6.059 |  | 1.352 |  | 0.889 |  |
|  | | | | | | | |
| Note.  Scale consists of items q6R, q9R, q17R, q27R | | | | | | | |

## Reliability Analysis

| **Scale Reliability Statistics** | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | | **mean** | | **sd** | | **Cronbach's α** | |
| scale |  | 6.179 |  | 0.823 |  | 0.827 |  |
|  | | | | | | | |
| Note.  Scale consists of items q2, q5, q13, q19, q26, q28, q30 | | | | | | | |

**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 JASP. Data were screened for multivariate assumptions (normality, linearity, homogeneity, and homoscedasticity), and all assumptions were met with slight problems of heteroscedasticity. Twelve outliers were detected using *z*-scores, and they were removed from further analyses. A small number of missing data values were replaced using the mean replacement. The following EFA analyses were conducted using guidelines outlined in Preacher and MacCallum (2003).

A parallel analysis and scree plot examination suggested three overall factors, and a 3-factor model was tested based on theory. Maximum likelihood estimation was used with direct oblimin rotation because of expected factor correlation. After testing all 32 questions, two items split across several factors (4, 15) 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 .10, 90%CI[.08, .10], while the TLI (.83) indicated room for improvement.

Factor 1 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”. 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 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”. The reliability of all three factors was very high with .95, .86, and .91 for Factors 1, 2, and 3 respectively. The mean scores for each factor were: Factor 1 *M* = 5.11 (*SD* = 1.08), Factor 2 *M* = 6.18 (*SD* = 0.85), and Factor 3 *M* = 6.08 (*SD* = 1.36).

Table 1. *3-Factor Model Loadings.*

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

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