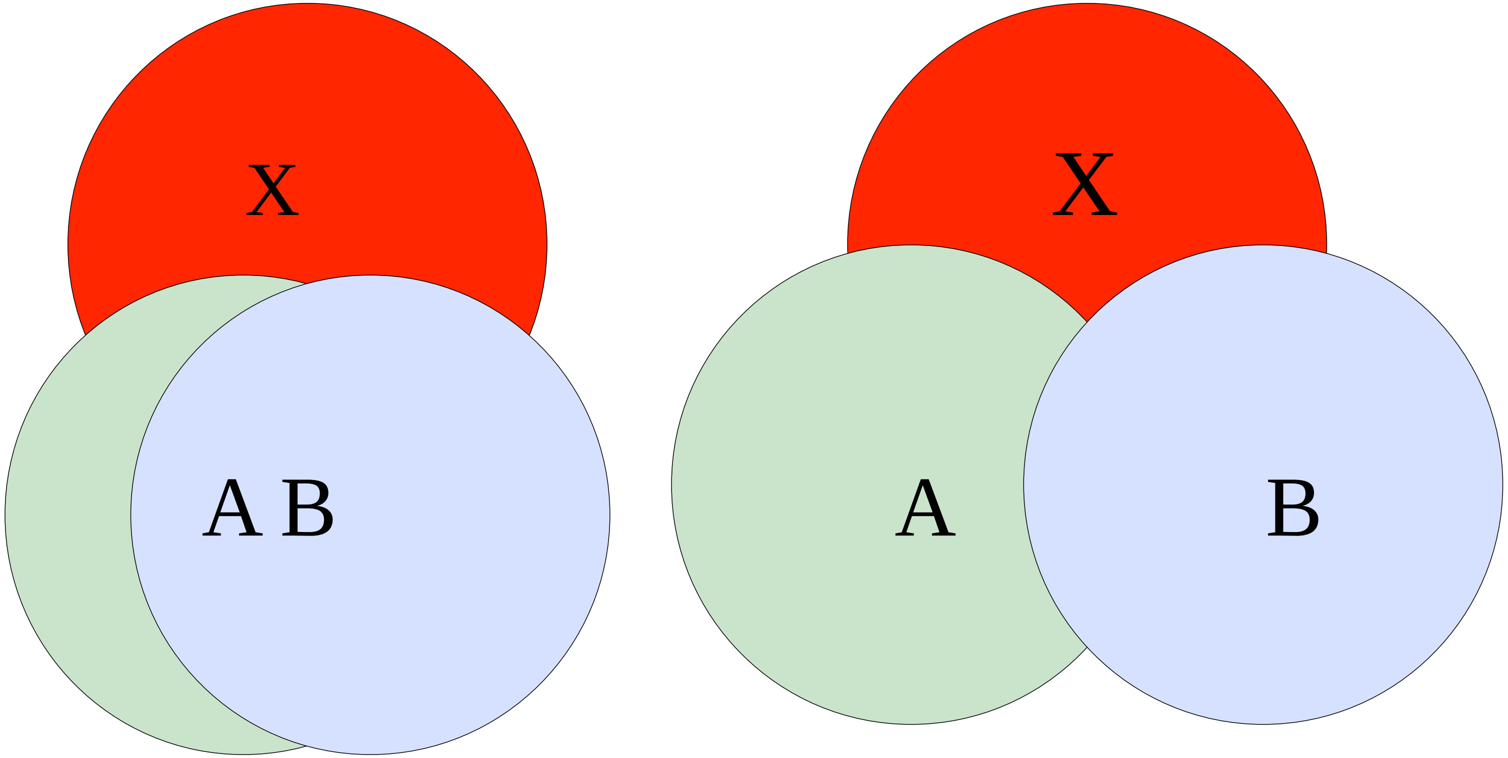
# Multicollinearity
# Data Summarization
# Data Reduction

# Multicollinearity




- high degree of correlation amongst IVs
  - ex: height and weight, household income and water consumption, mileage and price of a car

# Multicollinearity in IVs

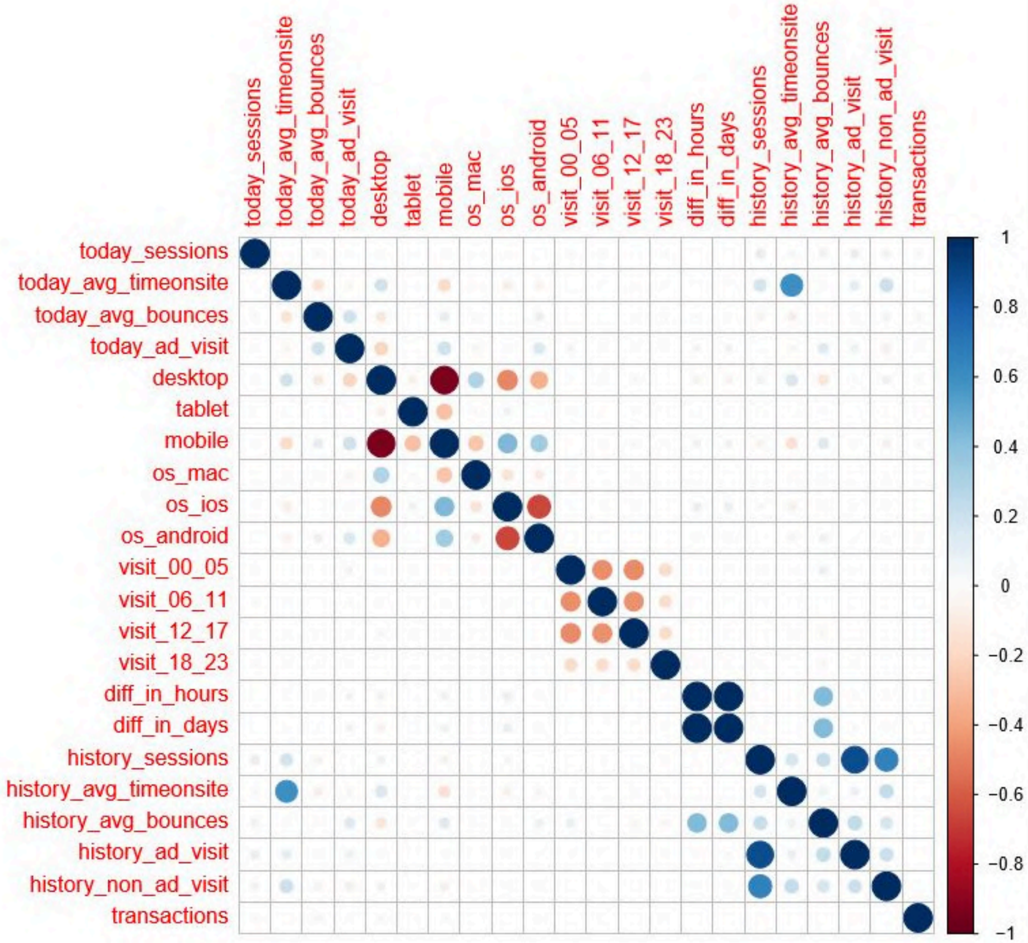
- causes unwanted effects
  - saps statistical power of the analysis
  - can cause switch in signs of the coefficients (in regression), overestimate standard errors, reduced precision in estimating the coefficients' effects, etc…
  - will result in less reliable statistical inferences
- higher number of IV –> increase in sample size required
- what can you do?
  - removing highly correlated IVs/features/items/predictors
  - combine them/uncover latent dimensions

# Multicollinearity

- Some Solutions:
    - **Feature or Variable Selection**
    - **Reduce by Combining Variables**
- choice depends upon
    - research inquiry
    - interpretability

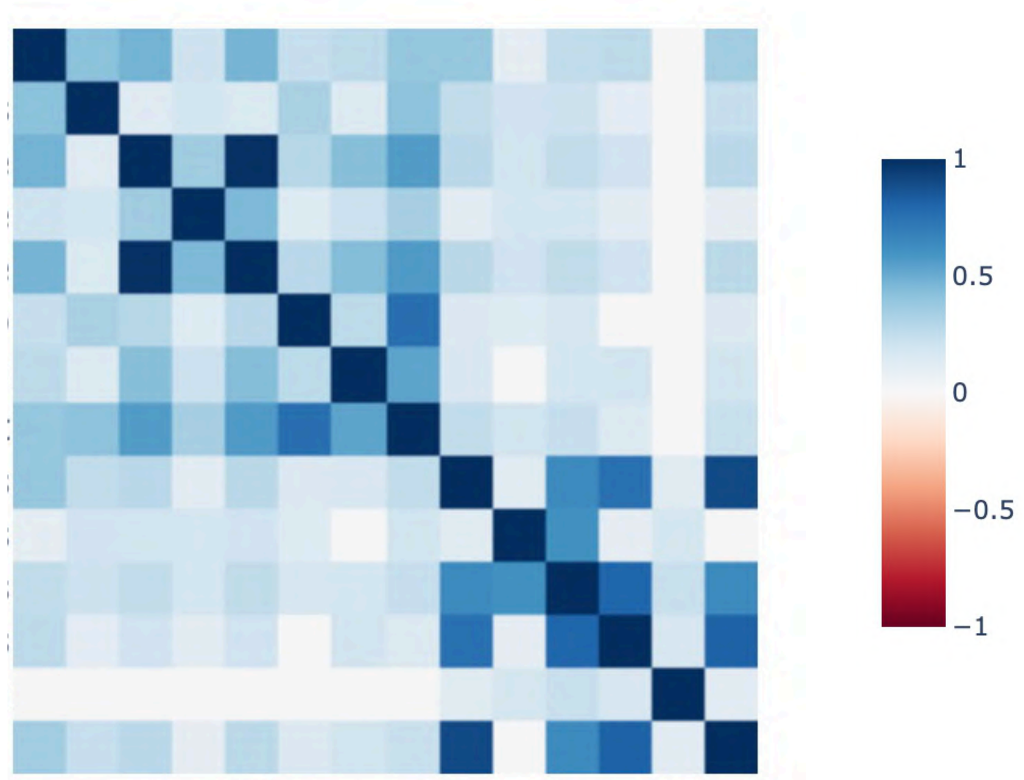
# Feature or Variable Selection

- **Correlation:** helps identify collinear variables




http://stackbug.com/google-analytics/identify-deal-multicollinearity-models/

# Is there collinearity?



# Feature or Variable Selection

- **Variance Inflation Factor (VIF)**
  - The R-square term tells us
    - how predictable one IV is from the set of other IVs
    - 1 = not correlated.
    - Between 1 and 5 = moderately correlated.
    - Greater than 5 = highly correlated.

$$VIF_1 = \frac{1}{1 - R^2_{1.2\ldots k}}$$

EXAMPLE

# Feature or Variable Selection

| | Gender | Age | Years of service | Education level | Salary |
|---|---|---|---|---|---|
| 0 | 0.0 | 27.0 | 1.7 | 0.0 | 39343.0 |
| 1 | 1.0 | 26.0 | 1.1 | 1.0 | 43205.0 |
| 2 | 1.0 | 26.0 | 1.2 | 0.0 | 47731.0 |
| 3 | 0.0 | 27.0 | 1.6 | 1.0 | 46525.0 |
| 4 | 0.0 | 26.0 | 1.5 | 1.0 | 40891.0 |

| | variables | VIF |
|---|---|---|
| 0 | Gender | 2.207155 |
| 1 | Age | 13.706320 |
| 2 | Years of service | 10.299486 |
| 3 | Education level | 2.409263 |

| | variables | VIF |
|---|---|---|
| 0 | Gender | 1.863482 |
| 1 | Years of service | 2.478640 |
| 2 | Education level | 2.196539 |

| | variables | VIF |
|---|---|---|
| 0 | Gender | 2.168068 |
| 1 | Education level | 2.407695 |
| 2 | Age_at_joining | 3.326991 |

(Age - Years of service)

https://www.analyticsvidhya.com/blog/2020/03/what-is-multicollinearity/

# Feature or Variable Selection

- **Squared Multiple Correlation (SMC)**
    - represent the maximal proportion of variance in each variable that can be explained by a linear combination of other variables (Harris, 2001).
    - inversely related to the *uniqueness* of the variable

$$SMC = 1 - (1/diag(1/R))$$

VIF

EXAMPLE

# Feature or Variable Selection

**37 participants x 100 audio examples**

| Bipolar Scales |
|---|
| Soft-Hard |
| Colorless-Colorful |
| Heavy-Light |
| Warm-Cold |
| Dark-Bright |
| Acoustic-Synthetic |
| Gentle-Harsh |
| Strong-Weak |
| Empty-Full |
| High Energy-Low Energy |

**10 scales x 100** (mean) **ratings**

EXAMPLE

# Feature or Variable Selection

| Bipolar Scales | Squared Multiple Correlation |
|---|---|
| Soft-Hard | .88 |
| Colorless-Colorful | .75 |
| Heavy-Light | .90 |
| Warm-Cold | .75 |
| Dark-Bright | .78 |
| Acoustic-Synthetic | .72 |
| Gentle-Harsh | .86 |
| Strong-Weak | .71 |
| Empty-Full | .51 |
| High Energy-Low Energy | .69 |

| | Mean Inter-Subject *r* | Cronbach Alpha |
|---|---|---|
| Soft-Hard | .50 | .88 |
| Heavy-Light | .41 | .83 |
| Warm-Cold | .28 | .72 |
| Dark-Bright | .32 | .76 |
| Acoustic-Synthetic | .55 | .90 |
| Gentle-Harsh | .52 | .87 |

