Universiteit Utrecht

Theory Construction and Statistical Modeling

Welcome!

Today

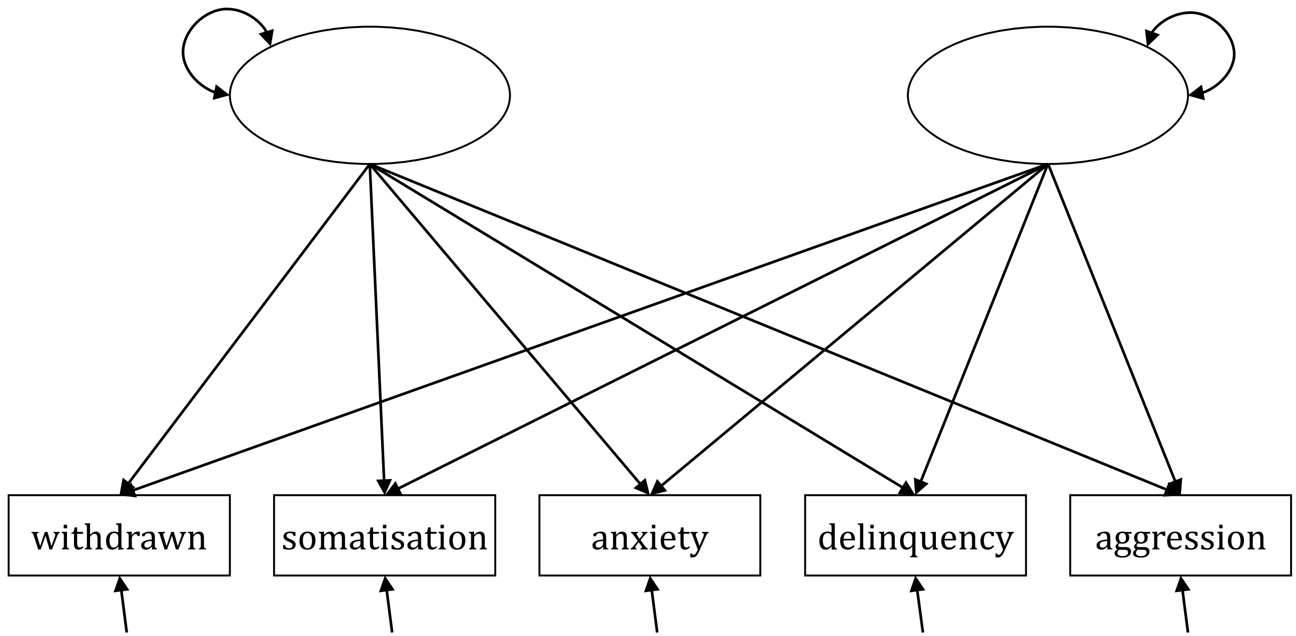
- Intro to Confirmatory Factor Analysis
  - EFA vs CFA
  - Giving Latent Variables a scale
- Model Fit 1
  - Complexity and Degrees of Freedom
  - The Chi-Square Test $\chi 2$
- Model Fit 2
  - Alternative Fit Measures
- Extensions
  - Second order factors
  - Means and intercepts

