# Data Reliability and Data Reduction

Samantha Estrada PhD

ORS Research Design & Data Analysis Lab Consultant
University of Texas at Tyler

02/25/2022

# Exploratory Factor Analysis

Exploratory factor analysis (EFA) is a statistical technique that is used to reduce data to a smaller set of summary variables and to explore the underlying theoretical structure of the phenomena. It is used to identify the structure of the relationship between the variable and the respondent.

- Exploratory Factor Analysis
- Confirmatory Factor Analysis

## Definitions

### 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 $R^2$ for that item
- EFA = describes the common variance

# Kinds of 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?

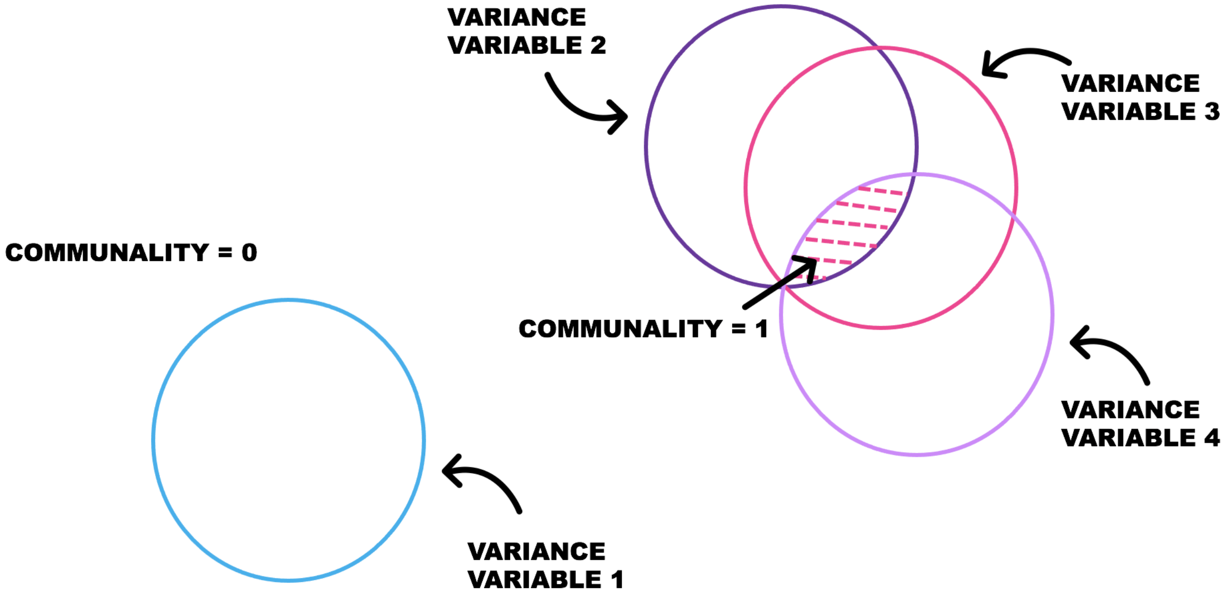


Figure 1: Factor Loadings

# Example: Self-determination Theory

## Data

- Questionnaire on self-determination theory with 32 items and N = 206
- The data file is called `EFA data.csv`.

## Research Question

In self-determination theory, is a theory of motivation that "is concerned with supporting our natural or intrinsic tendencies to behave in effective and healthy ways." Researchers in this area developed a scale for assessing why students are in college was developed.

- In other words, we want to see if the items in that scale fit a three factor structure.

## Items

1. Because I experience pleasure and satisfaction while learning new things.
2. Because I need a degree to get a good job.
3. To show to myself that I can succeed in college studies.
4. Because it's one of the ways that I have chosen to acquire skills in an area which is important to me.
5. Because college provides me with a better understanding of the profession which will be mine.
6. **I wonder what I'm doing in college; actually I find it boring.**
7. For the pleasure I experience while surpassing myself in my studies.
8. For the intense feeling I experience when I am communicating my own ideas to others.
9. **Honestly I don't know, I truly have the impression of wasting my time in college.**
10. Because going to college makes me feel important.
11. Because it allows me to learn about subjects which are of utmost importance to me.
12. Because going to college allows me to continue to learn about a lot of things that interest me.