# Multicollinearity

- **Solutions:**
  - *Feature or Variable Selection*
  - $\boxed{\textit{Reduce by Combining Variables}}$
- choice depends upon
  - research inquiry
  - interpretability vs model performance

# Feature Set Reduction

- Why?
  - increase in dimensions -> complex data -> harder to interpret
  - additional variables = additional processing time and space
  - avoid curse of dimensionality -> amount of data needed to support the result often grows exponentially with the dimensionality
  - reduce overfitting
  - help eliminate irrelevant features
  - easier visualisation

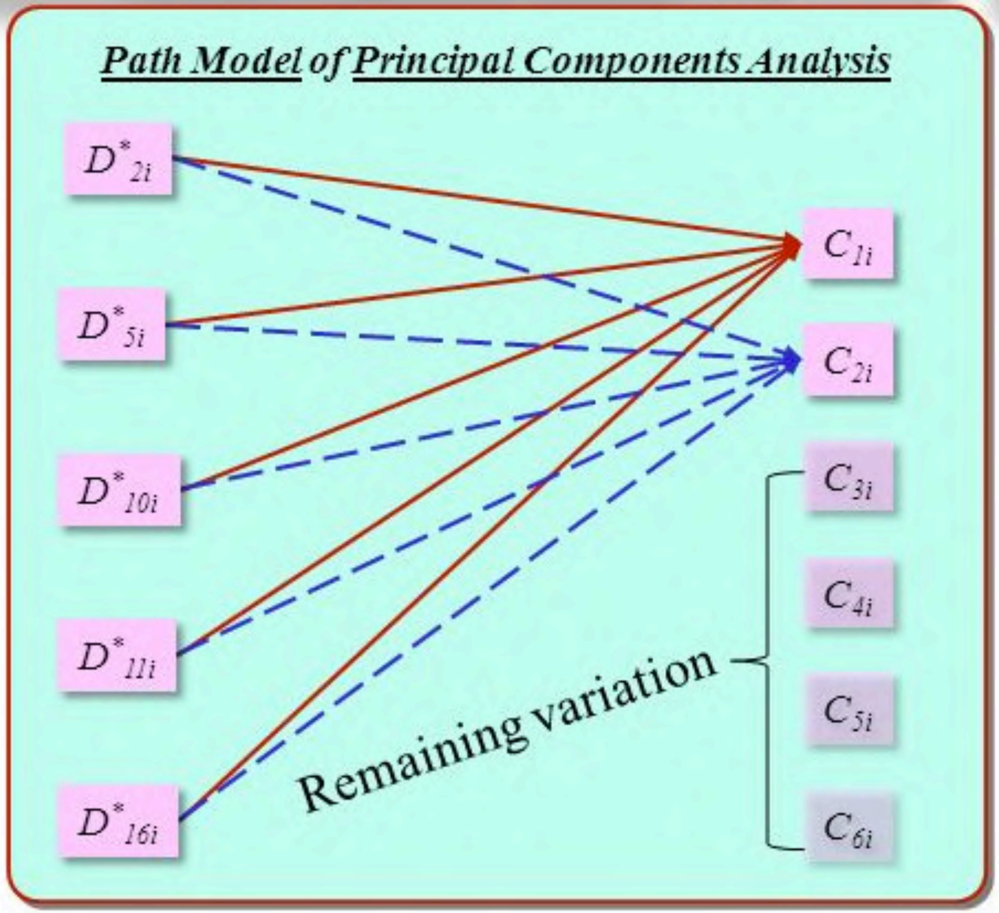
## *Research Question?*

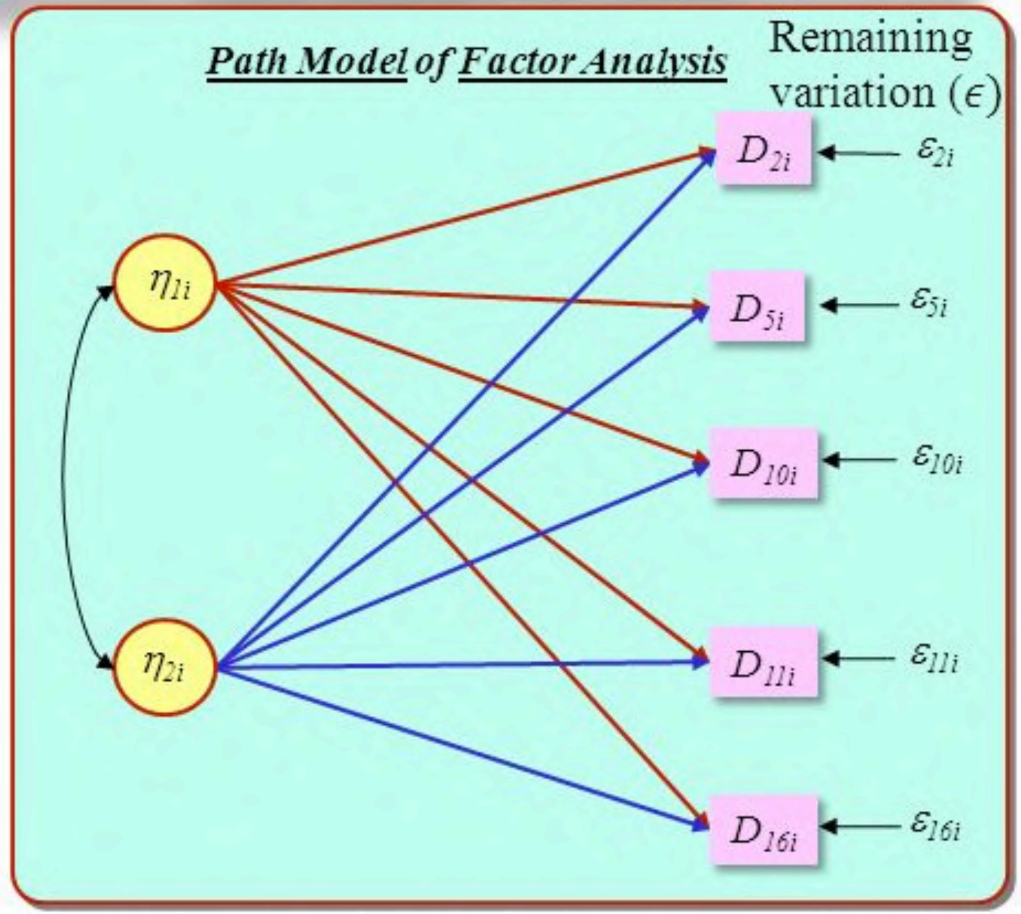
*Rather than asking … **"Can We <u>Forge These Several Indicators</u> Together Into A <u>Smaller Number Of Composites</u> With Defined Statistical Properties?"***

Then, we would need …
***<u>Principal Components Analysis</u> (PCA)***

*We could ask … **"Are There A Number Of <u>Unseen (Latent) Factors (Constructs)</u> Acting "<u>Beneath</u>" These Indicators To Forge Their Observed Values?"***

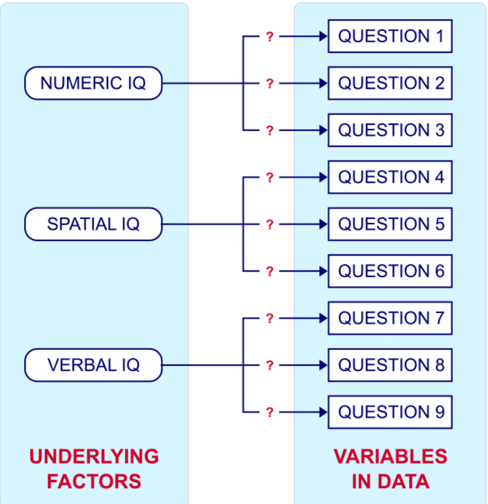
Instead, we would need …
***<u>Factor Analysis</u>** (**CFA** or **EFA?**)*



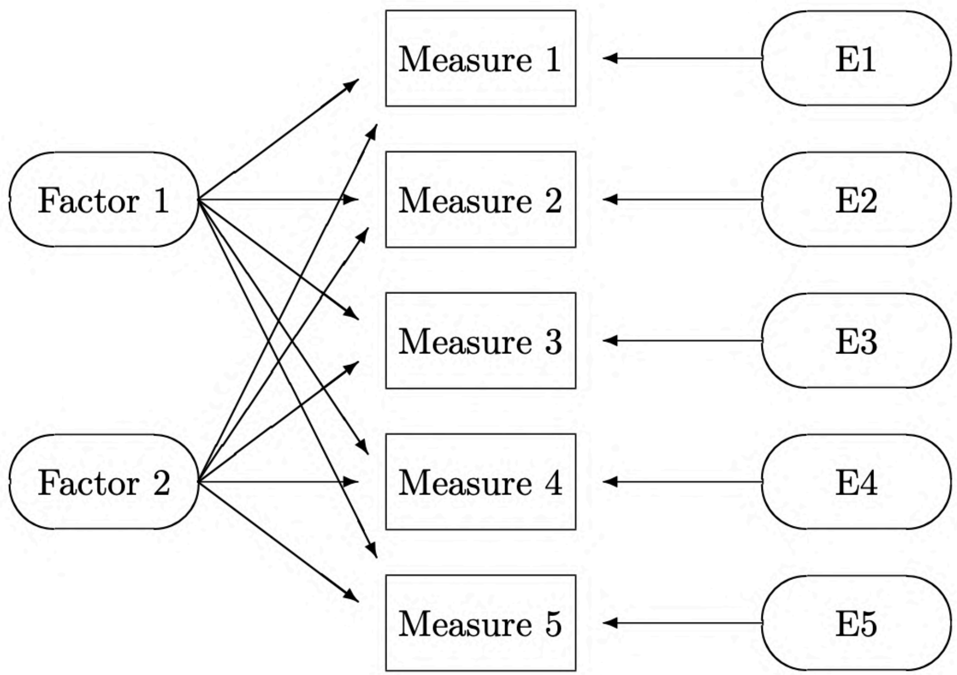
# Factor Analysis

- idea–> there are underlying “latent” variables or “factors”, and several variables might be measures of the same factor
- underlying/latent dimensions are not directly observable
- hidden constructs/factors give rise to observed variables



# Factor Analysis

- condense information into **factors** with minimum information loss
- predetermined no. of factors (intrinsic dimensionality estimation)



# Factor Analysis

$$
\begin{aligned}
X_1 &= \mu_1 + l_{11} f_1 + l_{12} f_2 + \cdots + l_{1m} f_m + \epsilon_1 \\
X_2 &= \mu_2 + l_{21} f_1 + l_{22} f_2 + \cdots + l_{2m} f_m + \epsilon_2 \\
&\vdots \\
X_p &= \mu_p + l_{p1} f_1 + l_{p2} f_2 + \cdots + l_{pm} f_m + \epsilon_p
\end{aligned}
$$

$$
\mathbf{L} = \begin{pmatrix} l_{11} & l_{12} & \dots & l_{1m} \\ l_{21} & l_{22} & \dots & l_{2m} \\ \vdots & \vdots & & \vdots \\ l_{p1} & l_{p2} & \dots & l_{pm} \end{pmatrix} = \text{matrix of factor loadings}
$$

$$
\epsilon = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_p \end{pmatrix} = \text{vector of specific factors}
$$

*error terms, what the Factors cannot explain in each variable*

$$
\mathbf{X} = \mu + \mathbf{L}\mathbf{f} + \epsilon
$$

# Before Performing Factor Analysis .....

- **Bartlett's Test of Sphericity** compares an observed correlation matrix to the identity matrix to check for redundancy between variables
- $H_0$ -> variables are orthogonal, i.e. not correlated.
- $H_A$ -> variables are not orthogonal, i.e. they are correlated enough to where the correlation matrix diverges significantly from the identity matrix
- returns chi-square statistic and *p*-value

