FACTOR ANALYSIS

Types of FA

- Exploratory Factor Analysis (EFA): Is used to discover the factor structure of a construct. It is data driven.

- Confirmatory Factor Analysis (CFA): is used to confirm the fit of the hypothesized factor structure to the observed (sample) data. It is theory driven.

Holzinger and Swineford Dataset

- The classic Holzinger and Swineford (1939) dataset consists of **mental ability test** scores of seventh- and eighth-grade children from two different schools (Pasteur and Grant-White).
- In the original dataset (available in the MBESS package), there are scores for 26 tests.
- However, a smaller subset with 9 variables is more widely used in the literature.
- The raw data analyzed here are on the first **six** psychological item (variable)s in for 301.

Holzinger and Swineford Dataset

- The classic Holzinger and Swineford (1939) dataset consists of **mental ability test** scores of seventh- and eighth-grade children from two different schools (Pasteur and Grant-White).
- In the original dataset (available in the MBESS package), there are scores for 26 tests.
- However, a smaller subset with 9 variables is more widely used in the literature.
- The raw data analyzed here are on the first **six** psychological item (variable)s in for 301.

Holzinger and Swineford Dataset

- Id: Identifier
- Sex: Gender
- Ageyr:Age, year part
- Agemo: Age, month part
- School: School (Pasteur or Grant-White)
- Grade: Grade
- X1:Visual perception
- X2: Cubes
- X3: Lozenges
- X4: Paragraph comprehension
- X5: Sentence completion
- X6: Word meaning
- X7: Speeded addition
- X8: Speeded counting of dots
- X9: Speeded discrimination straight and curved capitals

Model Covariance

coefficient

Predictor(observed variable)

$$y = b_0 + b_1 x + \epsilon$$

Factor (unobserved variable)

$$y_1 = \tau_1 + \lambda_1 \eta + \epsilon_1$$

loading

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = \begin{pmatrix} \tau_1 \\ \tau_2 \\ \tau_3 \end{pmatrix} + \begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \end{pmatrix} (\eta_1) + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \end{pmatrix}$$

$$\begin{aligned} y_1 &= \tau_1 + \lambda_1 \eta_1 + \epsilon_1 \\ y_2 &= \tau_2 + \lambda_2 \eta_1 + \epsilon_2 \\ y_3 &= \tau_3 + \lambda_3 \eta_1 + \epsilon_3 \end{aligned}$$