Confirmatory factor analysis

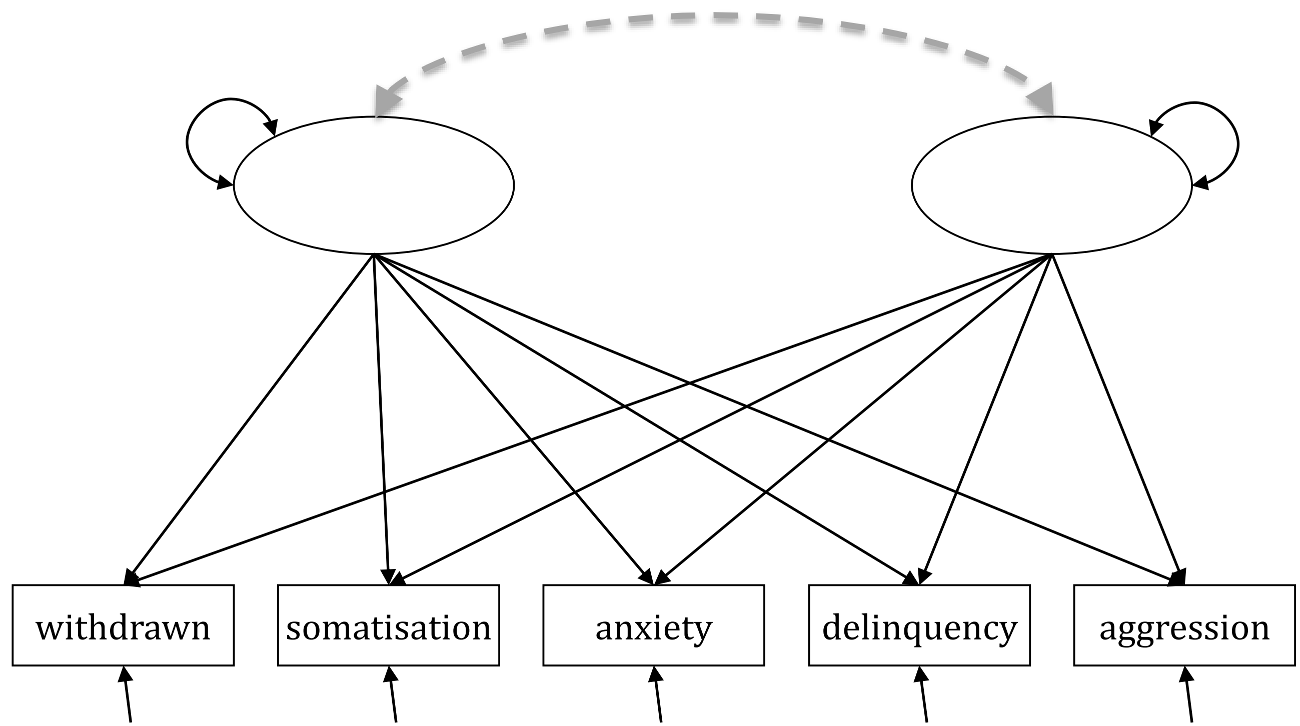
- EFA vs CFA

Exploratory factor analysis	Confirmatory factor analysis
Theory development **Inductive**	Theory testing **Deductive**
# factors a posteriori Data-driven	# factors a priori Theory-driven
All variables load on all factors	Not all variables load on all factors (usually only one)
Rotation needed for interpretation	No rotation needed
SPSS / R psych	SEM-programs (lavaan)