13. Because I have to go to college in order to complete my degree.
14. For the satisfaction I experience when I am in the process of achieving difficult academic activities.
15. Because, in my opinion, it is a good way to develop skills which will be useful to me later.
16. For the pleasure that I experience when I read interesting authors.
17. **I don't know, I can't understand what I am doing in college.**
18. For the pleasure I experience when I discover new things never seen before.
19. Because it is a prerequisite for getting the job I want.
20. For the pleasure that I experience while I am surpassing myself in one of my personals accomplishments.
21. For the pleasure that I experience when I feel completely absorbed by what certain authors have written.
22. To show to myself that I am intelligent person.
23. For the pleasure that I experience in knowing more about subjects which appeal to me.
24. Because the college experience is very meaningful to me.
25. Because college allows me to experience a personal satisfaction in my quest for excellence in my studies.
26. Because it is one of the ways I have chosen to take responsibility for my future career.
27. **I once had good reasons for registering in college; however, now I wonder whether I should continue.**
28. Because it was the only way to be considered for the career I want.
29. To prove to myself that I am adept in academic endeavors.
30. Because attending college is a good way to prepare myself for my future career.
31. Because attending college is what I really want to do for the time being.
32. For the "high" feeling that I experience while reading on various interesting subjects.

# Data Screening

- I usually screen the data using the `summary()` function in R.
- The output is as large as the number of variables so I am omitting it here.

```r
summary(EFA.data)
```

```r
dim(EFA.data)
```

```
## [1] 206  32
```

## Missing Data

- My next step is usually checking for missing data.
- You have the option to check for % missing, or utilize a type of data imputation method.

```r
## Missing Data
any(is.na(EFA.data))
```

```
## [1] TRUE
```

- I have used the `MICE` package in the past and found it easy to use.
- I will simply remove the data.

```r
nomiss=na.omit(EFA.data)
dim(nomiss)
```

```
## [1] 197  32
```

## Outliers

- I check for outliers using Mahalanobis distance (MD).

- I use the function `mahalanobis()`

```
df = ncol(nomiss)
cutoff = qchisq(0.999, df) ## Cutoff score
mahal = mahalanobis(nomiss,
                    colMeans(nomiss),
                    cov(nomiss) # Calculate Mahalanobis distance
                  )
summary(mahal < cutoff) # Remember FALSE is bad.
```

```
##    Mode   FALSE    TRUE
## logical      12     185
```

We have 12 outliers! I remove them using the following code:

```
##exclude outliers
noout = subset(nomiss, mahal < cutoff)

## Check
dim(noout)
```

```
## [1] 185  32
```

### Sample Size Determination

- Large sample sizes are needed for this 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. This is called crossvalidation.
- 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.

A couple of resources:

- MacCallum et al. (1999) : This paper discusses sample size determination for EFA.
- Kyriazos & others (2018) : Discusses sample size determination for EFA and CFA.

## Assumptions

### Correlation Matrix

- Examine the correlation matrix.
- Examine the determinats of the correlation matrix.
  - If these are positive, very likely you won't have issues later.
  - The package `Hmisc` provides the p-values along with correlations

```
# Use correlation matrix rather than raw data file
corfact = cor(noout)
```

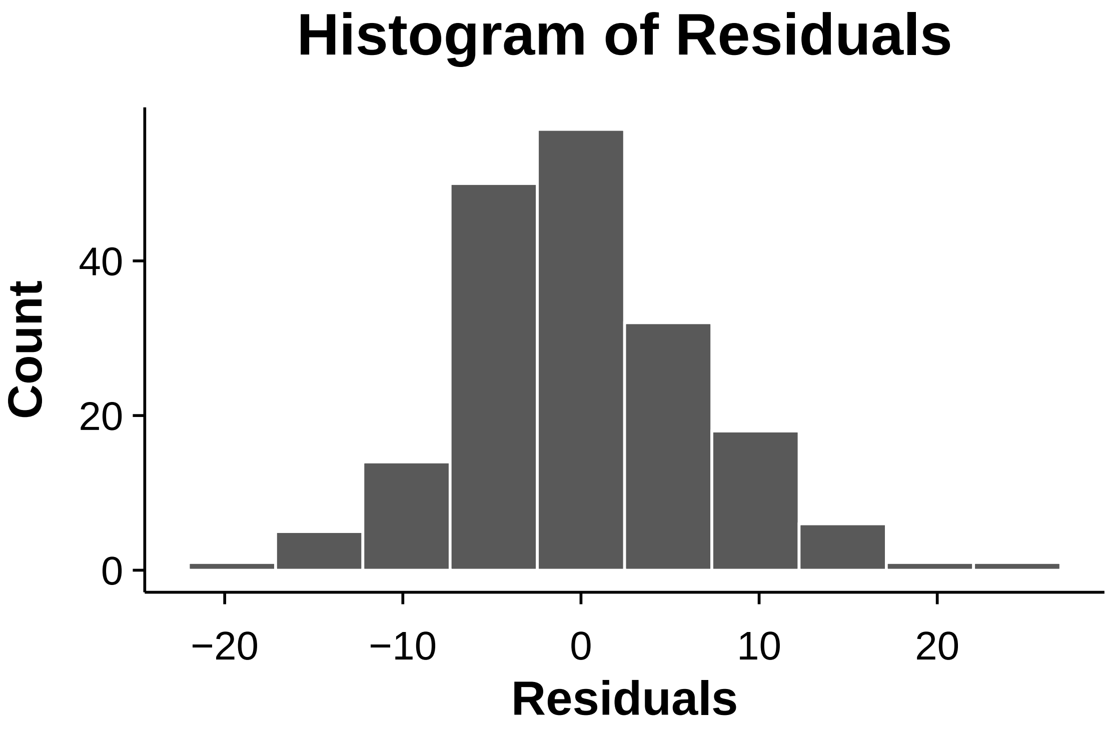
```
#Check that the determinant for the correlation matrix is positive.
det(corfact)
```

```
## [1] 2.84073e-12
```