# Before Performing Factor Analysis …..

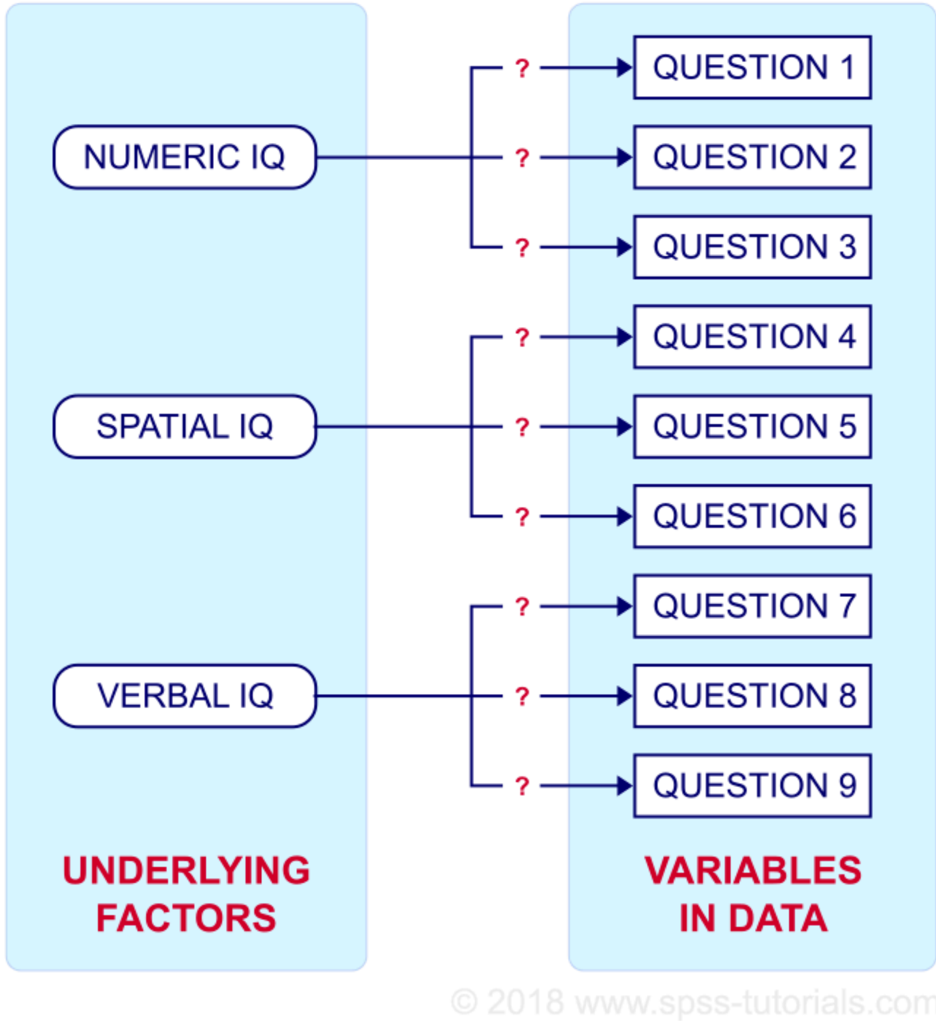
- **Kaiser-Meyer-Olkin** (KMO) Test for Sampling Adequacy: examine the strength of the partial correlation between the variables
- low KMO value indicative that variables are not sufficiently related to each other
- at least 10/15 cases per variable

$$KMO = \frac{\sum\sum_{j \neq k} r_{jk}^2}{\sum\sum_{j \neq k} r_{jk}^2 + \sum\sum_{j \neq k} p_{jk}^2}$$

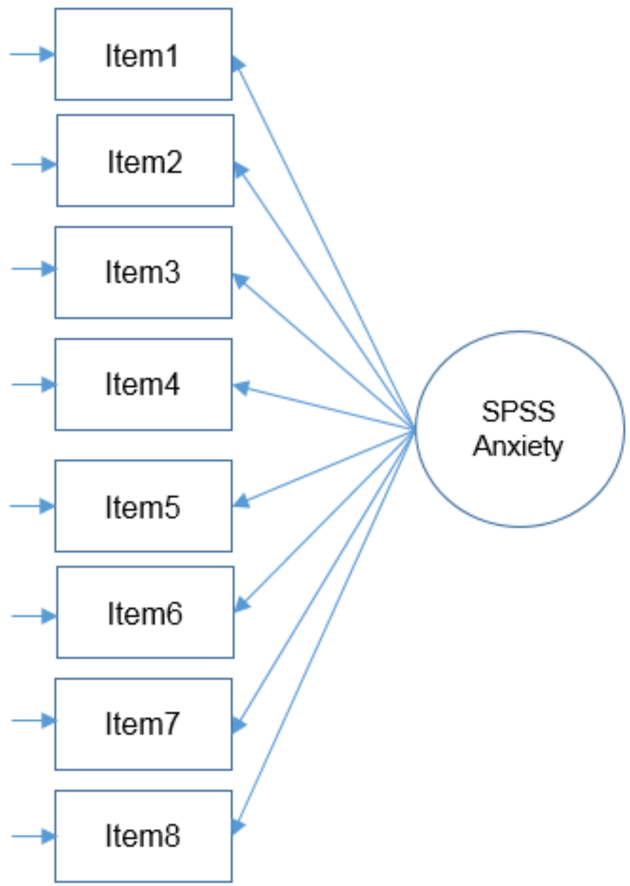
| KMO Value | Degree of Common Variance |
|---|---|
| 0.91 and above | Superb |
| 0.81 - 0.9 | Great |
| 0.71 - 0.8 | Good |
| 0.5 - 0.7 | Mediocre |
| 0.49 and below | Don't factor |

# Factor Analysis Types

- **R-Type** (commonly used)
    - covariation or correlation between variables




1. Statistics makes me cry
2. My friends will think I'm stupid for not being able to cope with SPSS
3. Standard deviations excite me
4. I dream that Pearson is attacking me with correlation coefficients
5. I don't understand statistics
6. I have little experience with computers
7. All computers hate me
8. I have never been good at mathematics




Do all these items actually measure what we call "SPSS Anxiety"?

EXAMPLE

| | Statistics makes me cry | My friends will think I'm stupid for not being able to cope with SPSS | Standard deviations excite me | I dream that Pearson is attacking me with correlation coefficients | I don't understand statistics | I have little experience with computers | All computers hate me | I have never been good at mathematics |
|---|---|---|---|---|---|---|---|---|
| Statistics makes me cry | 1 | | | | | | | |
| My friends will think I'm stupid for not being able to cope with SPSS | -.099 | 1 | | | | | | |
| Standard deviations excite me | -.337 | .318 | 1 | | | | | |
| I dream that Pearson is attacking me with correlation coefficients | .436 | -.112 | -.380 | 1 | | | | |
| I don't understand statistics | .402 | -.119 | -.310 | .401 | 1 | | | |
| I have little experience with computers | .217 | -.074 | -.227 | .278 | .257 | 1 | | |
| All computers hate me | .305 | -.159 | **-.382** | .409 | .339 | **.514** | 1 | |
| I have never been good at mathematics | .331 | -.050 | -.259 | .349 | .269 | .223 | .297 | 1 |

Inter-scale/item correlation

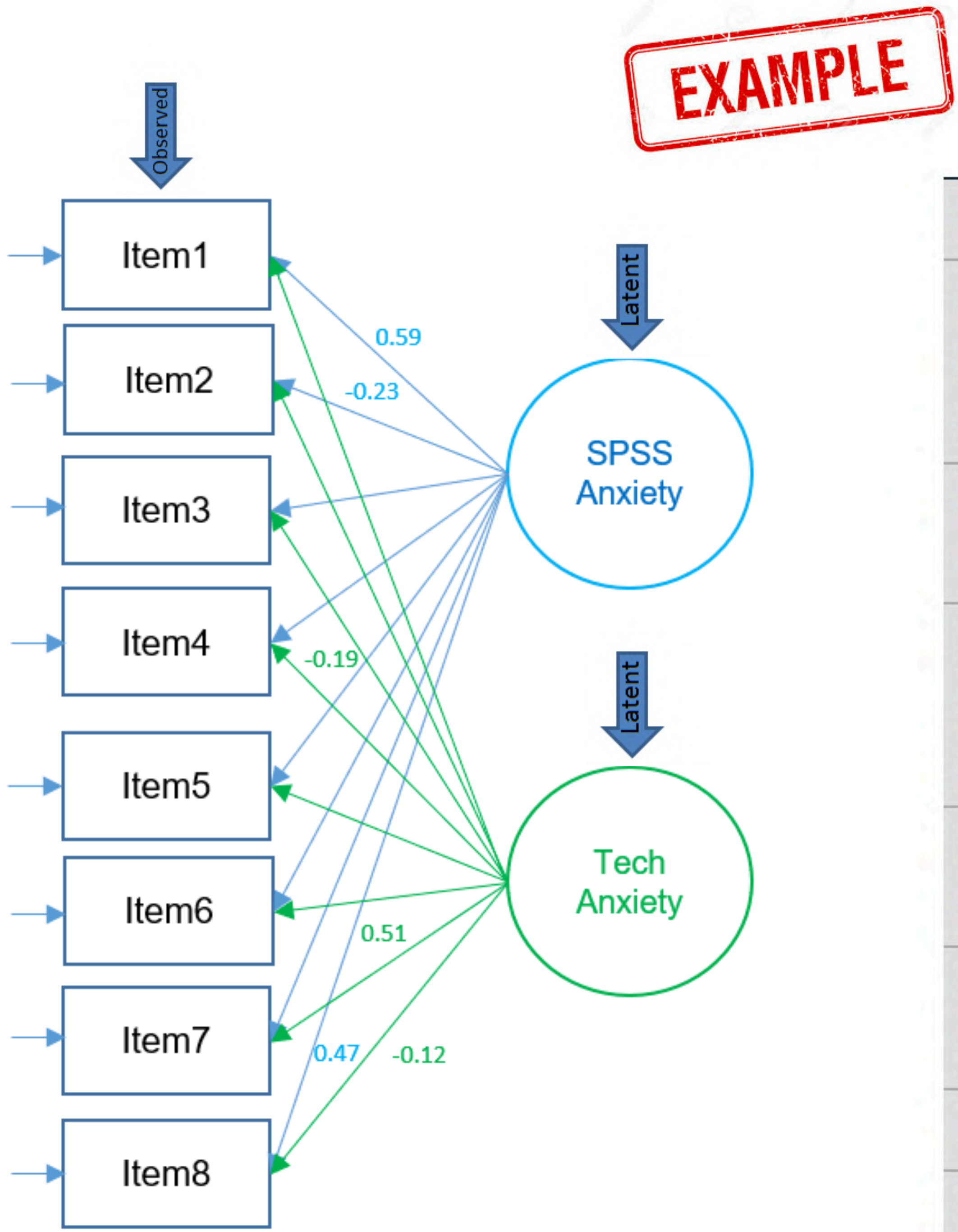
EXAMPLE

1. Statistics makes me cry
2. My friends will think I'm stupid for not being able to cope with SPSS
3. Standard deviations excite me
4. I dream that Pearson is attacking me with correlation coefficients
5. I don't understand statistics
6. I have little experience with computers
7. All computers hate me
8. I have never been good at mathematics

**Factor Loadings**: the weight of the factor in predicting the variable/correlations between variables and factors

**Factor Matrix[a]**

| | Factor 1 | Factor 2 |
|---|---|---|
| Statistics makes me cry | .588 | -.303 |
| My friends will think I'm stupid for not being able to cope with SPSS | -.227 | .020 |
| Standard deviations excite me | -.557 | .094 |
| I dream that Pearson is attacking me with correlation coefficients | .652 | -.189 |
| I don't understand statistics | .560 | -.174 |
| I have little experience of computers | .498 | .247 |
| All computers hate me | .771 | .506 |
| I have never been good at mathematics | .470 | -.124 |




Note: only selected loadings shown

**Factor Matrix[a]**

| | Factor 1 | Factor 2 |
|---|---|---|
| Statistics makes me cry | .588 | -.303 |
| My friends will think I'm stupid for not being able to cope with SPSS | -.227 | .020 |
| Standard deviations excite me | -.557 | .094 |
| I dream that Pearson is attacking me with correlation coefficients | .652 | -.189 |
| I don't understand statistics | .560 | -.174 |
| I have little experience of computers | .498 | .247 |
| All computers hate me | .771 | .506 |
| I have never been good at mathematics | .470 | -.124 |

EXAMPLE

# Factor Interpretation

- **R-Type**

| Bipolar Scales |
|---|
| Soft-Hard |
| Colorless-Colorful |
| Warm-Cold |
| Dark-Bright |
| Strong-Weak |
| Empty-Full |
| High Energy-Low Energy |

**F1**: Energy/Activity **F2** : Brightness **F3**: Fullness

| | Factor 1 Variance Explained (41%) | Factor 2 Variance Explained (33%) | Factor 3 Variance Explained (15%) |
|---|---|---|---|
| Colorless-Colorful | –.07 | .94 | .31 |
| Warm-Cold | .59 | –.58 | –.34 |
| Dark-Bright | .17 | .86 | .07 |
| Acoustic-Synthetic | .43 | –.67 | –.13 |
| Soft-Hard | .96 | –.18 | –.03 |
| Strong-Weak | –.91 | .06 | –.27 |
| Empty-Full | .33 | .36 | .87 |
| High Energy-Low Energy | –.90 | –.16 | –.33 |

EXAMPLE

# Factor Interpretation

**F1**: customer experience post boarding

**F2**: airline booking experience and related perks

**F3**: flight competitive advantage of the airline compared to its competition

| | Factor 1 | Factor 2 | Factor 3 |
|---|---|---|---|
| Great hospitality | 0.98 | -0.04 | 0.02 |
| Flight is on time | 0.95 | -0.01 | 0.18 |
| Great Food | 0.92 | 0.04 | -0.05 |
| Friendly atmosphere | 0.62 | 0.17 | -0.33 |
| Frequent flyer program | -0.03 | 0.97 | -0.01 |
| Flights are economic | -0.02 | 0.96 | 0.09 |
| No hassles in boarding | -0.07 | 0.95 | 0.09 |
| Good flight times | -0.09 | 0.19 | 0.96 |
| Seats are comfortable | 0.03 | 0.09 | 0.95 |
| Loyalty or attachment | -0.19 | -0.42 | -0.09 |

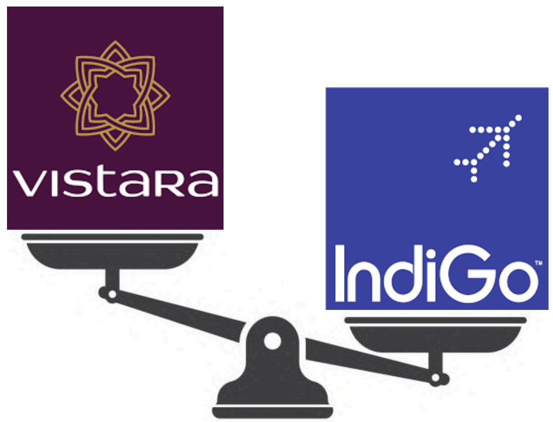
ex: factor loadings for an airlines survey

# Factor Scores

- composite scores represented by the latent variable which can be used in subsequent statistical analyses (ex: multiple regression, t-tests, etc.)