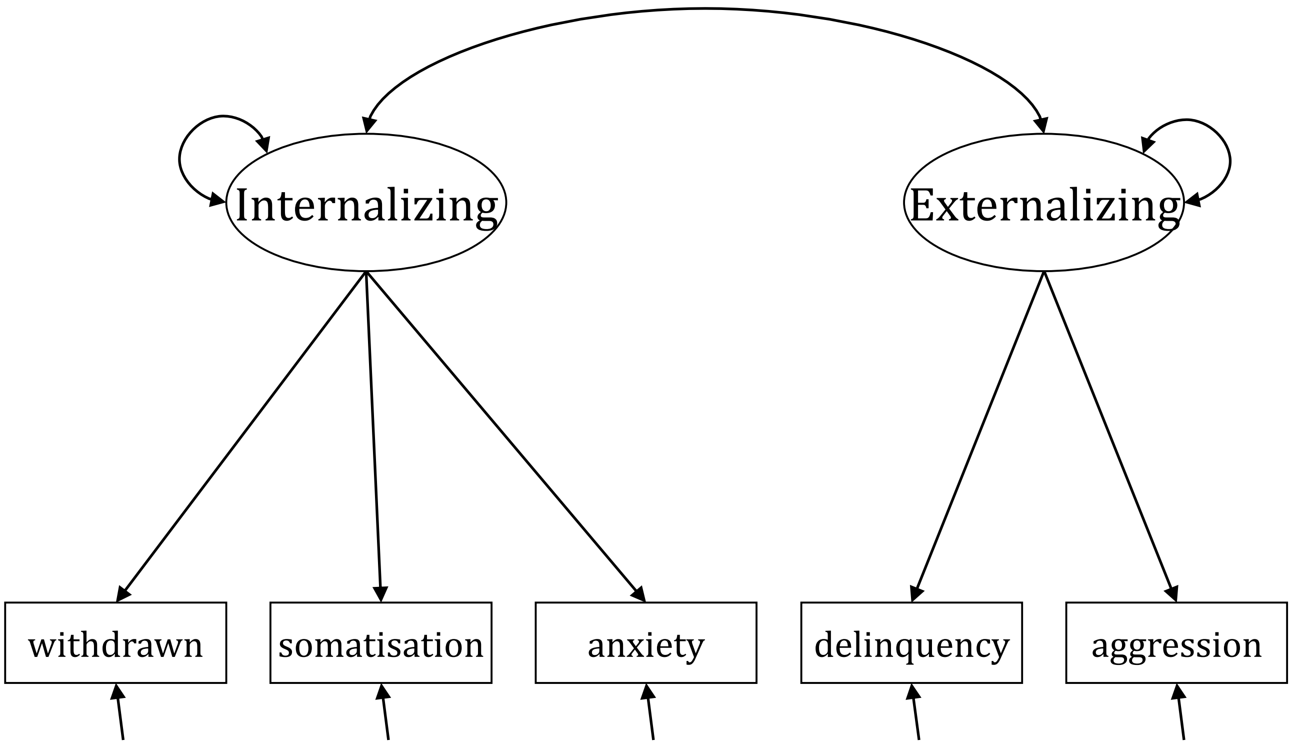
Exploratory factor model



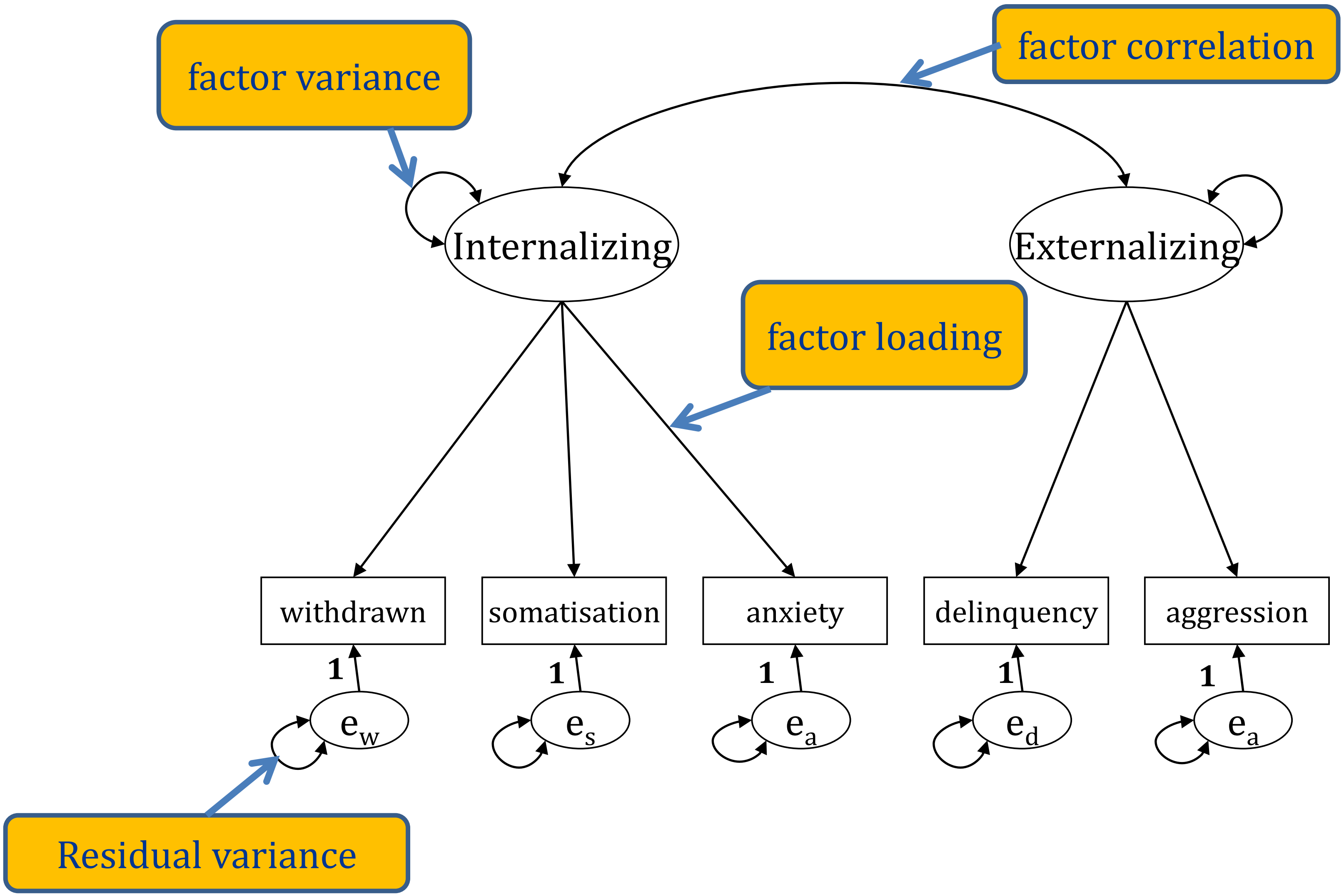
Exploratory factor model



Confirmatory factor model



Confirmatory factor model



Scaling