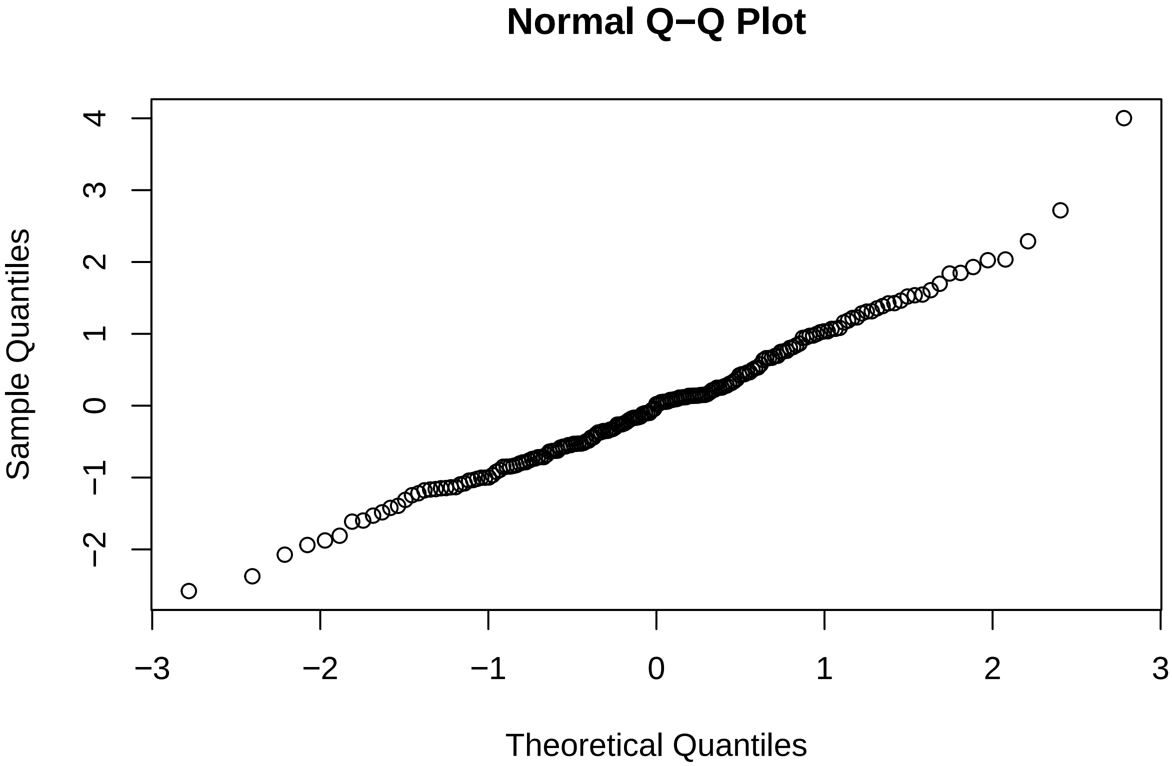
```
# Check the determinant for the covariance matrix is positive
det(cov(noout))
```

```
## [1] 0.003696887
```