Model implied Covariance

Data Covariance

Model

$$\Sigma(\theta) = \mathbf{\Lambda \Psi \Lambda}' + \Theta_\epsilon$$

versus

$$\Sigma$$

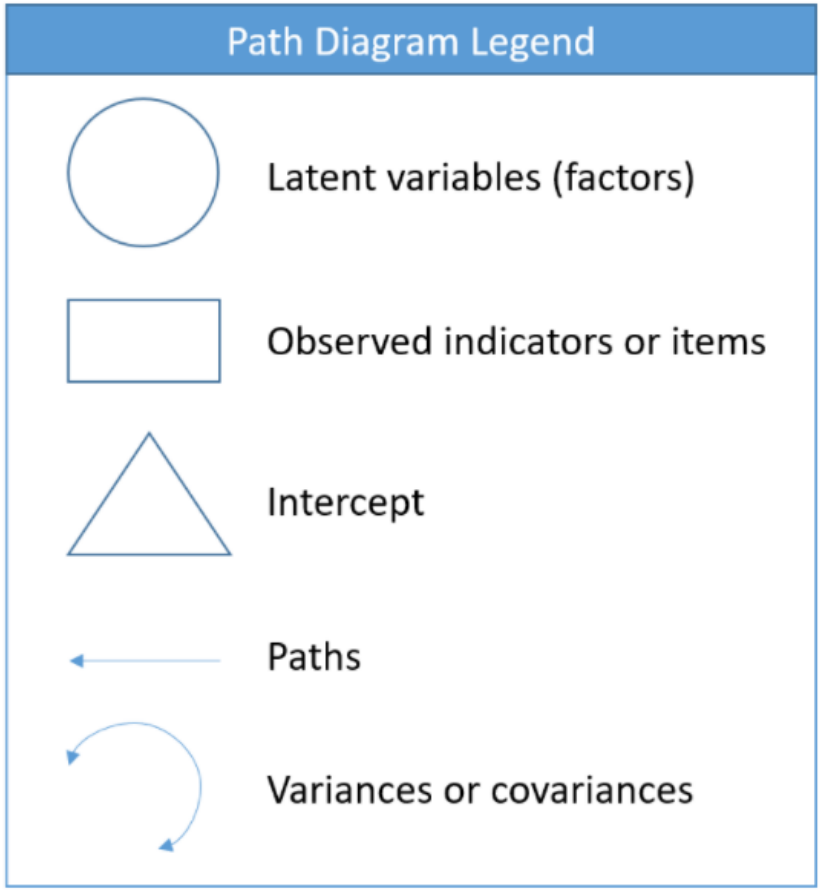
Population

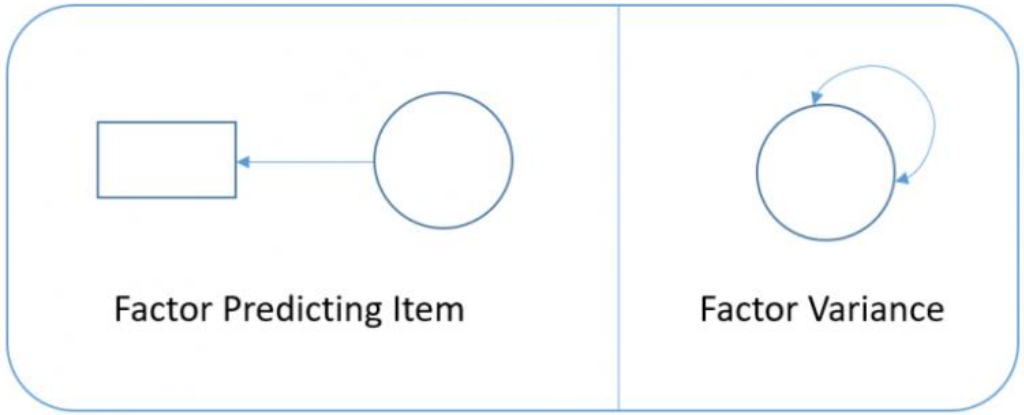
Covariance of factors

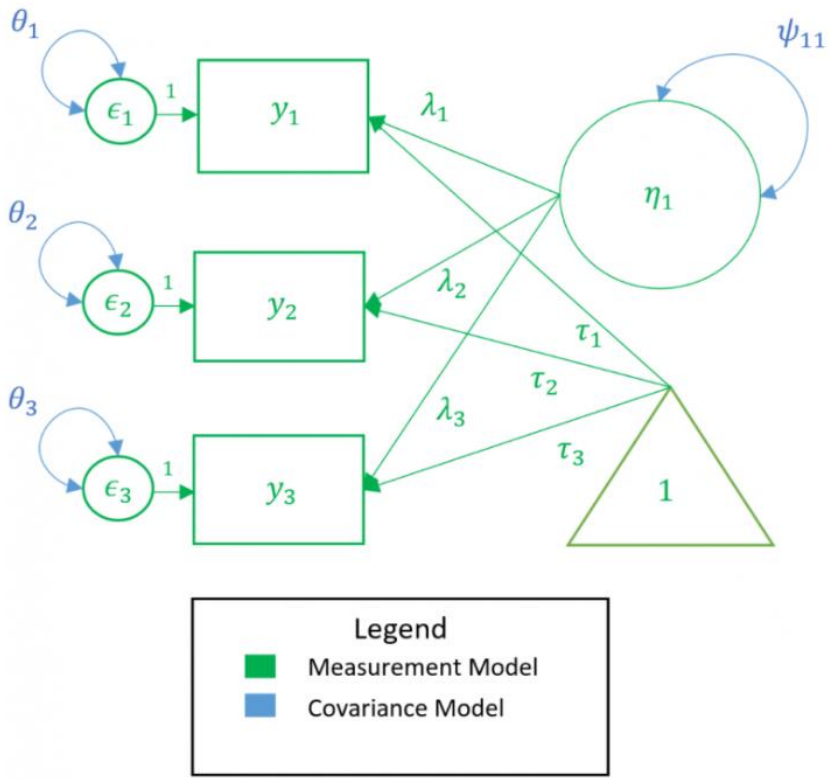
Covariance of residuals

$$\Sigma(\theta) = \begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \end{pmatrix} (\psi_{11}) \begin{pmatrix} \lambda_1 & \lambda_2 & \lambda_3 \end{pmatrix} + \begin{pmatrix} \theta_{11} & \theta_{12} & \theta_{13} \\ \theta_{21} & \theta_{22} & \theta_{23} \\ \theta_{31} & \theta_{32} & \theta_{33} \end{pmatrix}$$

Path Diagram







$$y_1 = \tau_1 + \lambda_1\eta_1 + \epsilon_1$$
$$y_2 = \tau_2 + \lambda_2\eta_1 + \epsilon_2$$
$$y_3 = \tau_3 + \lambda_3\eta_1 + \epsilon_3$$

$$\Sigma(\theta) = \begin{pmatrix}\lambda_1\\ \lambda_2\\ \lambda_3\end{pmatrix}(\psi_{11})\begin{pmatrix}\lambda_1 & \lambda_2 & \lambda_3\end{pmatrix} + \begin{pmatrix}\theta_{11} & 0 & 0\\ 0 & \theta_{22} & 0\\ 0 & 0 & \theta_{33}\end{pmatrix}$$

Sample Covariance matrix

$\Sigma(\theta) = \mathbf{\Lambda \Psi \Lambda}' + \Theta_\epsilon$ versus Σ versus $S = \hat{\Sigma}$

round(cov(HS[,7:9]),2)

Symmetric matrix

	x1	x2	x3
x1	1.36	0.41	0.58
x2	0.41	1.39	0.45
x3	0.58	0.45	1.28

Degree of Freedom

- Known values: total number of parameters

$$p(p+1)/2$$

For three items

$$3(4)/2 = 6$$

$$\Sigma(\theta) = \begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \end{pmatrix} (\psi_{11}) \begin{pmatrix} \lambda_1 & \lambda_2 & \lambda_3 \end{pmatrix} + \begin{pmatrix} \theta_{11} & \theta_{12} & \theta_{13} \\ \theta_{21} & \theta_{22} & \theta_{23} \\ \theta_{31} & \theta_{32} & \theta_{33} \end{pmatrix}$$

Highlight the unique parameters. Count 10.

Fixed vs. free parameters

- **Fixed parameters:** predetermined to have a specific value

$$\Sigma(\theta) = \begin{pmatrix} \lambda_1 = 1 \\ \lambda_2 = 1 \\ \lambda_3 = 1 \end{pmatrix} (\psi_{11} = 1) \, (\lambda_1 = 1 \quad \lambda_2 = 1 \quad \lambda_3 = 1) + \begin{pmatrix} \theta_{11} = 1 & \theta_{12} = 0 & \theta_{13} = 0 \\ \theta_{21} = 0 & \theta_{22} = 1 & \theta_{23} = 0 \\ \theta_{31} = 0 & \theta_{32} = 0 & \theta_{33} = 1 \end{pmatrix}$$

- **Free parameters**

number of free parameters= number of unique parameters –number of fixed parameters