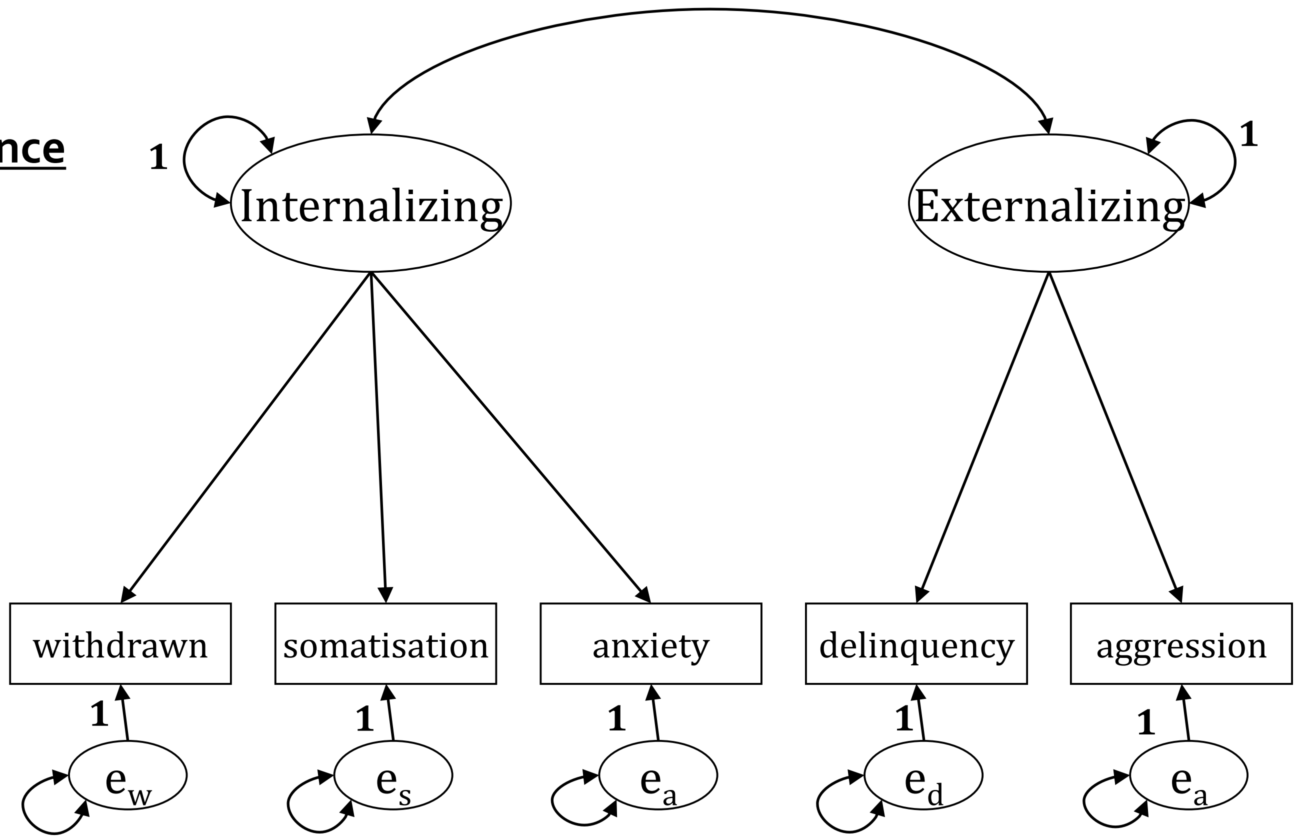
- What scale does your unmeasured latent variable have?
  - Height in cm has a scale,
  - Happiness from 1-7 has a scale
- But what is the scale of something you did not measure?