**F1**: customer experience post boarding


**F2**: airline booking experience and related perks



# Terminology

- **Communalities ($h^2$)**
  - proportion of each variable's variance that can be explained by the Factors
  - ex: CAD -> $0.889^2+0.106^2+(-0.108^2)$

| Variables | Factor 1 | Factor 2 | Factor 3 | Communalities |
|---|---|---|---|---|
| CAD | 0.889 | 0.106 | −0.108 | 0.813 |
| CTOPP–Elision | −0.086 | 0.619 | 0.022 | 0.391 |
| CTOPP–Nonword Repetition | 0.117 | 0.335 | 0.017 | 0.126 |
| PRF | 0.040 | 0.226 | 0.654 | 0.480 |
| RAN | 0.062 | 0.134 | −0.788 | 0.643 |
| Spelling Fluency | 1.018 | −0.085 | −0.007 | 1.000 |
| TOWRE–SWE | −0.043 | 0.117 | 0.891 | 0.809 |
| TOWRE–PDE | 0.002 | 0.632 | 0.398 | 0.558 |
| TSWRF | 0.682 | 0.016 | 0.172 | 0.495 |
| WIF | 0.047 | 0.209 | 0.762 | 0.627 |
| WJ III–Word Attack | −0.019 | 0.929 | −0.047 | 0.866 |
| WJ III–Letter-word Identification | 0.035 | 0.780 | 0.144 | 0.630 |

anomaly to be examined

Silverman, R. D., Speece, D. L., Harring, J. R., & Ritchey, K. D. (2013). Fluency has a role in the simple view of reading. *Scientific Studies of Reading*, *17*(2), 108-133

# Terminology

- **Communalities ($h^2$)**
  - proportion of each variable's variance that can be explained by the Factors
  - if very low (say <.30), a variable is “quite unique” and should be removed, as it is definitely measuring “something else.”
  - this may be also evident when the variable cross-loads with low and/or comparable values

# Terminology

- **Eigen Values:**
  - total variance explained by the factor

| *Variables* | *Factor 1* | *Factor 2* | *Factor 3* | *Communalities* |
|---|---|---|---|---|
| CAD | 0.889 | 0.106 | −0.108 | 0.813 |
| CTOPP–Elision | −0.086 | 0.619 | 0.022 | 0.391 |
| CTOPP–Nonword Repetition | 0.117 | 0.335 | 0.017 | 0.126 |
| PRF | 0.040 | 0.226 | 0.654 | 0.480 |
| RAN | 0.062 | 0.134 | −0.788 | 0.643 |
| Spelling Fluency | 1.018 | −0.085 | −0.007 | 1.000 |
| TOWRE–SWE | −0.043 | 0.117 | 0.891 | 0.809 |
| TOWRE–PDE | 0.002 | 0.632 | 0.398 | 0.558 |
| TSWRF | 0.682 | 0.016 | 0.172 | 0.495 |
| WIF | 0.047 | 0.209 | 0.762 | 0.627 |
| WJ III–Word Attack | −0.019 | 0.929 | −0.047 | 0.866 |
| WJ III–Letter-word Identification | 0.035 | 0.780 | 0.144 | 0.630 |

$\Sigma \text{Loadings}^2$

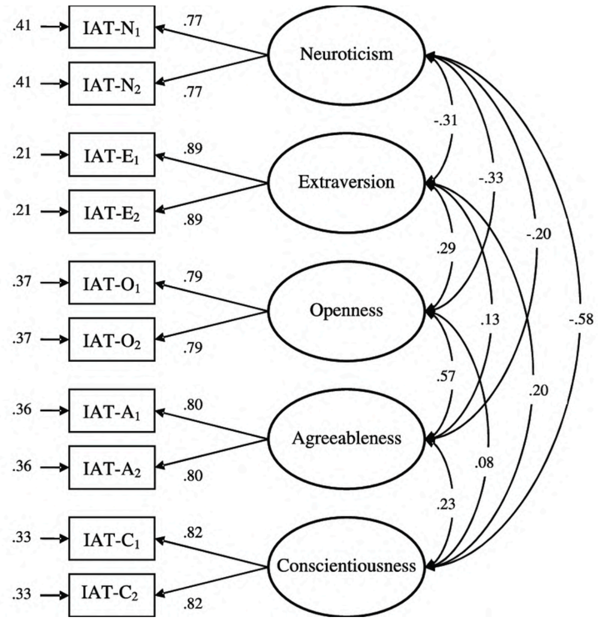
Silverman, R. D., Speece, D. L., Harring, J. R., & Ritchey, K. D. (2013). Fluency has a role in the simple view of reading. *Scientific Studies of Reading*, *17*(2), 108-133

# Terminology

- **Eigen Values:**
  - total variance explained by the factor
  - each successive component will account for less and less variance
  - factors with eigenvalues greater than 1.0 are retained (factors with a variance less than 1.0 are no better than a single variable)

# Factor Analysis Types

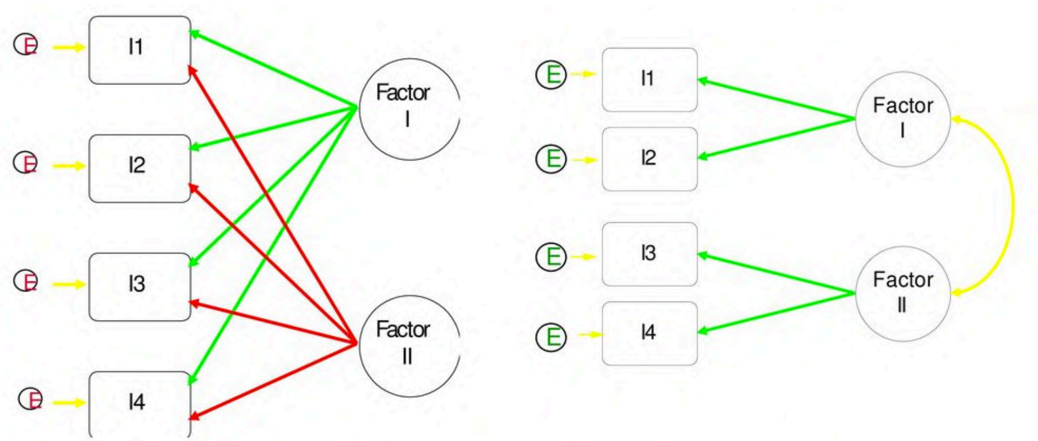
- **Q-Type**
  - similar to clustering of people
  - allows identification of groups
  - ex: participant X's responses are similar to Y's

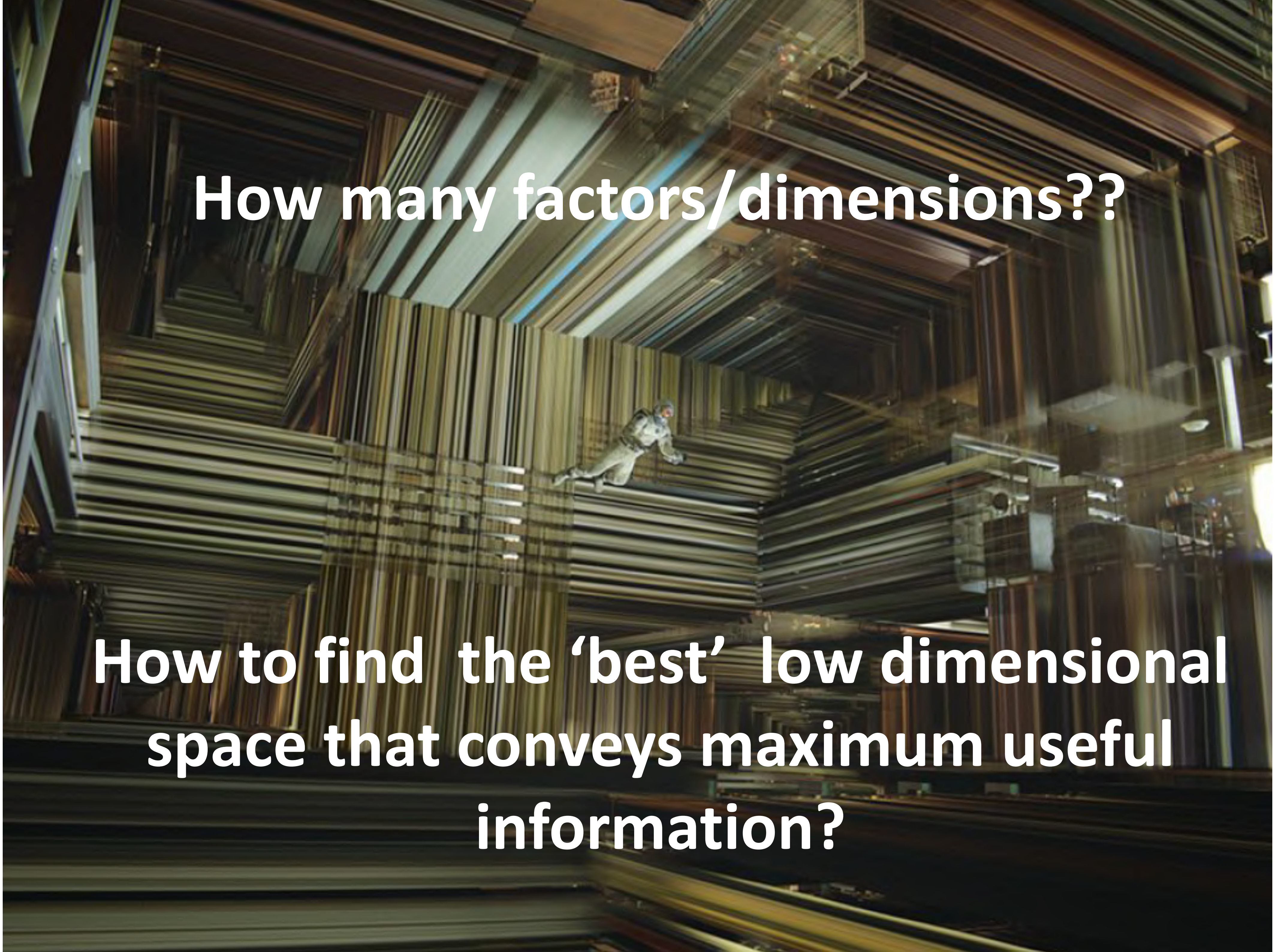


# Factor Analysis

- **Exploratory Factor Analysis**: *data-driven*
  - explore underlying structure

- **Confirmatory Factor Analysis**: *theory-driven*
  - confirm or reject pre-established theory





How many factors/dimensions??

How to find the 'best' low dimensional space that conveys maximum useful information?