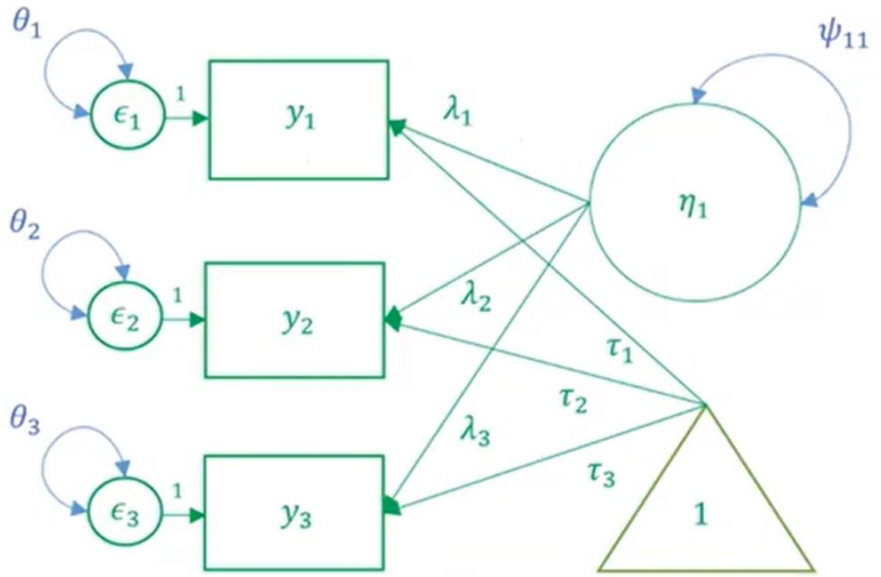
Fixed vs. free parameters

df= number of known values–number of free parameters

Calculated the degrees of freedom for our model. Should be 6.

- df negative, known<free (under-identified, cannot run model)
- df=0,known=free (just identified or saturated, no model fit)
- df positive, known>free (over-identified, model fit can be assessed)

Three-Item CFA



- Known values=6
- Free parameters=7-0
- Df=6-7=-1

Identification of Three-Item

- **Marker method**: fixes the first loading of each factor to 1

$$\Sigma(\theta) = \psi_{11} \begin{pmatrix} 1 \\ \lambda_2 \\ \lambda_3 \end{pmatrix} (1 \quad \lambda_2 \quad \lambda_3) + \begin{pmatrix} \theta_{11} & 0 & 0 \\ 0 & \theta_{22} & 0 \\ 0 & 0 & \theta_{33} \end{pmatrix}$$

- **Variance standardized method**: fixes the variance of each factor to 1 but freely estimates all loadings.

$$\Sigma(\theta) = (1) \begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \end{pmatrix} (\lambda_1 \quad \lambda_2 \quad \lambda_3) + \begin{pmatrix} \theta_{11} & 0 & 0 \\ 0 & \theta_{22} & 0 \\ 0 & 0 & \theta_{33} \end{pmatrix}$$

Lavvan syntax

- **~ predict** regression
- **=~ indicator** factor analysis
- **~~ covariance**
- **~1 intercept**
- **1* fixes parameter**
- **NA* frees parameter** useful to override default marker method
- **a* labels the parameter** 'a', model constraints

Y~X, Y<----X

f~~f

Y~1

f=~1*y1+y2+y3

Marker method in laavan

#one factor three items, default marker method

- vis <-'visual=~x1+x2+x3‘

fit1factor1 <- cfa(vis, data = HS)

summary(fit1factor1)

- visual<-'vis1=~1*x1+x2+x3'

fit2factor1<-cfa(visual,data=HS)

summary(fit2factor1)

Marker method in laavan

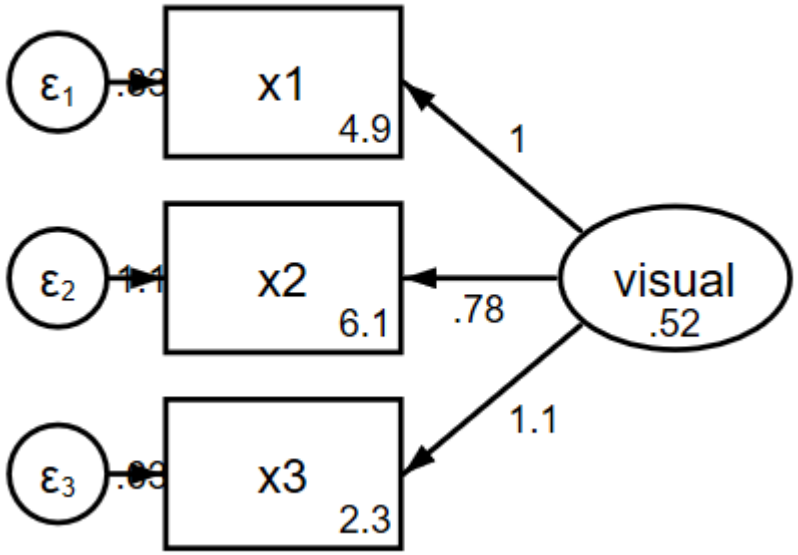
```
Latent Variables:

  visual =~
    x1
    x2
    x3
 Estimate  Std.Err    z-value  P(>|z|)

    1.000
    0.778    0.141      5.532    0.000
    1.107    0.214      5.173    0.000


Variances:

   .x1
   .x2
   .x3
    visual
 Estimate  Std.Err    z-value  P(>|z|)
    0.835    0.118      7.064    0.000
    1.065    0.105     10.177    0.000
    0.633    0.129      4.899    0.000
    0.524    0.130      4.021    0.000
```



Variance std method in laavan

#one factor three items, variance std method

```
visua1v<-'visv=~NA*x1+x2+x3+
          visv~~1*visv'
fit3factor1<-cfa(visualv,data=HS)
summary(fit2factor1)
```

```
visualv1 <-'visv1=~NA*x1+x2+x3'
fit4factor1<-cfa(visualv1,data=HS,std.lv=TRUE)
summary(fit4factor1)
```