### Normality

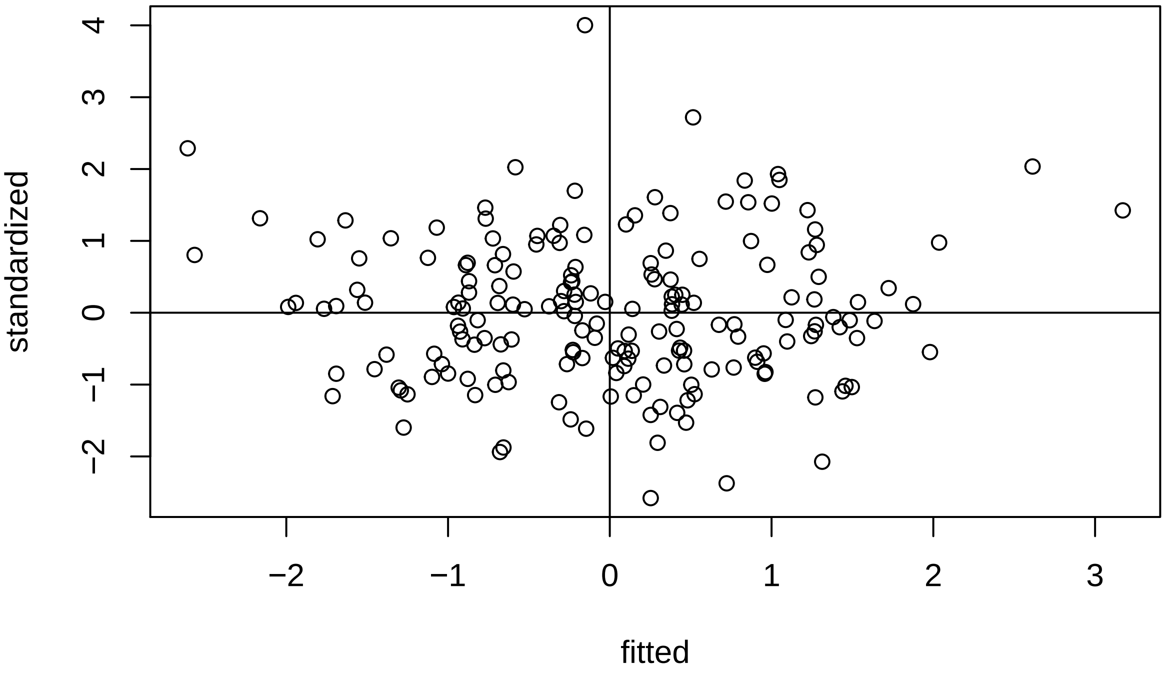


### Linearity (Heteroscedasticity)



### Homogeneity of Variance-Covariance Matrices

We can use the `heplots` package, but we don't have a grouping variable at the moment so I chose a visual method plotting the fitted residuals vs the standardized residuals.
* This can also be accomplished using Box's M test.



## KMO

- Kaiser-Meyer-Olkin Measure
- KMO is a measure of sampling adequacy.
  - In other words, how well suited your data is for EFA.
  - Measure of the proportion of variance among variables that might be common variance
- KMO ranges from 0-1.
- Closer to 1 indicating that the sample is adequate for the EFA.
  - Lower than .6 is a big nope!
- I use the `paf()` function to find KMO.
- The function `summary(paf.corfact)` will have a LOT of output, so it is easier to ask specifically for the KMO value. `summary(paf.pca)$KMO`
- The function `summary(paf.corfact)` can also provide the MSA.

```
# install.packages("rela")
library(rela)
paf.corfact = paf(as.matrix(noout), eigcrit=1, convcrit=.001)
# There's a lot of output, so I used $KMO
summary(paf.corfact)$KMO
```

```
## [1] 0.92304
```

### Write-up

The KMO is 0.92, close to 1, well above the recommended threshold of 0.6. Indicating that we have a sampling adequacy for EFA (or PCA).

### Correlation Adequacy: Bartlett's Test of Sphericity

Bartlett's test of sphericity tests the hypothesis that your correlation matrix is an identity matrix, which would indicate that your variables are unrelated and therefore unsuitable for structure detection.

Ho: The variables are unrelated
Ha: The variables are related

Rejecting the null hypothesis p <.05 indicate that a factor analysis may be useful with your data.

```
# Test significance of the Bartlett test
library(psych)
bt = cortest.bartlett(cor(noout), n = 185)
bt
```

```
## $chisq
## [1] 4586.3
##
## $p.value
## [1] 0
##
## $df
## [1] 496
```

#### Write-up

Bartlett's test yields a ($\chi^2$ (496) = 4586.25, $p$ <.001) indicating that there may be statistically significant interrelationship between variables in our dataset.

- KMO is .923, which is close to 1.0; thus, indicating that the sampling adequacy for factor analysis.

- Bartlett's test is significant. That results implies we adequate data for EFA.

### Do I have a good dataset for EFA?

Checking variables and sample size:
- Number of variables: 32 questions, so we are good.
- Types of variables: 1 to 7 Likert scales, which are at least interval.
- Sample Size: 185, also at least 100, not quite 320 for 10 for each item.