- Two choices:
  - Fix the factor variance at 1
    (+1 SD on latent variable associated with +1*factor loading on observed variable)
  - Fix one factor loading at 1
    Scale of latent variable is linked to scale of that observed variable

Scaling

Each factor needs to be assigned a **<u>scale</u>**:
 For example by **<u>fixing the variance at 1</u>**.

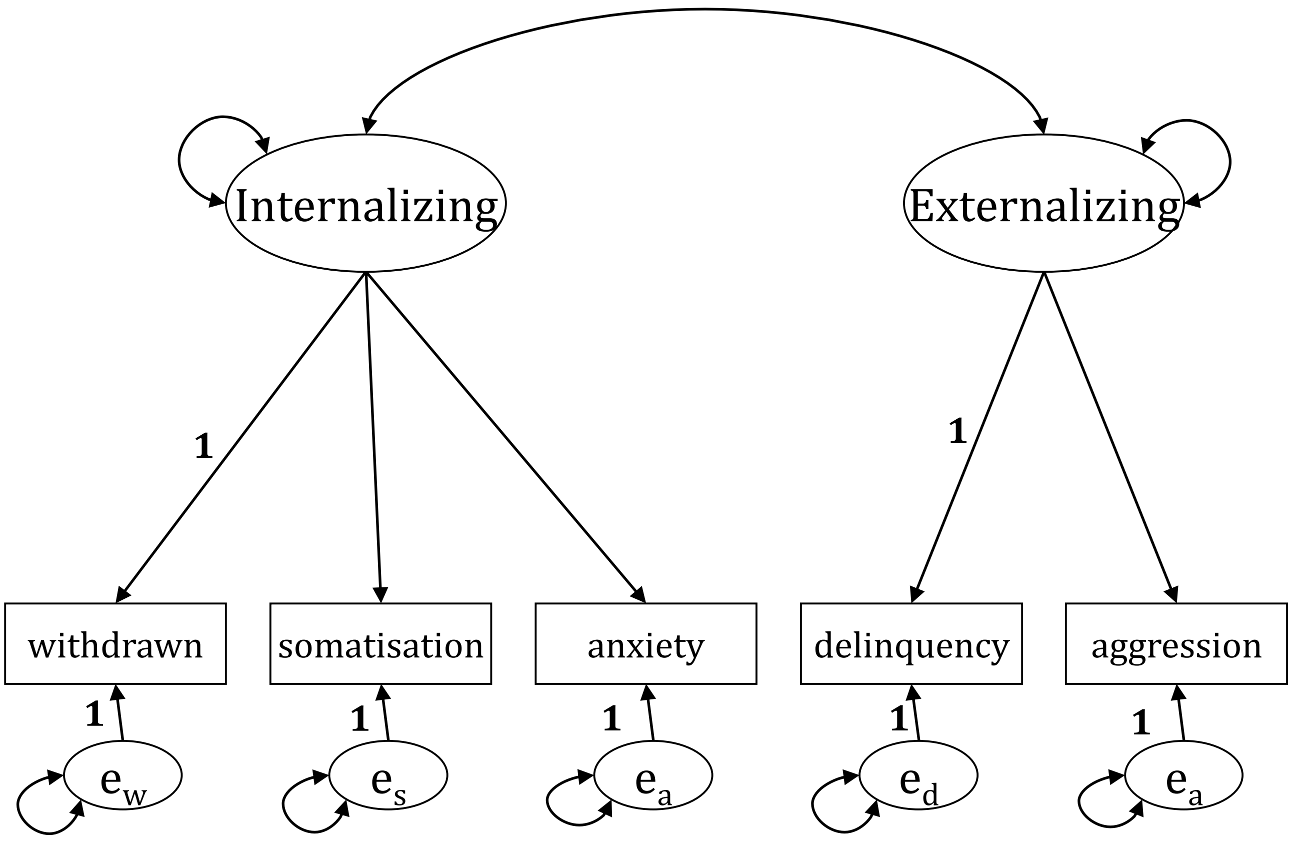
Number of parameters in model: 11
5 variances
5 regression coefficients
1 covariance