Variance std method in laavan

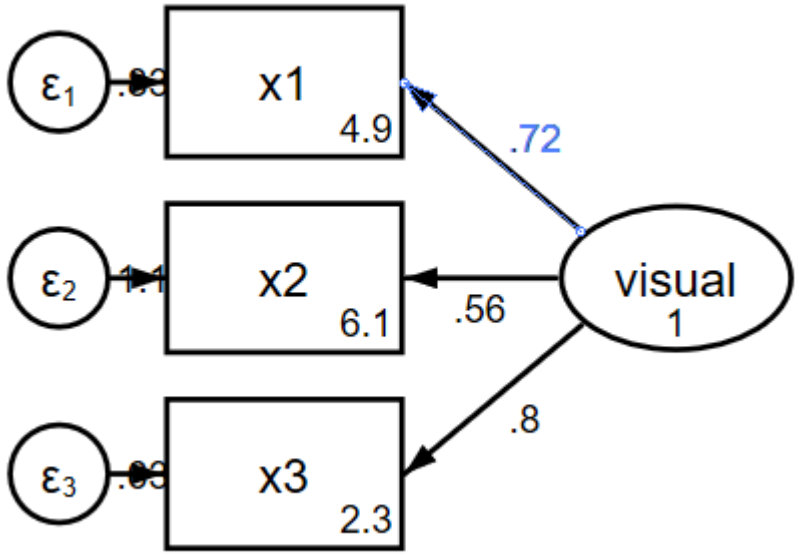
```
Latent Variables:

  visv =~
    x1
    x2
    x3
 Estimate  Std.Err  z-value  P(>|z|)

    0.724    0.090    8.043    0.000
    0.563    0.082    6.847    0.000
    0.801    0.093    8.626    0.000


 Variances:

     visv
    .x1
    .x2
    .x3
  Estimate  Std.Err  z-value  P(>|z|)
     1.000
     0.835    0.118    7.064    0.000
     1.065    0.105   10.177    0.000
     0.633    0.129    4.899    0.000
```



Automatic standardization in lavaan

```
visualv1 <-'visv1=~NA*x1+x2+x3'
fit4factor1<-cfa(visualv1,data=HS,std.lv=TRUE)
summary(fit4factor1, )
```

Automatic standardization in lavaan

```
Latent Variables:

  visual =~
    x1
    x2
    x3
 Estimate  Std.Err   z-value  P(>|z|)   Std.lv  Std.all

    1.000                               0.724    0.621
    0.778    0.141     5.532    0.000    0.563    0.479
    1.107    0.214     5.173    0.000    0.801    0.710
```

solution standardizes the factor loadings by the standard deviation of both the predictor (the factor, X) and the outcome (the item, Y)

we only standardize by the predictor (the factor, X).

For one standard deviation increase in Visual, x2 goes up by 0.479 standard deviation unites.

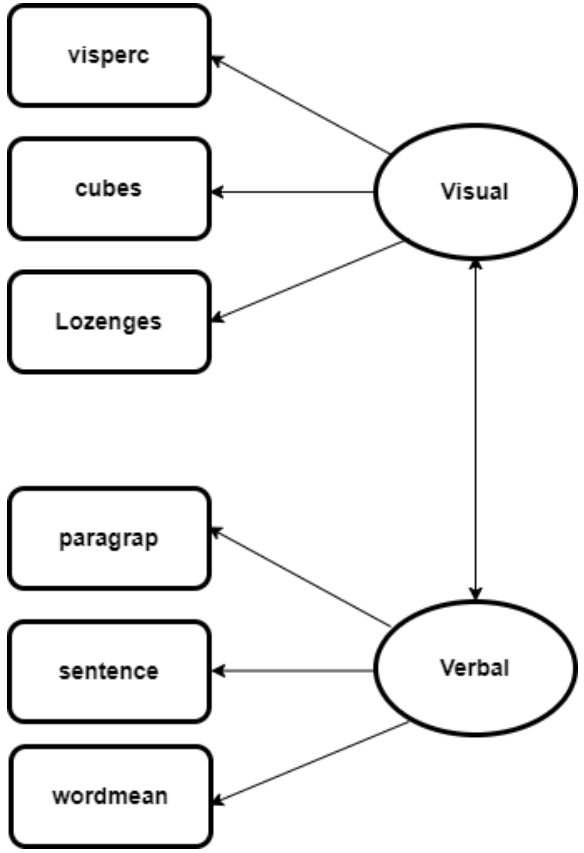
```
 Variances:

 Estimate  Std.Err   z-value  P(>|z|)   Std.lv  Std.all
    0.835    0.118     7.064    0.000    0.835    0.614
    1.065    0.105    10.177    0.000    1.065    0.771
    0.633    0.129     4.899    0.000    0.633    0.496
    0.524    0.130     4.021    0.000    1.000    1.000
```

0.614^2=0.377

Only 37.7 of variance in x1 can be explained by visual.

Two Factor Confirmatory Factor Analysis



Two Factor Confirmatory Factor Analysis

```
#correlated two factor solution, marker method
HS.model1 <-"visual=~x1+x2+x3
          textual =~ x4 + x5 + x6"
fit1factor2 <- cfa(HS.model1, data = HS)
summary(fit1factor2, standardized = TRUE)
```

```
#correlated two factor solution, variance std method
HS.model2 <-"visual=~x1+x2+x3
          textual =~ x4 + x5 + x6"
fit2factor2 <- cfa(HS.model2,std.lv=TRUE, data = HS)
summary(fit2factor2, standardized = TRUE)
```

Two Factor Confirmatory Factor Analysis

#correlated two factor solution, marker method

```
Covariances:

  visual ~~
    textual
 Estimate  Std.Err   z-value  P(>|z|)  Std.lv  Std.all

    0.414    0.074     5.562    0.000   0.461    0.461
```

covariance

correlation

Two Factor Confirmatory Factor Analysis

```
#uncorrelated two factor solution, marker method
HS.model3 <-"visual=~x1+x2+x3
        textual =~ x4 + x5 + x6
        visual~~0*textual"
fit2factor3 <- cfa(HS.model3,std.lv=TRUE, data = HS)
summary(fit2factor3, standardized = TRUE)
```

Model Fit Statistics

$$H_0 : \Sigma(\theta) = \Sigma$$

accept-support test

versus

$$H_1 : \Sigma(\theta) \neq \Sigma$$

reject-support test

$$\Sigma(\theta) = \mathbf{\Lambda \Psi \Lambda}' + \Theta_\epsilon \quad \text{versus} \quad \Sigma$$