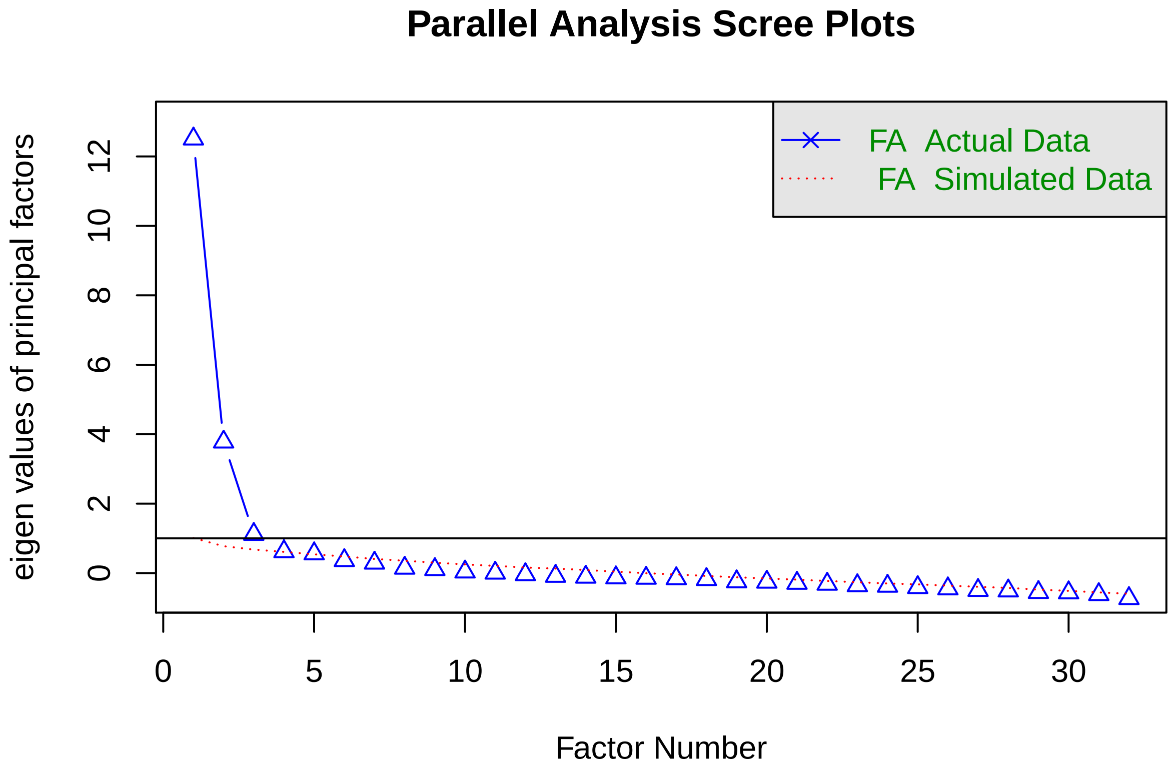
## Data Analysis

### Parallel Analysis

- Parallel Analysis is a procedure used to determine the number of Factors or Principal Components to retain in the initial stage of Exploratory Factor Analysis.
- Based on the Monte Carlo simulation.
  - A data set of random numbers, but having the same sample size and number of variables as the users' research data, are subjected to analysis, and the Eigenvalues obtained are recorded. This is repeated many times (often between 50 and 100 iterations).
- The `fa.parallel()` function in the `psych` package can also produce a scree plot of actual and simulated data based on eigenvalues of the factor analysis

- There's a lot more output here, simplifying due to time.

```
# fm What factor method to use
nofactors = fa.parallel(cor(noout), n.obs=185, fm="ml", fa="fa")
```



```
## Parallel analysis suggests that the number of factors =  5  and the number of components =  NA
```

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

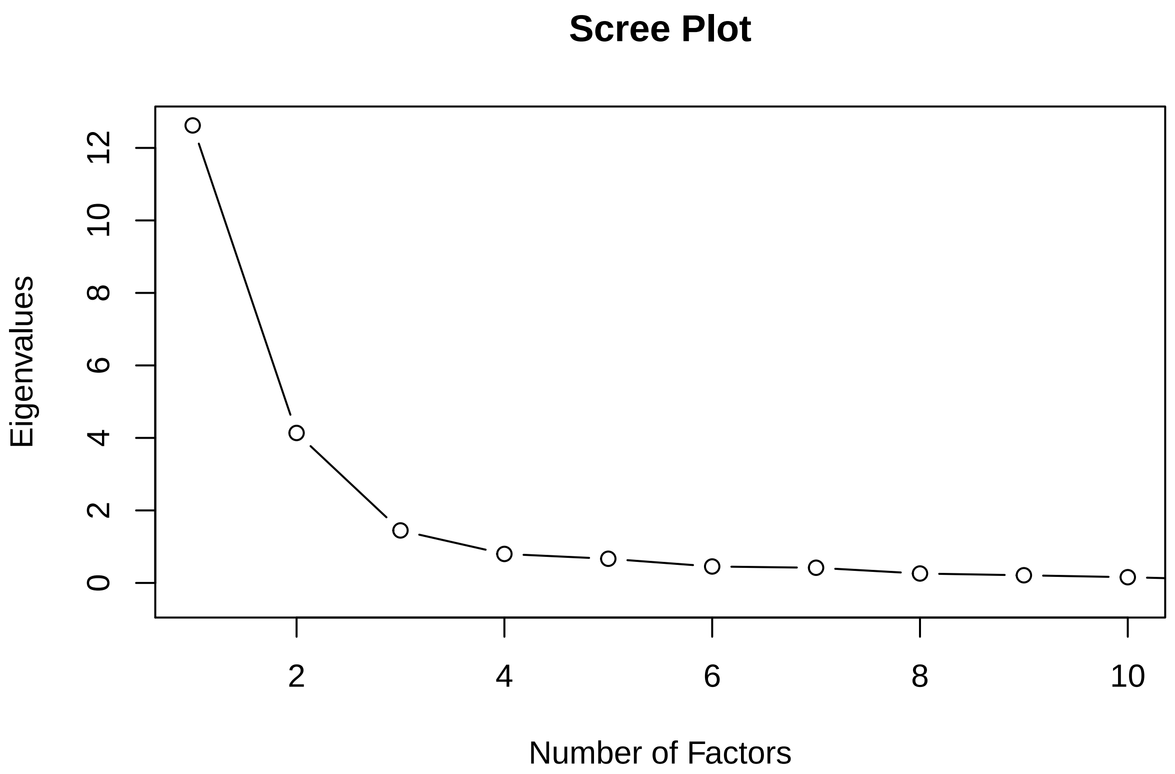
- Theory: self-determination theory suggests three factors.
- Parallel analysis: five factors

```
## Factor Analysis
fa.noout = fa(r = noout, nfactors = 3, rotate = "promax", fm = "ml")
```