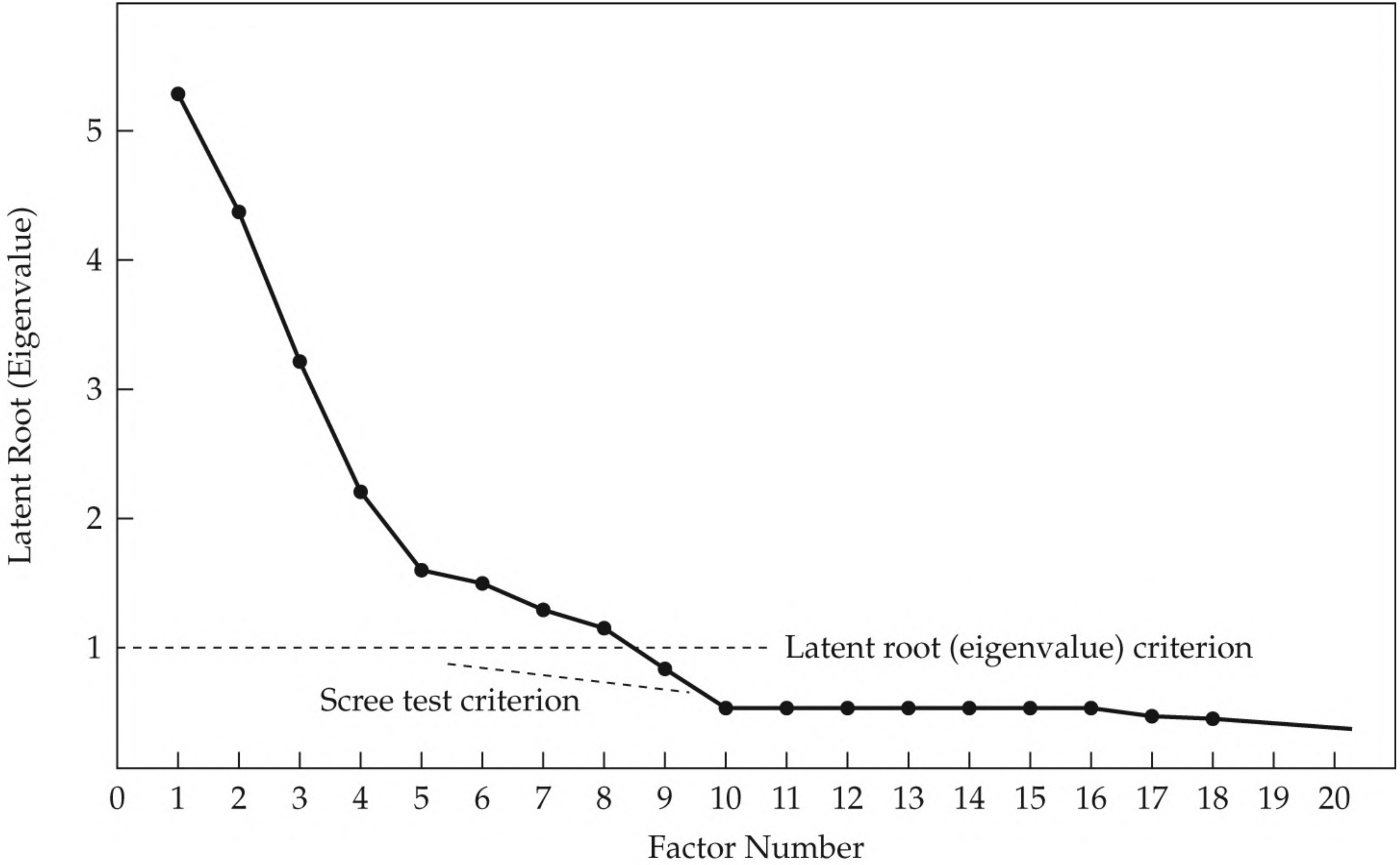
# Dimensionality Estimation

- **a priori criterion**
  - define a priori the number of factors to be extracted (testing a hypothesis about the number of factors)
  - trade off - representativeness vs parsimony
- **latent Root criterion**
  - any individual factor should account for the variance of at least one single variable – latent root or eigenvalue >1
- **scree plot/test**
  - point of inflexion in latent root plot

# Terminology

- **Scree Plot**
  - plots eigenvalue against component number
  - components with eigenvalues greater than 1 are retained (they are the ‘principal’ components)
  - components with eigenvalues less than 1 are of little use because they account for less of the variance than the original variable

# Scree Plot


Latent Root (Eigenvalue)
5
4
3
2
1
Latent root (eigenvalue) criterion
Scree test criterion
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
Factor Number

# Dimensionality Estimation

- **parallel Analysis** (widely used)
  - based on the Monte Carlo simulation
  - creating a random dataset with the same numbers of observations and variables as the original data
  - compare eigenvalues from the random data with original data

# Dimensionality Estimation Example

**Healthy-Unhealthy Music Scale (HUMS)**

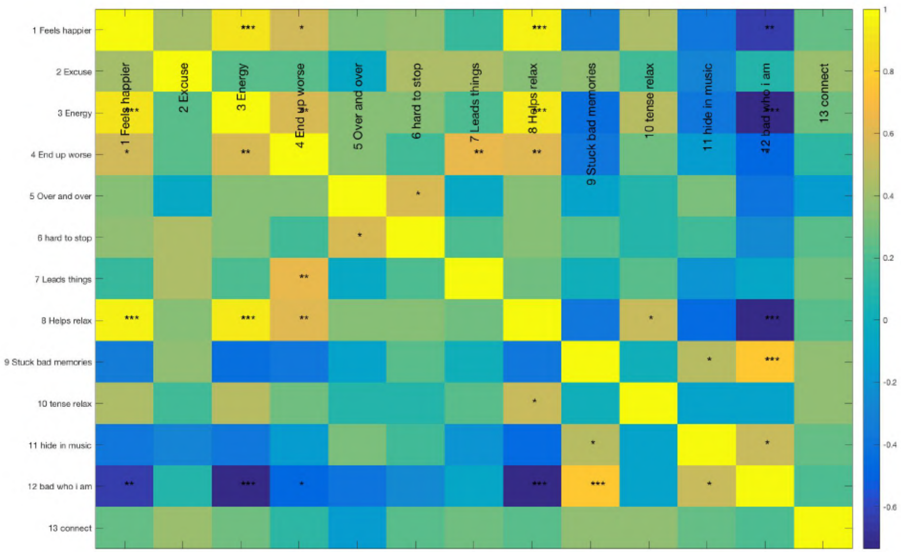
Most people believe that music is a helpful part of their lives, but sometimes it's not. When you answer the questions below, please try to recall actual moments when music has been helpful and when it has not.

Please read each statement and mark how much it applies to you. Mark only one answer for each question.

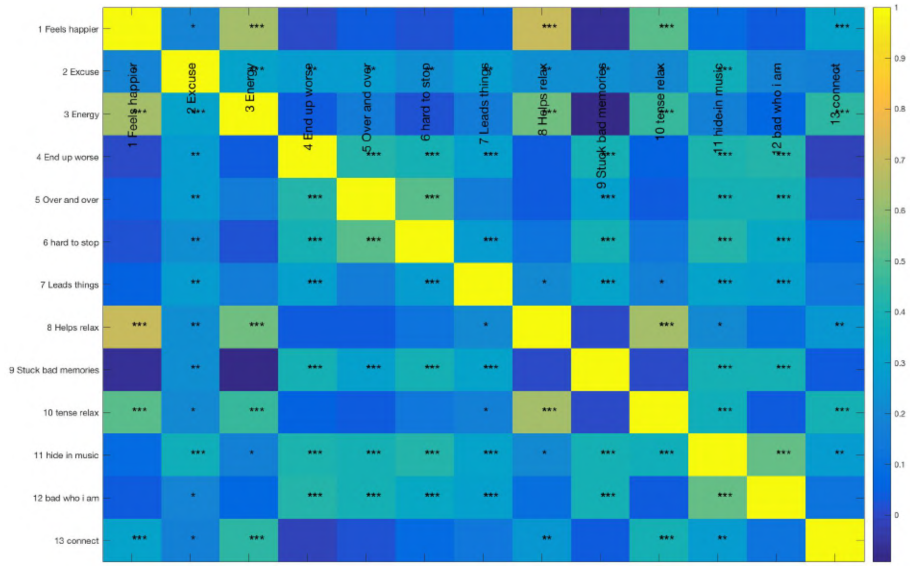
| | | Never | Rarely | Some-times | Often | Always |
|---|---|---|---|---|---|---|
| 1. | When I listen to music I get stuck in bad memories | ☐ | ☐ | ☐ | ☐ | ☐ |
| 2. | I hide in my music because nobody understands me, and it blocks people out | ☐ | ☐ | ☐ | ☐ | ☐ |
| 3. | Music helps me to relax | ☐ | ☐ | ☐ | ☐ | ☐ |
| 4. | When I try to use music to feel better I actually end up feeling worse | ☐ | ☐ | ☐ | ☐ | ☐ |
| 5. | I feel happier after playing or listening to music | ☐ | ☐ | ☐ | ☐ | ☐ |
| 6. | Music gives me the energy to get going | ☐ | ☐ | ☐ | ☐ | ☐ |
| 7. | I like to listen to songs over and over even though it makes me feel worse | ☐ | ☐ | ☐ | ☐ | ☐ |
| 8. | Music makes me feel bad about who I am | ☐ | ☐ | ☐ | ☐ | ☐ |
| 9. | Music helps me to connect with other people who are like me | ☐ | ☐ | ☐ | ☐ | ☐ |
| 10. | Music gives me an excuse not to face up to the real world | ☐ | ☐ | ☐ | ☐ | ☐ |
| 11. | It can be hard to stop listening to music that connects me to bad memories | ☐ | ☐ | ☐ | ☐ | ☐ |
| 12. | Music leads me to do things I shouldn't do | ☐ | ☐ | ☐ | ☐ | ☐ |
| 13. | When I'm feeling tense or tired in my body music helps me to relax | ☐ | ☐ | ☐ | ☐ | ☐ |