Scaling

Or by **fixing 1 factor loading** per factor at 1

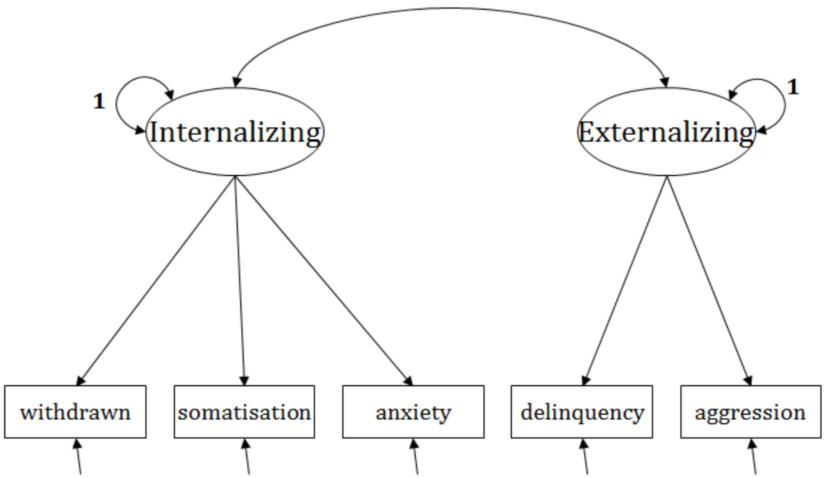
Number of parameters in model: 11
7 variances
3 regression coefficients
1 covariance



Default in lavaan

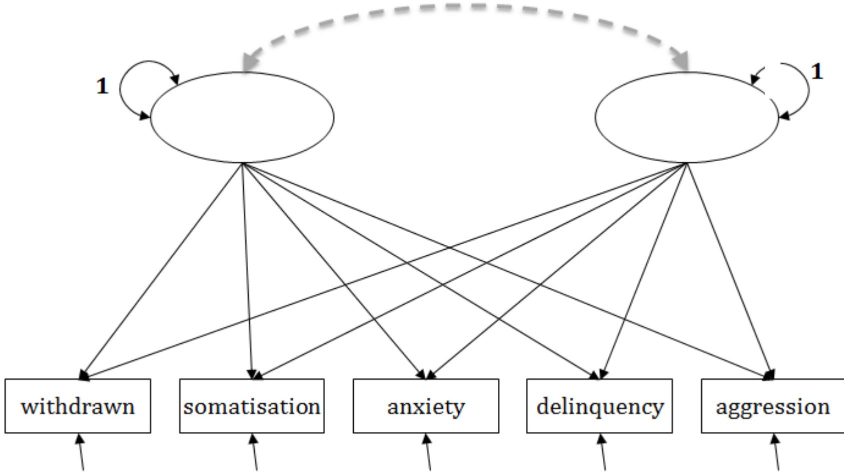
Technical Intermezzo

- CFA model must be **identified** to be fit, by having $\mathbf{df} \geq \mathbf{0}$
  - df = number of observed variances and covariances – number of parameters in model
  - Df = number of known pieces of information – number of estimated pieces of information
  - Latent variables must be given a **scale** by **fixing** certain parameter
  - Remember example:
    a = 5 – 2 is **identified**
    a = 5 – b is **not identified**



Technical Intermezzo

- EFA is identified by other restrictions
  - Factor variances fixed to 1
  - Factor covariances fixed to 0
  - Functions of multiple loadings fixed to a constant



Model complexity

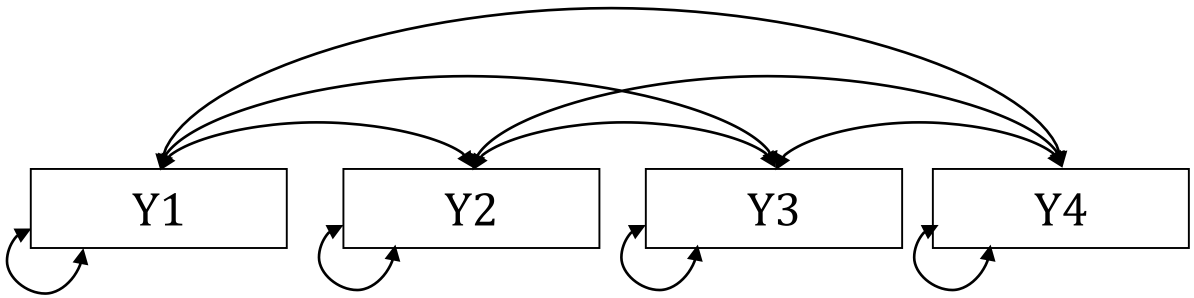
The model explains the covariances between observed variables. A good model is:

- Simple
- A good description of reality

- The larger the degrees of freedom, the more simple the model (good). But… the worse the model will fit to the data.