### Scree Plot

- A scree plot is often used to find the most desirable number of factors.
- The y-axis is eigenvalues, which essentially stand for the amount of variation.
- An ideal curve should be steep, then bends at an "elbow" — this is your cutting-off point — and after that flattens out.

```
# type b puts the line at the bottom
plot( fa.noout$values, type = 'b', xlim=c(1,10),
      main = "Scree Plot", xlab="Number of Factors",
      ylab="Eigenvalues")
```



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

- Theory: self-determination theory suggests three factors.
- Parallel analysis: five factors
- Scree Plot: three factors

*Note*: There are other measures such as focusing simply on eigenvalues.

## Factor Analysis

We already calculated this when we used the `fa()`, here I am just printing what's "inside" our variable `fa.noout`

### Factor Rotation

Primary Distinction: Orthogonal vs. Oblique

**Orthogonal Rotation**

- Rotated factors remain perpendicular
  - Factors restricted to being uncorrelated
  - Method: Varimax rotation

**Oblique Rotation**

- Rotated factors are not perpendicular
  - Allow factors to be correlate
  - Method: Promax, Oblimin

**Communalities $h_2$**

- $h2$ column indicates that the common or shared variance contributed to the factor structure
- $u2$ column indicates the unique or residual variance.
- We desire that $h2$ values, called commonality estimates, be larger than $u2$ values, called residual estimates.
- Tabachnick & Fidell recommend $>$ .300 for communalities $h2$ (Tabachnick & Fidell, 2019)