# Inter-Scale/Item Correlation

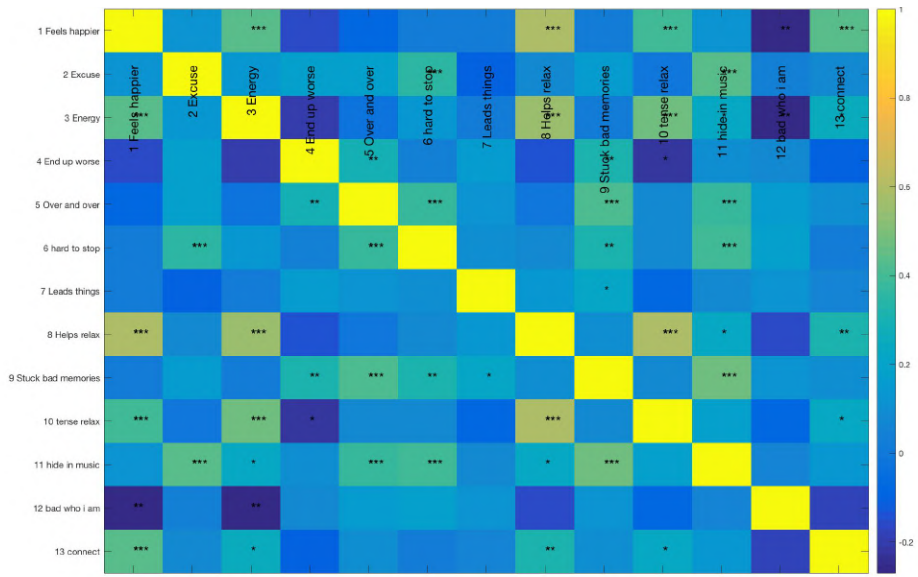
RM class 2018
25 students




141 Indians

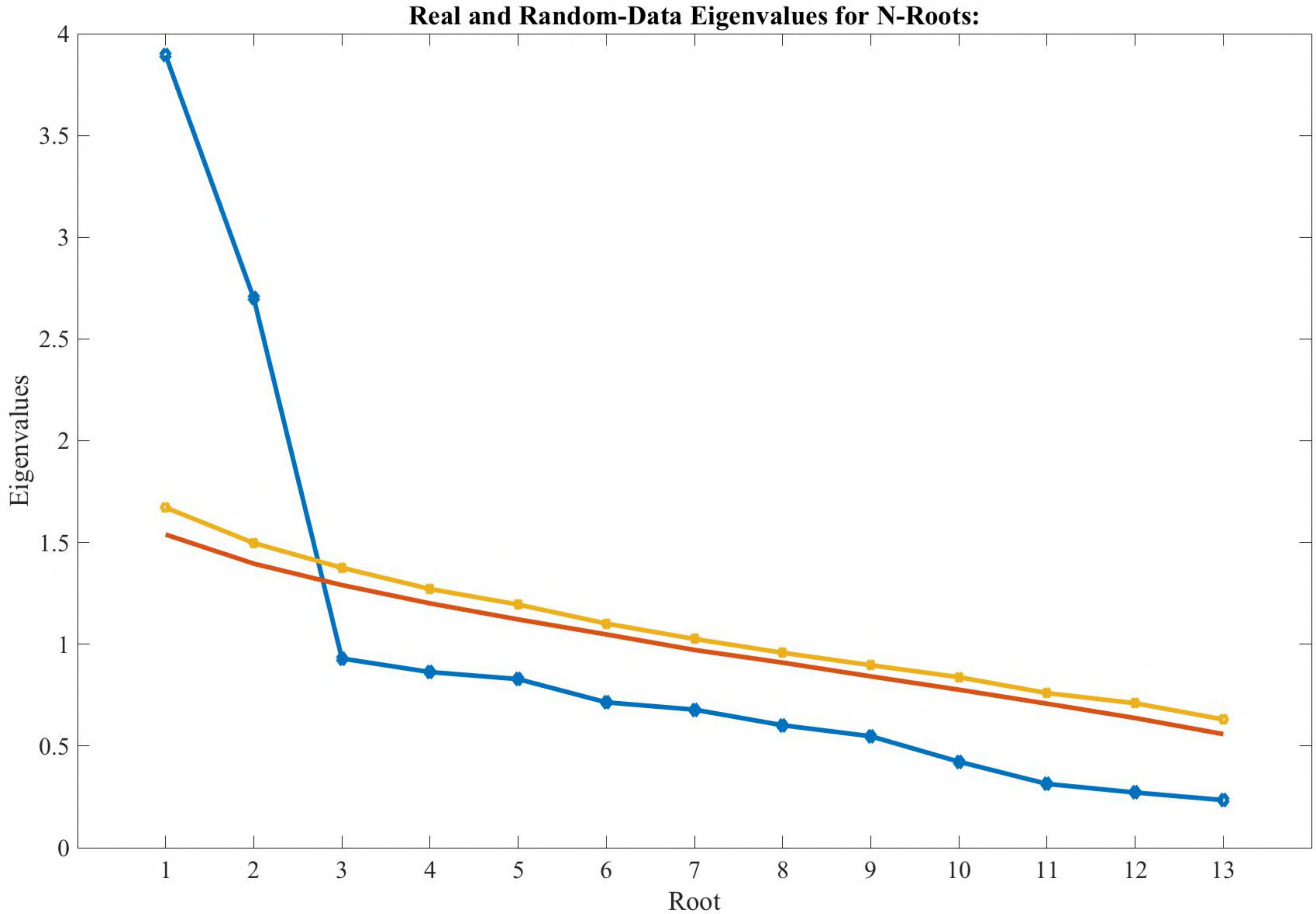



102 British



# Parallel Analysis

**141 Indians**



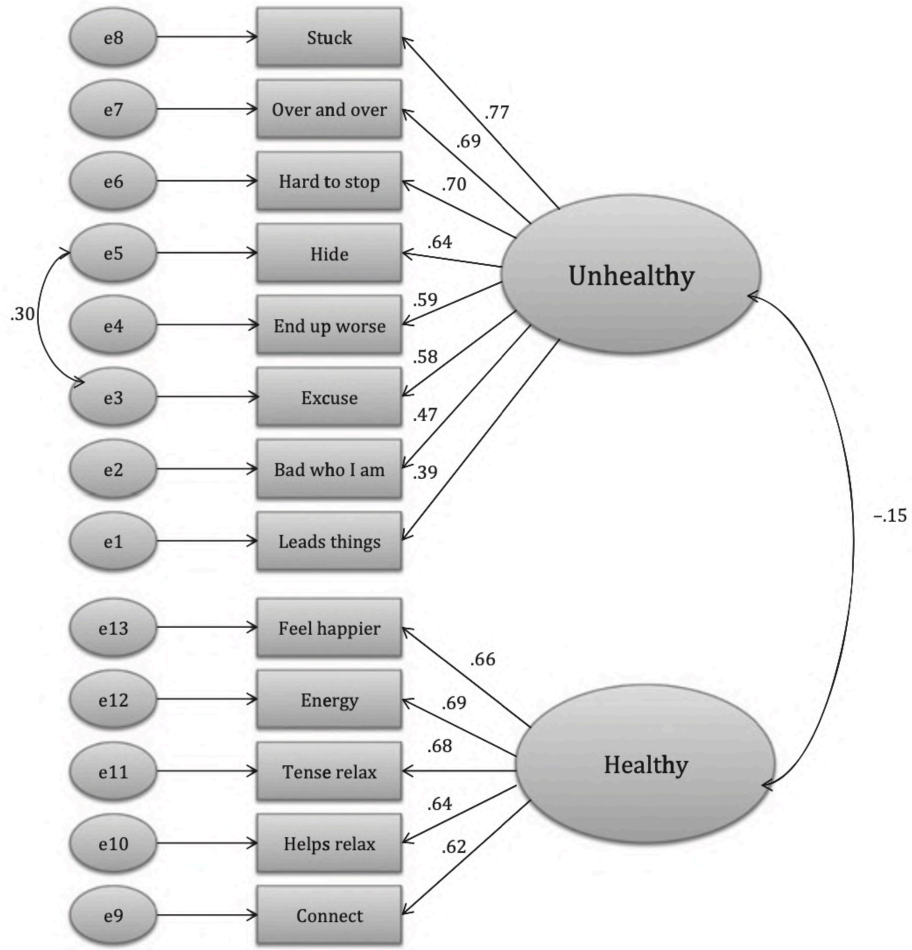
# Number of Dimensions

| |  |  |  |
|---|---|---|---|
| • the Kaiser/Eigen Value (1960) criterion | 2 | 2 | 5 |
| • Cattell's scree plot test | 2 | 2 | 4 |
| • Velicer's Minimum Average Partial (1976) | 2 | 2 | 1 |
| • Parallel Analysis | 2 | 2 | 1 |

# Factor Interpretation

**Table 2.** The factor loadings (pattern matrix) of the final version of Healthy-Unhealthy Music Scale

| Items | F1 | F2 |
|---|---|---|
| When I listen to music I get stuck in bad memories | **.760** | −.033 |
| I like to listen to songs over and over even though it makes me feel worse | **.714** | −.092 |
| It can be hard to stop listening to music that connects me to bad memories | **.658** | .187 |
| I hide in my music because nobody understands me, and it blocks people out | **.639** | .156 |
| When I try to use music to feel better I actually end up feeling worse | **.627** | −.163 |
| Music gives me an excuse not to face up to the real world | **.571** | .249 |
| Music makes me feel bad about who I am | **.521** | −.186 |
| Music leads me to do things I shouldn't do | **.428** | −.103 |
| I feel happier after playing or listening to music | −.157 | **.708** |
| Music gives me the energy to get going | −.005 | **.692** |
| When I'm feeling tense or tired in my body music helps me to relax | −.028 | **.667** |
| Music helps me to relax | .040 | **.621** |
| Music helps me to connect with other people who are like me | −.061 | **.608** |



# Factor Rotation

- the reference axes of the factors are tuned about the origin until some other position has been reached
- the ultimate effect of rotating the factor matrix is to redistribute the variance from earlier factors to later ones to achieve a simpler, theoretically more meaningful factor pattern

# Factor Rotation

- **Orthogonal**: (varimax, quartimax, & equamax)
  - are the most widely used rotational methods.
  - preferred method when the research goal is data reduction to either a smaller number of variables or a set of uncorrelated measures for subsequent use in other multivariate techniques.
- **Oblique**: (premix, direct oblimin)
  - best suited to the goal of obtaining several theoretically meaningful factors or constructs because, realistically, very few constructs in the "real world" are uncorrelated.

# Orthogonal Factor Rotation

- **Varimax**
  - minimises number of variables with high loading on a factor
- **Quartimax**
  - maximizes the variance across the rows of the factor matrix

- **Quartimax (simplify rows).**
- **Varimax (simplify columns).**
- **Equimax (combination).**

# Factor Interpretation

- variables that cross-load (load highly on two or more factors) are usually deleted unless theoretically justified or the objective is strictly data reduction.

- re-specification of a factor analysis can include options such as:
    - deleting a variable(s) (ex: based on *SMC*, *VIF*)
    - changing rotation methods
    - increasing or decreasing the number of factors.

# Principal Component Analysis

# ***Research Question?***

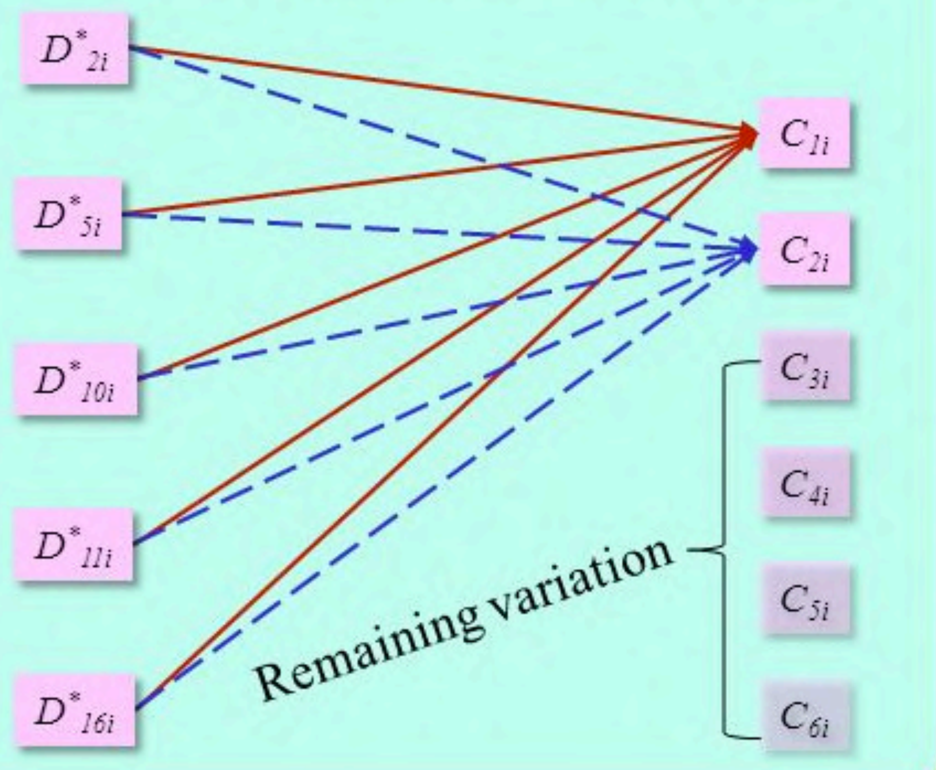
*Rather than asking … **"Can We <u>Forge These Several Indicators</u> Together Into A <u>Smaller Number Of Composites</u> With Defined Statistical Properties?"***

Then, we would need …
***<u>Principal Components Analysis</u> (PCA)***

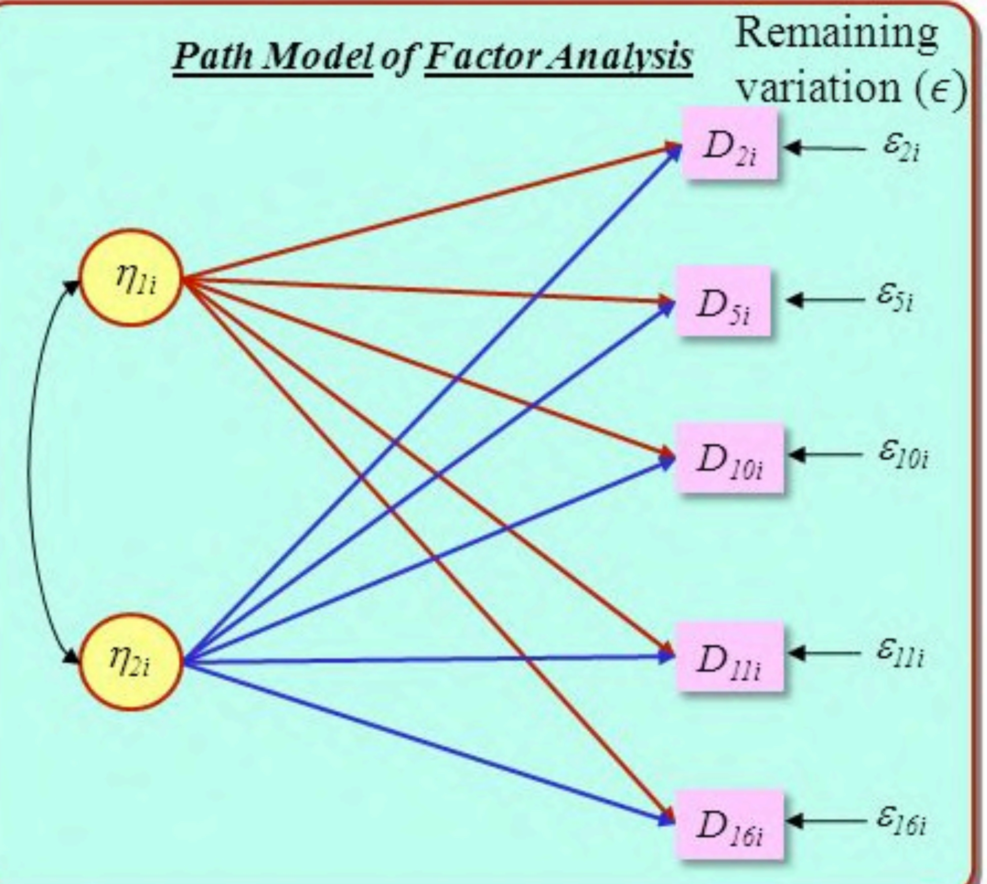
*We could ask … **"Are There A Number Of <u>Unseen (Latent) Factors (Constructs)</u> Acting "<u>Beneath</u>" These Indicators To Forge Their Observed Values?"***