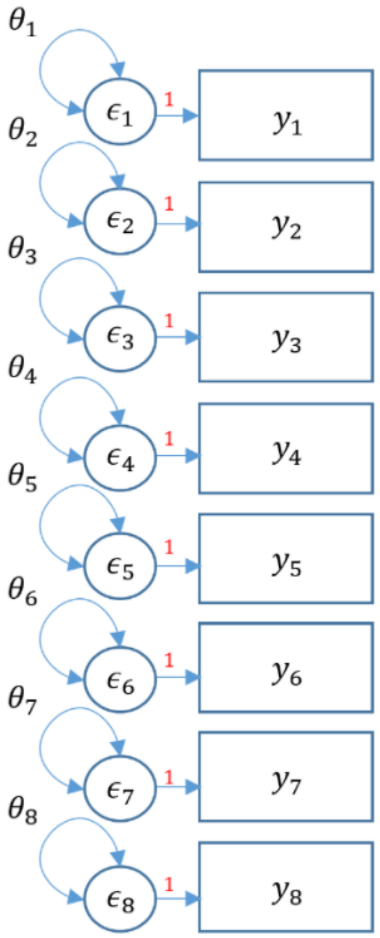
$$\Sigma(\hat{\theta}) \quad \text{versus} \quad S = \hat{\Sigma}$$

Model Fit Statistics

1. **Model chi-square** is the chi-square statistic we obtain from the maximum likelihood statistic (in lavaan, this is known as the Test Statistic for the Model Test User Model)
2. **CFI** is the *Comparative Fit Index* – values can range between 0 and 1 (values greater than 0.90, conservatively 0.95 indicate good fit)
3. **TLI** *Tucker Lewis Index* which also ranges between 0 and 1 (if it's greater than 1 it should be rounded to 1) with values greater than 0.90 indicating good fit. If the CFI and TLI are less than one, the CFI is always greater than the TLI.
4. **RMSEA** is the *root mean square error of approximation*
    - In **lavaan**, you also obtain a *p*-value of close fit, that the RMSEA < 0.05. If you reject the model, it means your model is not a close fitting model.

```
HS.model1 <-"visual=~x1+x2+x3
                textual =~ x4 + x5 + x6"
fit1factor2 <- cfa(HS.model, data = HS)
summary(fit1factor2, standardized = TRUE, fit.measures = TRUE)
```

Baseline



How many free parameters? Count 8.
How many degrees of freedom? Count 28.

8(9)/2 – 8.

Worst model.
Compare with saturated model.

Incremental versus absolute fit index

- For over-identified models, there are many types of fit indexes available to the researcher.

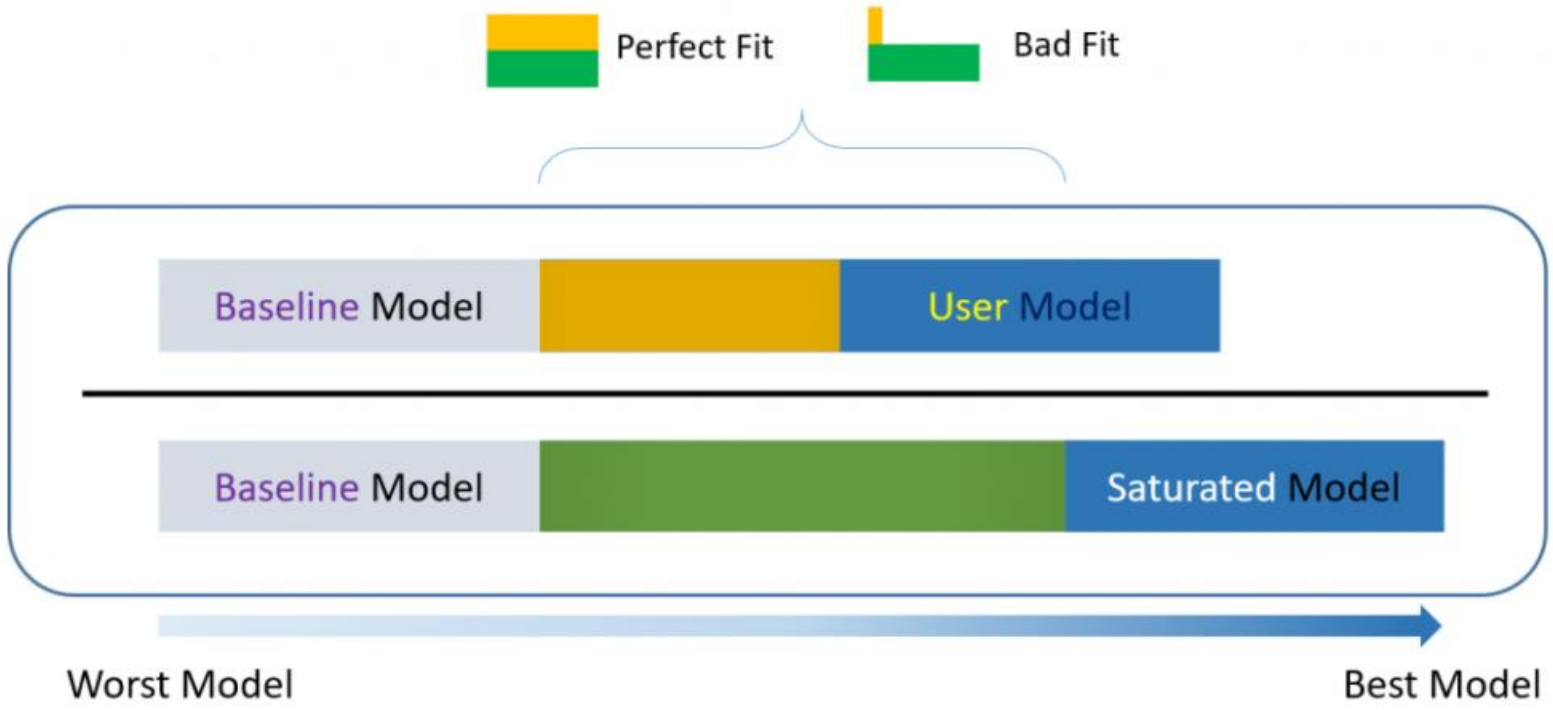
- Historically, model chi-square was the only measure of fit but in practice the null hypothesis was often rejected due to the chi-square's heightened sensitivity under large samples. To resolve this problem, *approximate* fit indexes that were not based on accepting or rejecting the null hypothesis were developed.