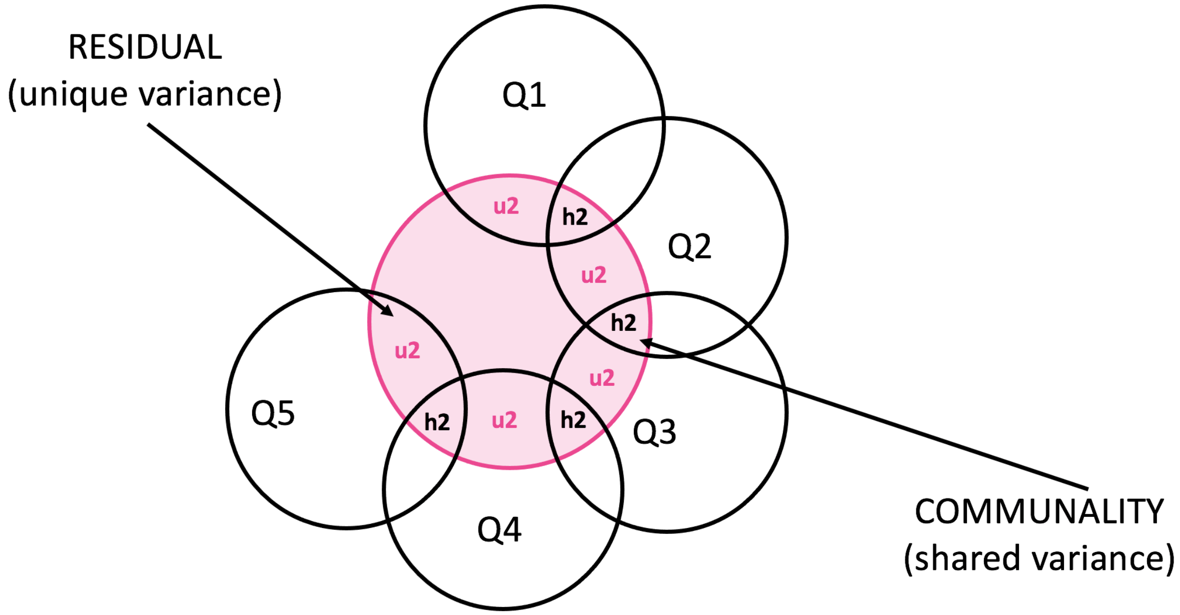


Figure 2: h2 vs u2

```
## Factor Analysis
fa.noout = fa(r = noout, nfactors = 3, rotate = "promax", fm = "ml")
```

```
print(fa.noout$loadings,cutoff = 0.3, sort = TRUE)
```

```
##
## Loadings:
##     ML1    ML3    ML2
## q1   0.535
## q3   0.623
## q7   0.894
## q8   0.822
## q10  0.520
## q14  0.748
## q16  0.905
## q18  0.717
## q20  0.766
## q21  0.928
## q22  0.743
## q23  0.551  0.323
## q25  0.611
## q29  0.620
## q32  0.826
## q2          0.556
## q5          0.565
## q13         1.036
## q19         0.775
## q26         0.713
## q28         0.739
## q30         0.884
## q6                 0.867
## q9                 1.073
## q17                1.069
## q27                0.816
## q4          0.362
## q11  0.479
## q12  0.462  0.311
## q15         0.416
## q24  0.369  0.302
## q31  0.464
##
##                  ML1   ML3   ML2
## SS loadings    9.159 5.264 4.250
## Proportion Var 0.286 0.164 0.133
## Cumulative Var 0.286 0.451 0.584
```

- The proportion of variance explained by each factors ( From `Proportion Var` `0.28 0.16 0.12`) is:
  - Factor 1: 28%
  - Factor 2: 12%
  - Factor 3: 16%
- Item **NA** This item is crossloading in factor 2 and 3 so I would like to drop it from the scale and run the EFA again.

```
#Sample code to drop items
newdat =  noout[,-c(24)]
# OR
```

```
fa.noout = fa(noout[,-c(24)], nfactors = 3, rotate = "promax", fm = "ml")
# This print option will hide anything less than .300
# ALSO sort
print(fa.noout$loadings,cutoff = 0.3, sort = TRUE)
```

```
##
## Loadings:
##      ML1    ML3    ML2
## q1   0.539
## q3   0.622
## q7   0.888
## q8   0.815
## q10  0.517
## q14  0.748
## q16  0.894
## q18  0.717
## q20  0.765
## q21  0.918
## q22  0.739
## q23  0.558  0.311
## q25  0.616
## q29  0.619
## q32  0.817
## q2          0.551
## q5          0.555
## q13         1.027
## q19         0.778
## q26         0.694
## q28         0.734
## q30         0.869
## q6                 0.860
## q9                 1.065
## q17                1.057
## q27                0.810
## q4          0.359
## q11  0.484
## q12  0.469  0.303
## q15         0.407
## q31  0.466
##
##                  ML1   ML3   ML2
## SS loadings    8.961 5.023 4.180
## Proportion Var 0.289 0.162 0.135
## Cumulative Var 0.289 0.451 0.586
```

## Internal Consistency Reliability

We will use Cronbach's $\alpha$ (alpha) to measure internal consistency reliability.

### Cronbach's $\alpha$

- Cronbach's $\alpha$ (alpha) is a widely-used measure of reliability used to quantify the amount of random measurement error that exists in a multi-item measurement scale as estimated by the sum of the score or the average.
- Ranges from 0-1. Closer to 1 indicates better reliability.
- Tends to underestimate reliability.
- Other options MacDonald's $\omega$ and/or Guttman's $\lambda$

```
##reliability
# I excluded item 24
factor1 = c(1, 3, 7, 8, 10:12, 14, 16, 18, 20:23, 25, 29, 31, 32)
factor2 = c(2, 5, 13, 19, 26, 28, 30)
factor3 = c(4, 6, 9, 15, 17, 27)
psych::alpha(noout[ , factor1])$total
```

```
##  raw_alpha std.alpha G6(smc) average_r    S/N       ase   mean    sd median_r
##    0.94428   0.94722 0.96174   0.49925 17.946 0.0059664 5.0727 1.087   0.49746
```

```
psych::alpha(noout[ , factor2])$total$raw_alpha
```

```
## [1] 0.86629
```

```
# ITEM 6,8,17,27 are reverse coded.
psych::alpha(noout[ , factor3], check.keys= TRUE)$total$raw_alpha
```

```
## [1] 0.8655
```

## Factor Descriptives

- Find mean and standard deviations for each item. Report them.
- Also customary to report the correlations between the factors.

```
##      f1    f2    f3
## f1 1.00  0.49  0.08
## f2 0.49  1.00 -0.22
## f3 0.08 -0.22  1.00
##
## n= 185
##
##
## P
##    f1     f2     f3
## f1        0.0000 0.2847
## f2 0.0000        0.0022
## f3 0.2847 0.0022
```