Instead, we would need …
***<u>Factor Analysis</u> (CFA** or **EFA?)***

***<u>Path Model</u> of <u>Principal Components Analysis</u>***

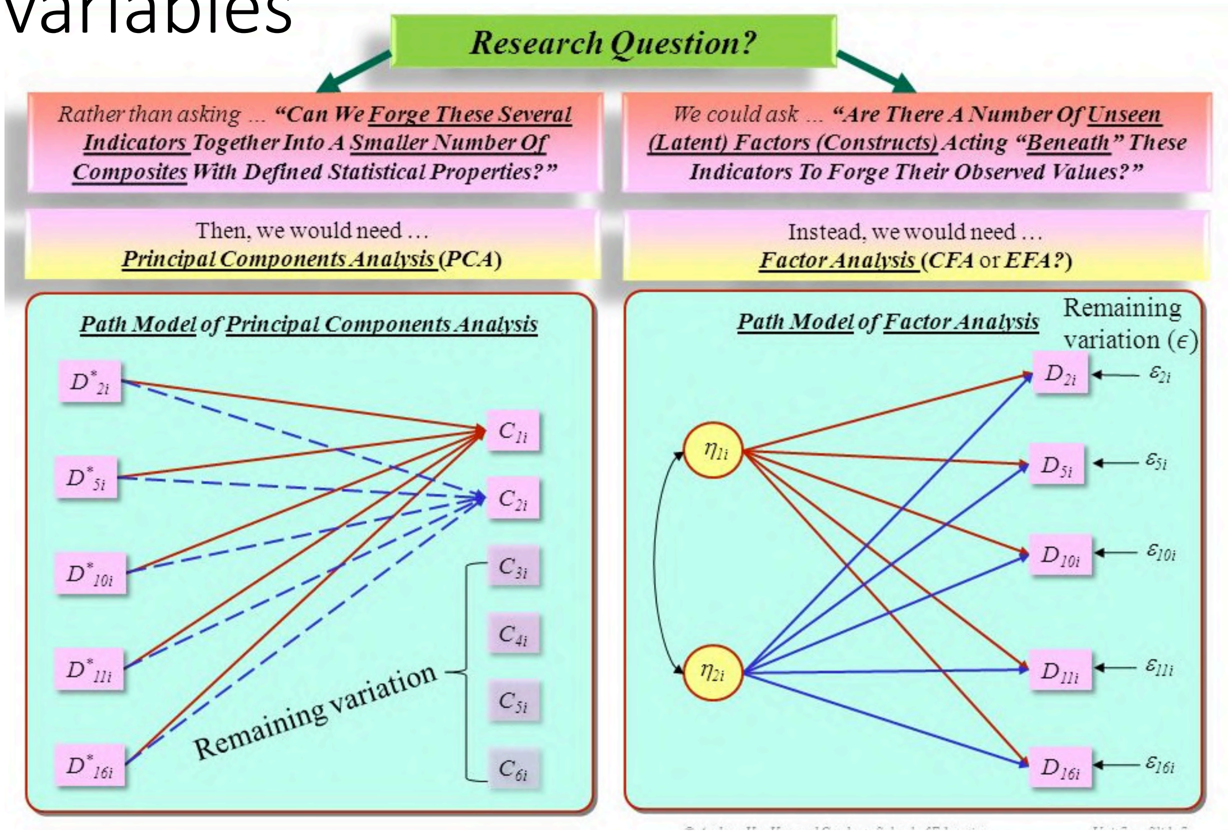



***<u>Path Model</u> of <u>Factor Analysis</u>***

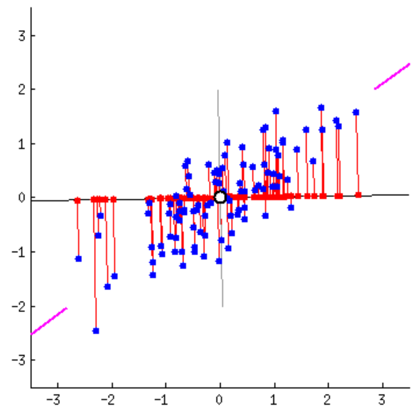


# PCA

- idea —> reduce the number of variables of a data set while preserving as much information as possible.
- dimensionality reduction by creating linear combinations of variables



# PCA




- Example: Combining two variables into a single component
  - Fit a regression line that represents the ’best’ summary of the linear relationship between the variables
  - This line, representing a new component, would capture most of the ’essence’ of the two variables

# PCA

- If there are more than two variables...
  - this process is repeated until all variables have been assigned to a component
  - gives as many components as variables in decreasing order of variance explained
  - however, only the first few components are likely to be useful..

# PCA

- Assumptions:
  - at least interval level data
  - a linear relationship between all variables
  - sampling adequacy (KMO, ~15 cases/ variable), Bartlett's test of sphericity
  - normally distributed (no outliers)

# PCA

- Subtract mean from data (center X)
- (Typically) scale each dimension by its variance
  - Helps to pay less attention to magnitude of dimensions
- Compute covariance matrix S $\quad \mathbf{S} = \frac{1}{N}\mathbf{X}^\intercal\mathbf{X}$
- Compute k largest eigenvectors of S
- These eigenvectors are the k principal components

https://www.youtube.com/watch?v=g-Hb26agBFg

https://www.youtube.com/watch?v=PFDu9oVAE-g

# Principal Components

- **principal components:** linear combinations of original variables that result in an axis or a set of axes that explain most of the variability in the dataset
- variables that correlate highly with each other are grouped together into underlying variables, or components
- In mathematical terms, we can say that the first Principal Component is the eigenvector of the covariance matrix corresponding to the maximum eigenvalue

# Component Scores & Loadings

- each original variable is assigned a component score and a component loading
- **Component scores** = score/projection on a given component (can be used in subsequent statistical analyses, e.g., regression)
- **Component loadings** = correlation of the original variable with a given component - can be used to determine the importance of a particular variable to a component (Higher loadings = more important)

# Dimensionality Estimation

- **Percentage of Variance** criterion
  - achieving a specified cumulative percentage of total variance.
    - typical values – natural sciences ~95%;
    - typical values – social sciences > ~60%
- **Parallel Analysis** (widely used)
  - based on the Monte Carlo simulation
  - creating a random dataset with the same numbers of observations and variables as the original data
  - compare eigenvalues from the random data with original datas

# Dimensionality Estimation

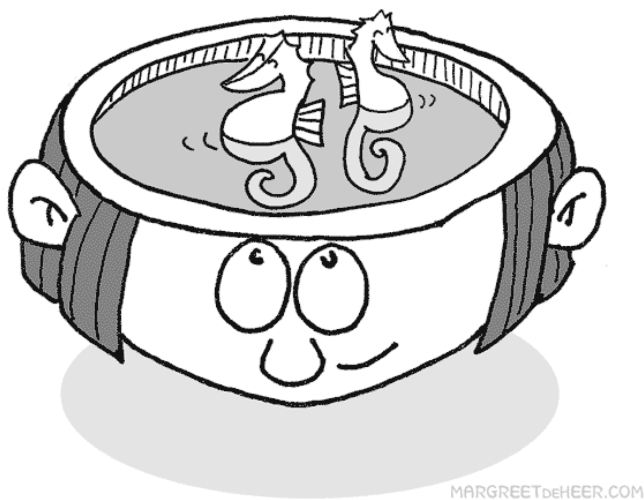
- **latent Root criterion**
  - any individual factor should account for the variance of at least one single variable – latent root or eigenvalue >1
- **scree plot/test**
  - point of inflexion in latent root plot

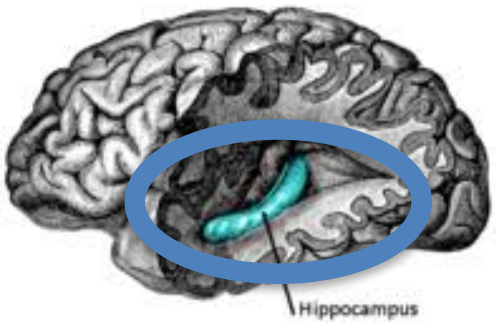
# Rotation (similar to FA)

- the reference axes of the factors are tuned about the origin until some other position has been reached
- the ultimate effect of rotating the factor matrix is to redistribute the variance from earlier factors to later ones to achieve a simpler, theoretically more meaningful factor pattern

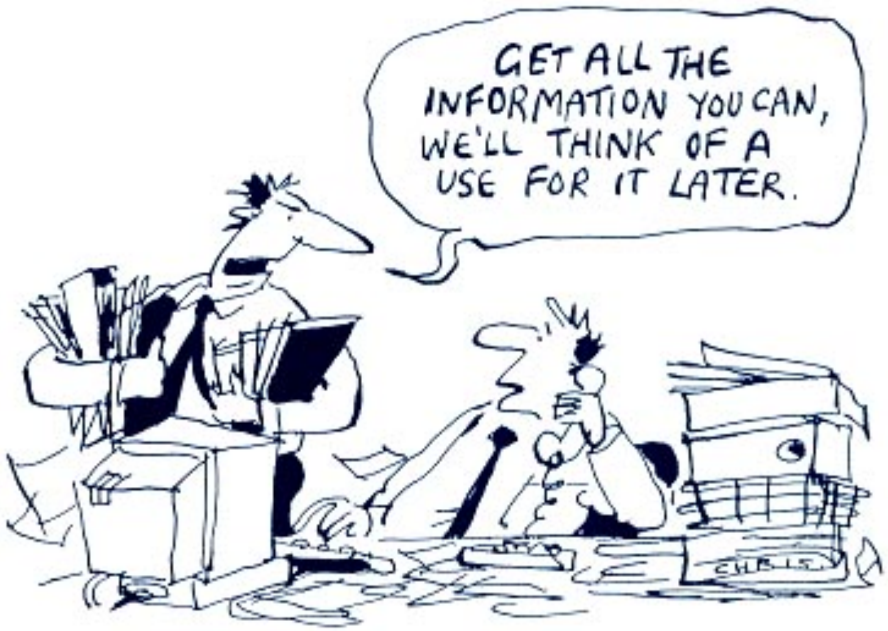
# PCA EXAMPLE

## Hippocampus structure
## vs.
## Affective personality dimensions




Hippocampus

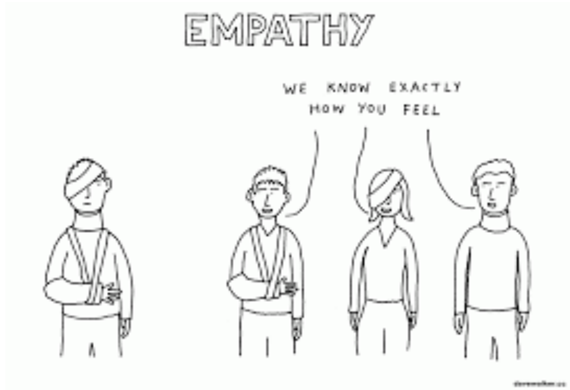

Personality Traits
model
higher
grades
reliably
dutifully
curiosity
cooperative
CANOE
develops
differences
compared
predictors
Job
positive
activity
preference
generally
sociability
high
withdrawn
global
nervous
pole
related
five traits
alone
analysis
talkativeness
adolescence
neuroticism
measures
company
cluster
success
unusual
vulnerability
synthesis
conscientiousness
FFM
Dominance
anger
seeking
person
emotional
outcome
anxiety
OCEAN
refers
around
specific
inhibited
performance
Acronyms


GET ALL THE INFORMATION YOU CAN, WE'LL THINK OF A USE FOR IT LATER.