- Approximate fit indexes can be further classified into a) absolute and b) incremental or relative fit indexes.

- An **incremental fit index** assesses the ratio of the deviation of the user model from the worst fitting model (the baseline model) against the deviation of the saturated model from the baseline model. Conceptually, if the deviation of the user model is the same as the deviation of the saturated model (best fitting model), then the ratio should be 1. Alternatively, the more discrepant the two deviations, the closer the ratio is to 0 (see figure below). Examples of incremental fit indexes are the CFI and TLI.

- An **absolute fit index** on the other hand, does not compare the user model against a baseline model, but instead compares it to the observed data. An example of an absolute fit index is the RMSEA (see figure above).

Incremental versus absolute fit index



CFI (Comparative Fit Index)

The CFI or *comparative fit index* is a popular fit index as a supplement to the model chi-square. Let $\delta = \chi^2 - df$ where df is the degrees of freedom for that particular model. The closer δ is to zero, the more the model fits the data. The formula for the CFI is:

$$CFI = \frac{\delta(\text{Baseline}) - \delta(\text{User})}{\delta(\text{Baseline})}$$

Model Test User Model:

Test statistic	24.361
Degrees of freedom	8
P-value (Chi-square)	0.002

Model Test Baseline Model:

Test statistic	668.643
Degrees of freedom	15
P-value	0.000

$\delta(User) = 24.361 - 8 = 16.361$

$\delta(Baseline) = 668.643 - 15 = 653.643$

$$CFI = \frac{653.643 - 16.361}{653.643} = 0.975$$

User Model versus Baseline Model:

Comparative Fit Index (CFI)	0.975
Tucker-Lewis Index (TLI)	0.953

TLI (Tucker Lewis Index)

The Tucker Lewis Index is also an incremental fit index that is commonly outputted with the CFI in popular packages such as Mplus and in this case **lavaan**. The term used in the TFI is the **relative chi-square** (a.k.a. normed chi-square) defined as $\frac{\chi^2}{df}$. Compared to the model chi-square, *relative* chi-square is less sensitive to sample size. To understand relative chi-square, we need to know that the expected value or mean of a chi-square is its degrees of freedom (i.e., $E(\chi^2(df)) = df$).

$$TLI = \frac{\chi^2(\text{Baseline})/df(\text{Baseline}) - \chi^2(\text{User})/df(\text{User})}{\chi^2(\text{Baseline})/df(\text{Baseline}) - 1}$$

$$CFI = \frac{\min\left(\frac{668.643}{15}, 1\right) - \min(\frac{24.361}{8}, 1)}{\min\left(\frac{668.643}{15}, 1\right) - 1} = \frac{44.576 - 3.045}{44.576 - 1} = \frac{41.531}{43.576} = 0.953$$

```
User Model versus Baseline Model:

  Comparative Fit Index (CFI)                    0.975
  Tucker-Lewis Index (TLI)                       0.953
```

The CFI is always greater than the TLI. CFI pays a penalty of one for every parameter estimated. Because the TLI and CFI are highly correlated, only one of the two should be reported.

RMSEA

The *root mean square error of approximation* is an **absolute** measure of fit because it does not compare the discrepancy of the user model relative to a baseline model like the CFI or TLI. Instead, RMSEA defines δ as the non-centrality parameter which measures the degree of misspecification. Recall from the CFI that $\delta = \chi^2 - df$ where df is the degrees of freedom for that particular model. The greater the δ the more misspecified the model.

$$RMSEA = \sqrt{\frac{\delta}{df(n-1)}}$$

where n is the total number of observations. The cutoff criteria as defined in Kline (2016, p.274-275)

- ≤ 0.05 (*close-fit*)
- between $.05$ and $.08$ (*reasonable approximate fit*, fails close-fit but also fails poor-fit)
- $>= 0.10$ (*poor-fit*)

$$\delta(User) = 24.361 - 8 = 16.361$$

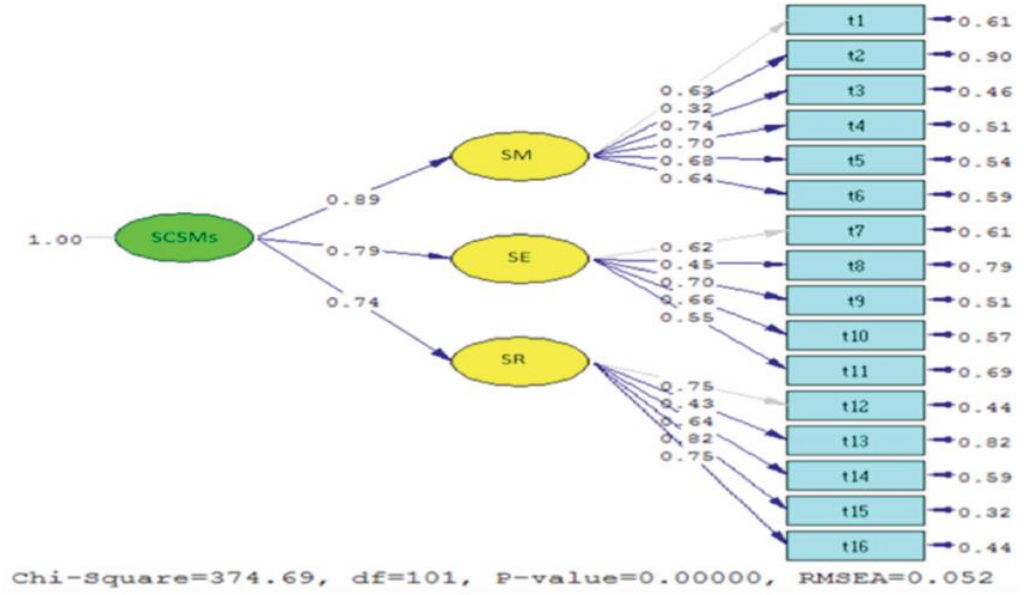
$$RMSEA = \sqrt{\frac{16.361}{8(301)}} = 0.082$$

```
Root Mean Square Error of Approximation:

  RMSEA                                          0.082
  90 Percent confidence interval - lower         0.046
  90 Percent confidence interval - upper         0.121
  P-value RMSEA <= 0.05                          0.067
```

Second-Order CFA

- Suppose the Principal Investigator believes that the correlation between **Visual** and **Textual** are first-order factors is caused more by the second-order factor, overall **mental ability**.
- In order to understand the model, we have to understand **endogenous** and **exogenous** factors.
- An endogenous factor is a factor that is being predicted by another factor (or variable in general).
- An exogenous factor is a factor that is not being predicted by another.
- The main difference is that endogenous factors now have a residual variance as it is not being predicted by another latent variable known as ζ. The residual variance is essentially the variance of ζ, which we classify here as ψ.