## Naming the factors

| Factor 1 | Factor 2 | Factor 3 |
|---|---|---|
| 1. Because I experience pleasure and satisfaction while learning new things. | 2. Because I need a degree to get a good job. | 4. Because it's one of the ways that I have chosen to acquire skills in an area which is important to me. |

| Factor 1 | Factor 2 | Factor 3 |
|---|---|---|
| 3. To show to myself that I can succeed in college studies. | 5. Because college provides me with a better understanding of the profession which will be mine. | 6. I wonder what I'm doing in college; actually I find it boring. |
| 7. For the pleasure I experience while surpassing myself in my studies. | NA | 9. Honestly I don't know, I truly have the impression of wasting my time in college. |
| 8. For the intense feeling I experience when I am communicating my own ideas to others. | NA | NA |
| 10. Because going to college makes me feel important. | NA | NA |
| NA | NA | NA |
| NA | NA | |
| NA | | |
| NA | | |
| NA | | |
| NA | | |
| NA | | |
| NA | | |
| NA | | |
| NA | | |
| NA | | |
| NA | | |

- Factor 1: Measured the intrinsic motivation
- Factor 2: Career goals
- Factor 3: Doubt about motivation for college

## Write-up

- Report what type of analysis you conducted.
- Report what software and version you used.

> An exploratory factor analysis (EFA) was used to analyze the underlying factors in the self-determination motivation for college questionnaire using the psych package in R (Revelle & Revelle (2015); R Core Team, 2019).

- Report what type of data screening you conducted.

> Data were screened for multivariate assumptions (normality, linearity, homogeneity, and homoscedasticity), and all assumptions were met with slight problems of heteroscedasticity. Twelve multivariate outliers were detected using Mahalanobis distance ($\chi^2$ (32) = 62.49), and they were removed from further analyses. Bartlett's test indicated correlation adequacy, ($\chi^2$ (496) = 4586.25, $p$ <.001), and the KMO test indicated sampling adequacy, $KMO = 0.92$.

- It is always good to report on the parallel analysis and scree plot.
  - Do not worry that they don't all have the same answer (e. g. 3 factors vs. 5 factors). This is common!

  - Decide on the number of factors based on theory!
- Report the type of rotation you used.
  - Report if you tested multiple.

> 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 promax rotation because of expected factor correlation. [*if we had removed items we would explain here* ] After testing all 32 questions, one item crossloaded on several factor according to the criterion that the loadings must be greater than .300, Item "NA" (Tabachnick & Fidell, 2019).

- Go into details regarding the factors
  - Number of items. Give examples if needed.
  - Report means and standard deviation of the items
  - This is a good place to present reliability estimates.

> 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". Internal consistency reliability of the scores was assessed through Cronbachs $\alpha$ 0.94, 0.87, and .0.87 for Factors 1, 2, and 3 respectively. [*Include* $\omega$ or $\lambda2$ ] The mean scores for each factor were: Factor 1 $M$ = 5.07 ($SD$ = 1.09), Factor 2 $M$ = 6.16($SD$ = 6.16), and Factor 3 $M$ = 3.29 ($SD$ = 3.29).

### EFA Resources

- Page 218 of the APA Manual (*Publication Manual of the American Psychological Association*, 2020) has a sample EFA table.
- Osborne (2015) : Indepth discussion on what rotating means for EFA
- Costello & Osborne (2005) : Best practices for EFA

## Resources

- Research Design & Data Analysis Lab: https://www.uttyler.edu/research/ors-research-design-data-analysis-lab/
- Schedule a consultant appointment with me: https://www.uttyler.edu/research/ors-research-design-data-analysis-lab/ors-research-design-data-analysis-lab-consultants/
- Check out Lab Resources (including recording of this webinar): https://www.uttyler.edu/research/ors-research-design-data-analysis-lab/resources/

## References

Costello, A. B., & Osborne, J. (2005). Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Practical Assessment, Research, and Evaluation*, *10*(1), 7.

Kyriazos, T. A., & others. (2018). Applied psychometrics: Sample size and sample power considerations in factor analysis (efa, cfa) and sem in general. *Psychology*, *9*(08), 2207.

MacCallum, R. C., Widaman, K. F., Zhang, S., & Hong, S. (1999). Sample size in factor analysis. *Psychological Methods*, *4*(1), 84.

Osborne, J. W. (2015). What is rotating in exploratory factor analysis? *Practical Assessment, Research, and Evaluation*, *20*(1), 2.

*Publication manual of the american psychological association.* (2020). American Psychological Association. https://doi.org/10.1037/0000165-000

Revelle, W., & Revelle, M. W. (2015). Package “psych”. *The Comprehensive R Archive Network.*

Tabachnick, B. G., & Fidell, L. S. (2019). *Using multivariate statistics.* Pearson.