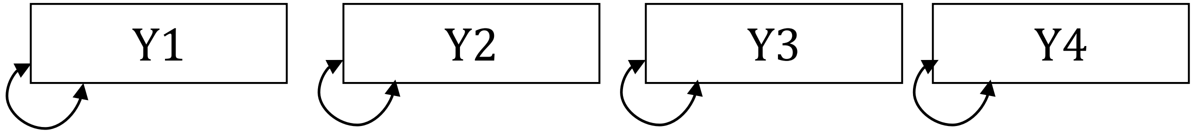
Model complexity

- Perfectly fitting (but very complex) model:



(Saturated model)

- Very simple (but ill fitting) model:



(Independence model)

Degrees of freedom

- Keep track of the balance between known and estimated quantities
- Degrees of freedom (df) = p – q
- p: Observed pieces of information
- q: Unknown pieces of information
- An **identified** model has
  fewer parameters (q) than
  observed variances and covariances (p)

Observations

- Input for SEM is a **variance/covariance** (vcov) matrix
- Number of observations is the number of unique elements in the vcov matrix
- Lower triangular formula:
  p = nvar*(nvar+1) / 2

	Y1	Y2	Y3	Y4
Y1	4.5			
Y2	2.1	3.9		
Y3	1.9	2.6	4.1	
Y4	2.8	2.5	2.0	4.8

Parameters

These are parameters in a SEM:

- Variances of exogenous (predictor) variables
- Covariances among exogenous (predictor) variables
- Regression (or covariance) between exogenous (predictor) and endogenous (outcome) variables
- Residual variances
- Covariances between residual variances

5*6 /2 = 15

var = 2 (y1, y2)