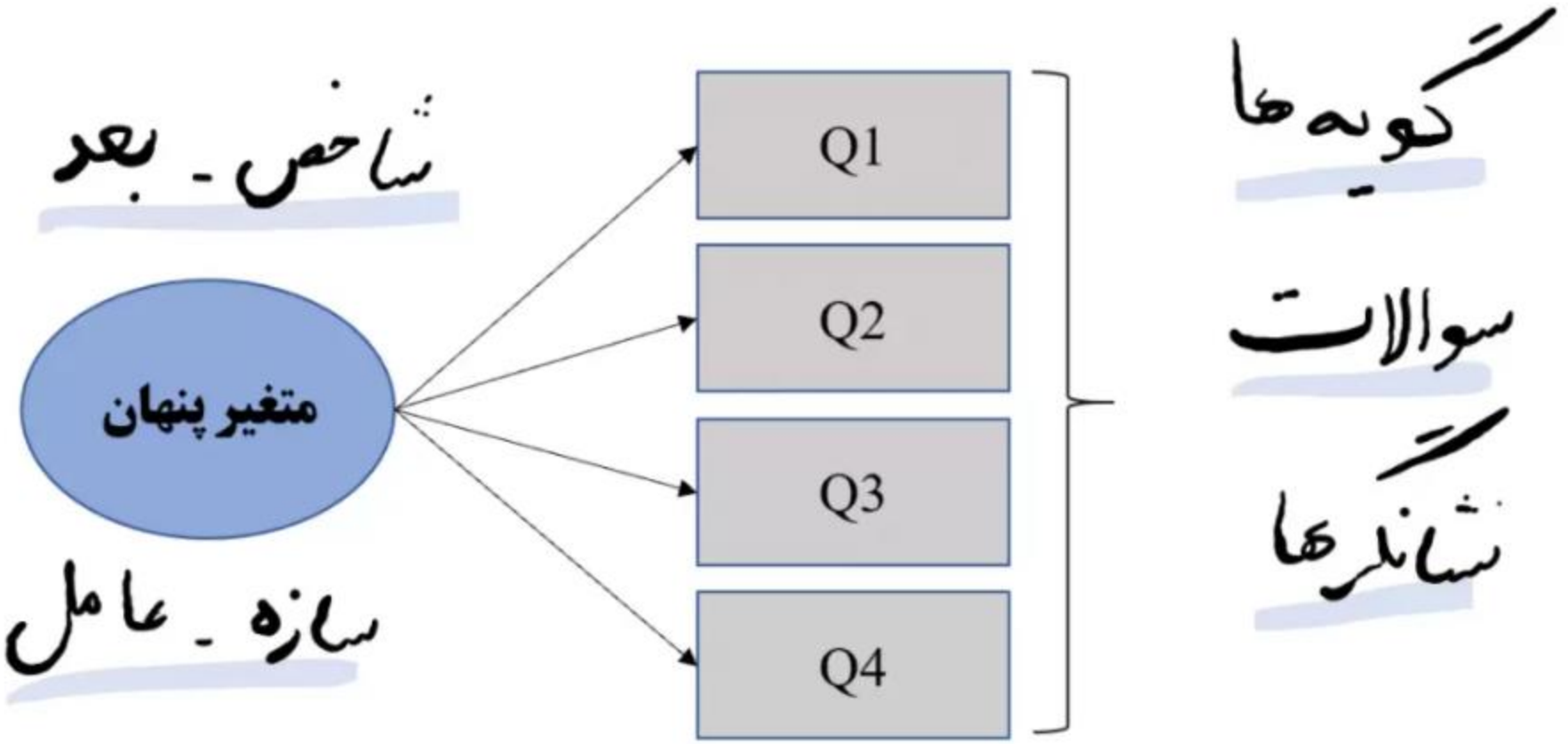
Second-Order CFA

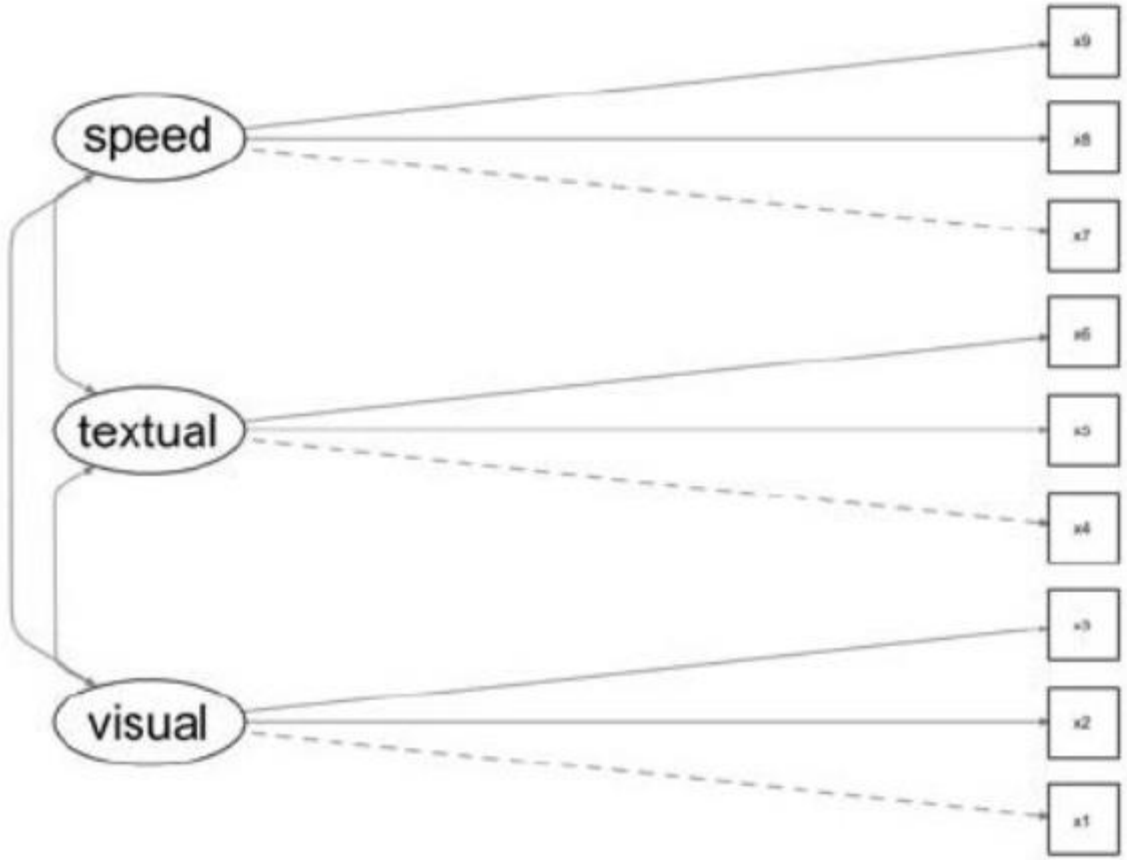
```
##second order three factor solution, marker method
HS.model5 <- "visual=~x1+x2+x3
          textual =~ x4 + x5 + x6
          x2 ~~ x3
          mental =~ 1*visual + 1*textual
           mental ~~  mental"
secondorder <- cfa(HS.model5, data=HS)
summary(secondorder,fit.measures=TRUE,standardized=TRUE)
```

Assumptions of the factor analysis model

In traditional confirmatory factor analysis or structural equation modeling, the

- mean of the intercepts is zero $E(\tau) = 0$ (not tenable, this is no longer true with modern full information CFA/SEM, see Kline 2016)
- mean of the factor is zero $E(\eta) = 0$
- mean of the residual is zero $E(\epsilon) = 0$
- covariance of the factor with the residual is zero $Cov(\eta, \epsilon) = 0$


شاخص - بعد
متغیر پنهان
سازه - عامل
Q1
Q2
Q3
Q4
گویه‌ها
سوالات
نشانگرها


speed
textual
visual
x9
x8
x7
x6
x5
x4
x3
x2
x1

Model Fit Indices

- There are two types of model fit indices available for CFA;
- **Global:** Measure the global recovery of empirical observations without considering the mean and covariance structure.
  - **Absolute:** Absolute fit indices assess the overall theoretical model against the observed data.
    - Chi-square (χ^2) statistic
    - Goodness-of-fit index (GFI)
    - Adjusted GFI
    - Root mean square error of approximation (RMSEA)
    - Root mean square residual and standardized root mean square residual (SRMR)
  - **Incremental** (also known as comparative or relative): represent the improved fit for the model compared to the assumption of independence of variables.
    - Comparative fit index (CFI),
    - Normed-fit index (NFI)
    - Non-normed fit index
  - **Parsimony fit indices:** Parsimonious fit indices aim to address this issue by adding a penalty for model complexity.
    - Parsimony goodness-of-fit index
    - Parsimony normed fit index
- **Local fit indices**: examine model components including but not limited to factor correlations, inter-item residual covariance, and suggested model re-specification statistics.

The lowest possible RMSEA is 0. Values < .05 are considered indicative of close fit. Values up to .08 are considered acceptable (Pituch & Stevens, 2016).