# Dimensionality Estimation


**28 personality measures**



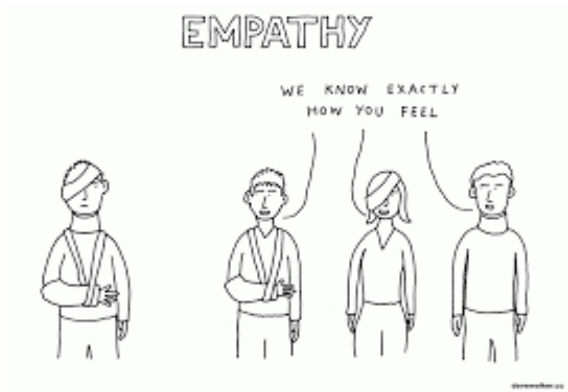





# Normality Check

Lilliefors test and Box Cox Transformation

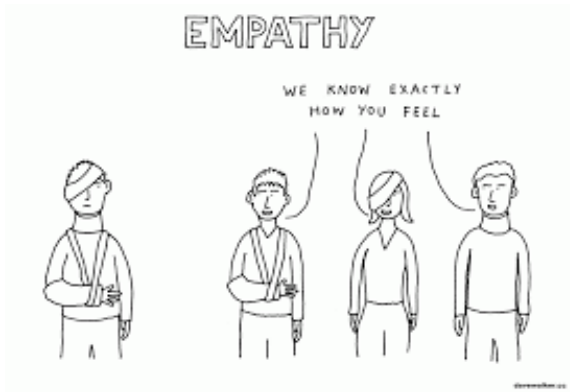






# Problem of Variable Collinearity

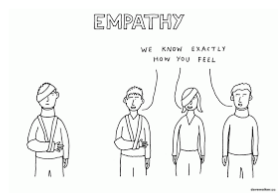




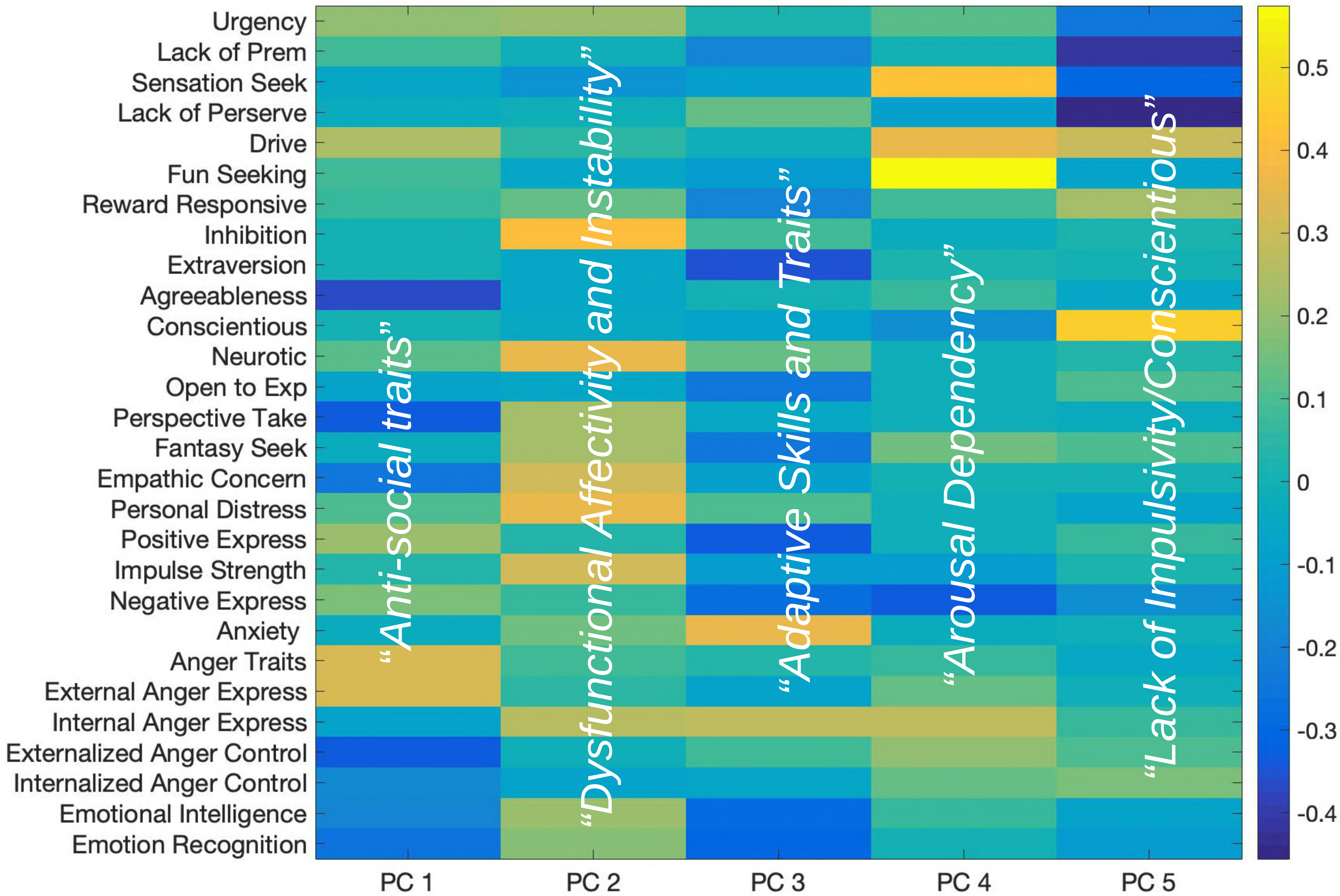


# Dimensionality Estimation

- the Kaiser (1960) criterion **9**
- Cattell's scree plot test **5/9**
- Parallel Analysis (PA) **5**
- Velicer's Minimum Average Partial (1976) **5**

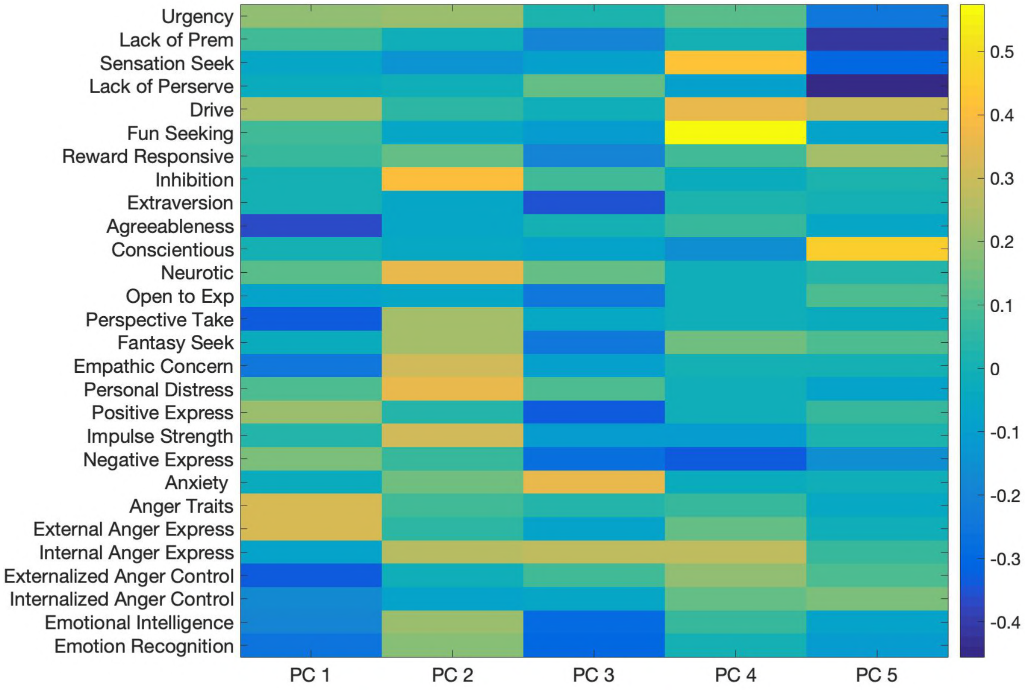


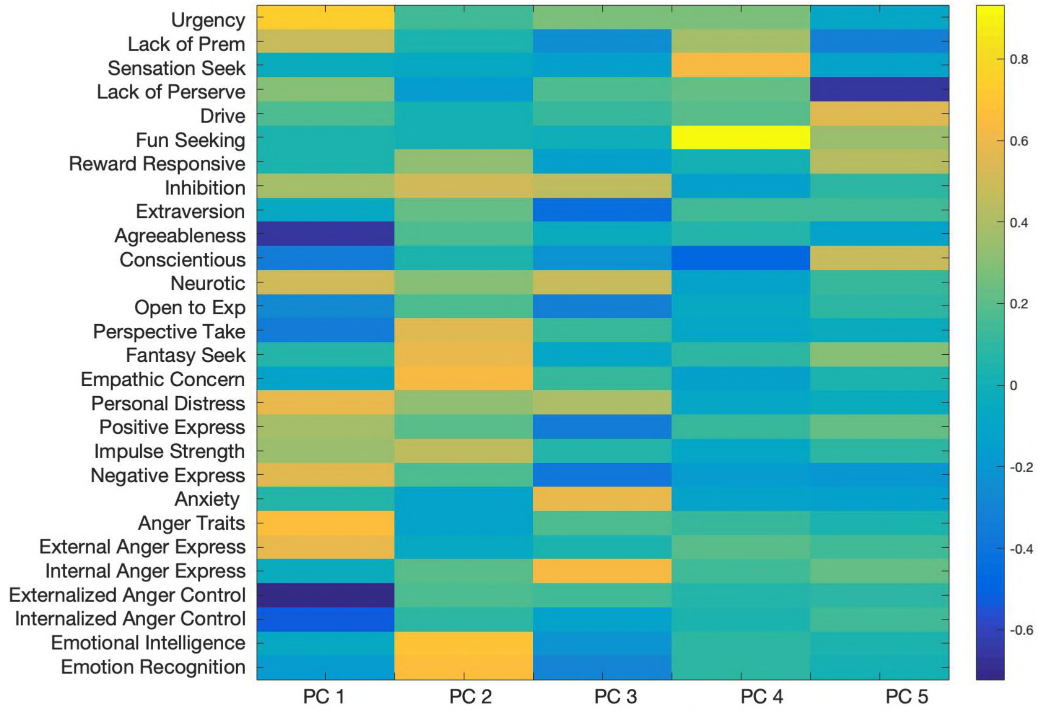
# PCA Loading Matrix




58% variance

# PCA vs FA Loading Matrix



# 5 final variables

"*Anti-social traits*"

*"Dysfunctional Affectivity and Instability"*

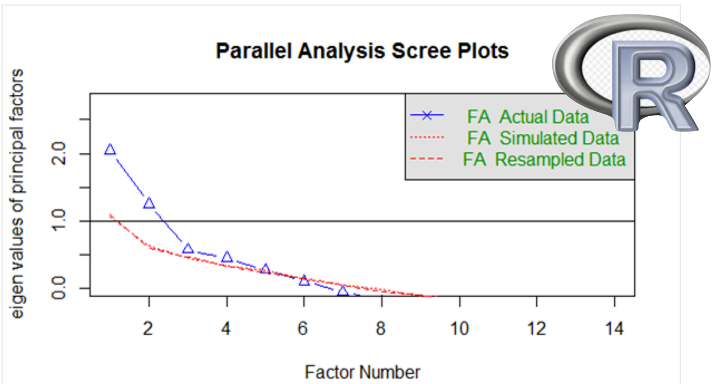
*"Adaptive skills and traits"*

*"Arousal Dependency"*

*"Impulsivity*"

# But ........

**50% variance**


Brain Structure and Function
December 2017, Volume 222, Issue 9, pp 3915–3925 | Cite as

Amygdala structure and core dimensions of the affective personality

# So ……..

"*Anti-social traits*"

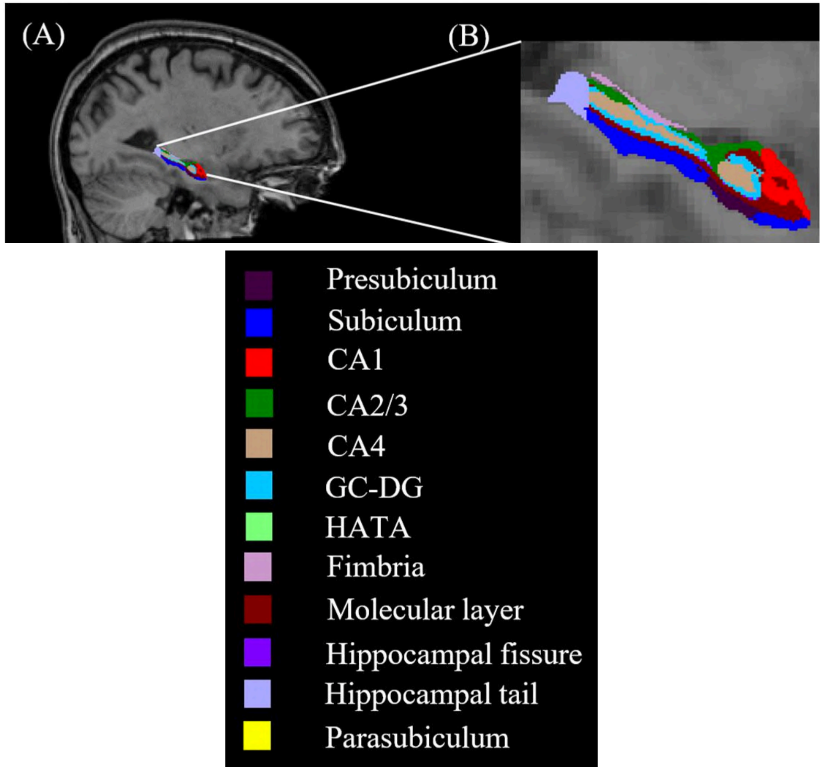
*"Arousal Dependency"*

*"Impulsivity*"



*"Arousal Dependency & Disinhibition"* (ADD)

# Now what??




*“Dysfunctional Affectivity and Instability”*

*“Adaptive skills and traits”*

*“Arousal Dependency and Disinhibition”*

**3 (components) x 12 (subfields) x 2 (hemispheres) = 72 comparisons !!!!!!**

| Factor analysis | Principal component analysis |
| --- | --- |
| Number of factors pre-determined | Number of components evaluated ex post |
| Many potential solutions | Unique mathematical solution |
| Factor matrix is estimated | Component matrix is computed |
| Factor scores are estimated | Component scores are computed |
| More appropriate when searching for an underlying structure | More appropriate for data reduction (no prior underlying structure assumed) |
| Factors are not necessarily sorted | Factors are sorted according to the amount of explained variability |
| Only common variability is taken into account | Total variability is taken into account |
| Estimated factor scores may be correlated | Component scores are always uncorrelated |
| A distinction is made between common and specific variance | No distinction between specific and common variability |
| Preferred when there is substantial measurement error in variables | Preferred as a preliminary method to cluster analysis or to avoid multicollinearity in regression |
| Rotation is often desirable as there are many equivalent solutions | Rotation is less desirable, unless components are difficult to be interpreted and explained variance is spread evenly across components |