$$\text{RMSEA} = \sqrt{\frac{\chi^2 - \text{df}_M}{\text{df}_M \times (N-1)}}$$

The pclose is a test of whether the model departs significantly from one that is a close-fit to the data (i.e., RMSEA< or = .05).

The CFI and TLI are both incremental fit indices. Values > .95 for these indices indicate very good fit (Schumaker & Lomax, 2016). Values .90 or above are considered evidence of acceptable fit (Pituch & Stevens, 2016).

$$\text{CFI} = 1 - \frac{\chi^2_M - \text{df}_M}{\chi^2_B - \text{df}_B} \qquad \text{TLI} = \left[\frac{\chi^2_B}{\text{df}_B} - \frac{\chi^2_M}{\text{df}_M}\right] / \left[\frac{\chi^2_B}{\text{df}_B} - 1\right]$$

SRMR values up to .05 are considered indicative of a close-fitting model. Values between .05 up to .10 suggest acceptable fit (Pituch & Stevens, 2016).

The SRMR is the ratio of the sum of the squared differences between the correlations for the observed variable and the correlations implied by our model divided by the number of variances and covariances. This is given by the formula below

$$\text{SRMR} = \sqrt{\frac{\sum_{i<=j}(r_{i,j} - \rho_{i,j})^2}{v(v+1)/2)}}$$

- The criterion of Fornell-Larcker (1981) has been commonly used to assess the degree of shared variance between the latent variables of the model.
- According to this criterion, the convergent validity of the measurement model can be assessed by the Average Variance Extracted (AVE) and Composite Reliability (CR).
- AVE measures the level of variance captured by a construct versus the level due to measurement error, values above 0.7 are considered very good, whereas, the level of 0.5 is acceptable.
- CR is a less biased estimate of reliability than Chonbachs Alpha, the acceptable value of CR is 0.7 and above.

References

- Kline suggests that at a minimum the following indices should be reported and assessed in combination: chi-square; RMSEA; CFI; and SRMSR.
- Kline, R. B. (2015). *Principles and practice of structural equation modeling*. Guilford publications.
- Confirmatory Factor Analysis (CFA) in R with lavaan (ucla.edu)
- Hair JF, Black WC, Babin BJ, Anderson RE, Tatham RL (2006). Multivariate data analysis. Pearson Prentice Hall Upper Saddle River