res.var = 3

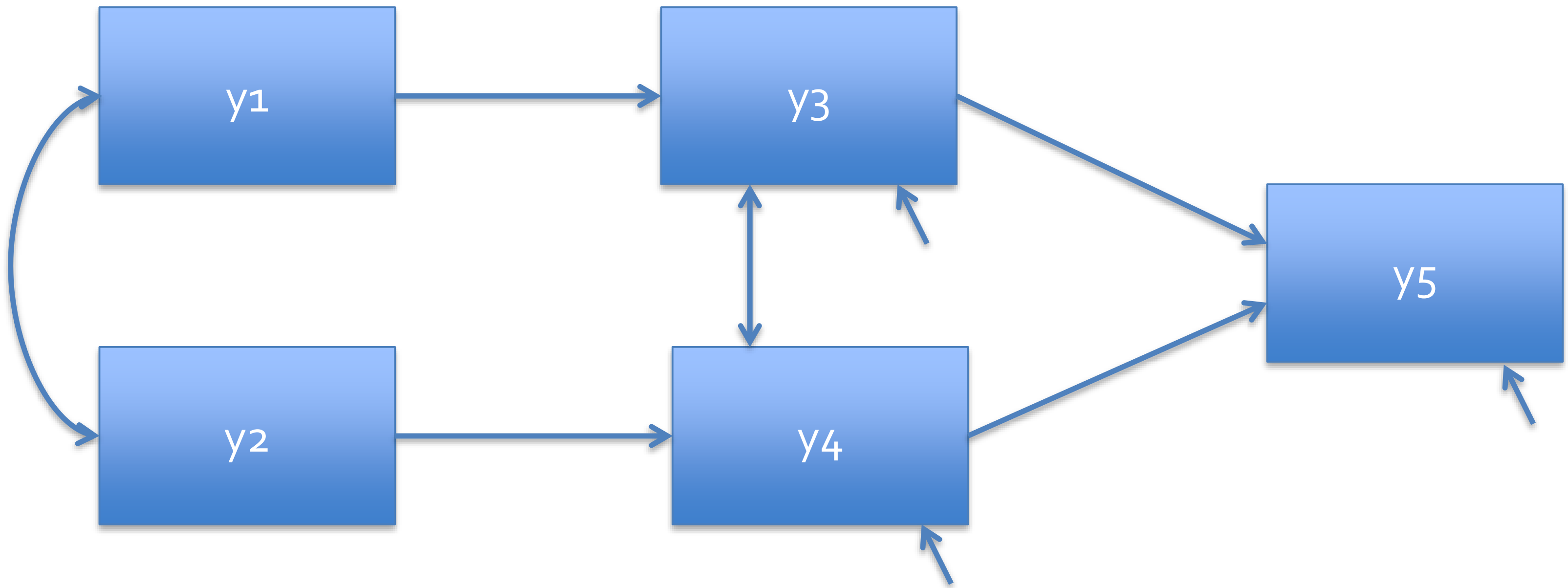
cov = 2

reg = 4

DF = 15 – 11

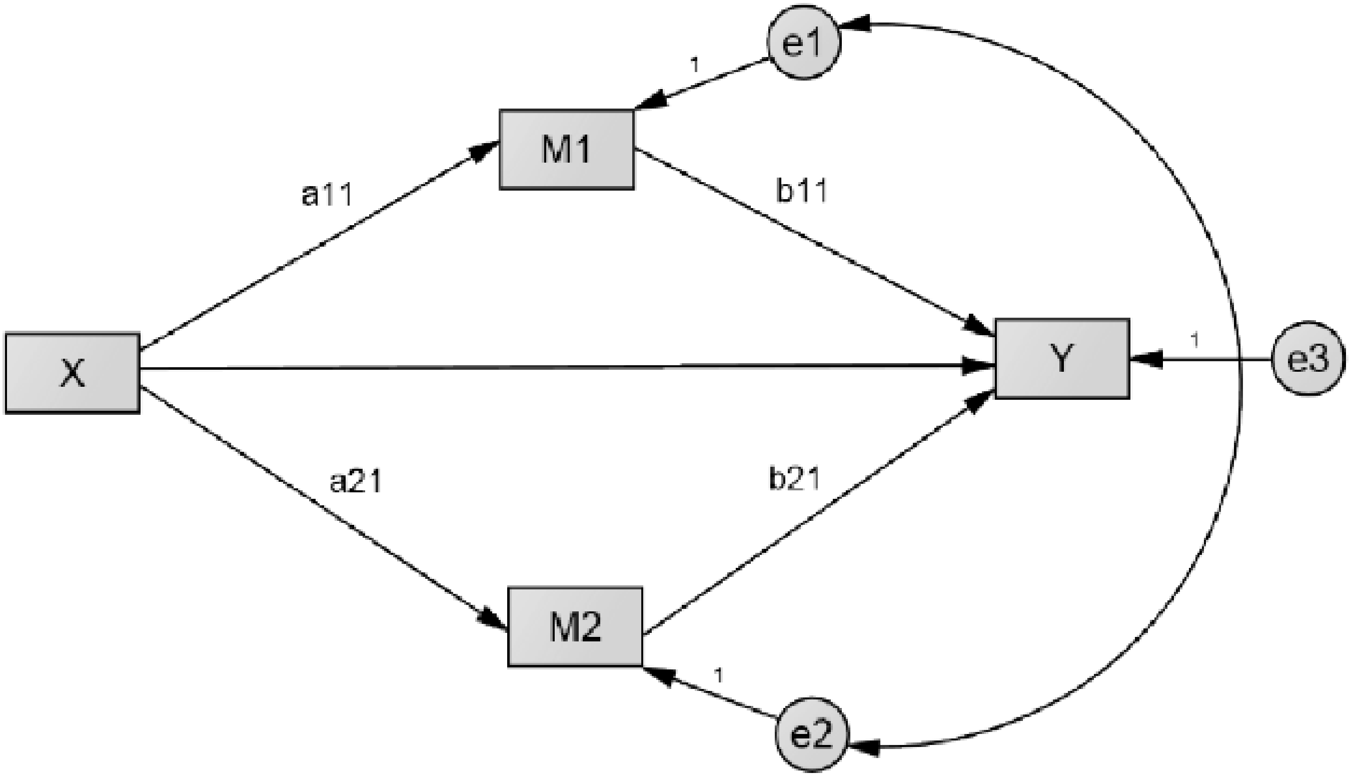
= 4

DF?



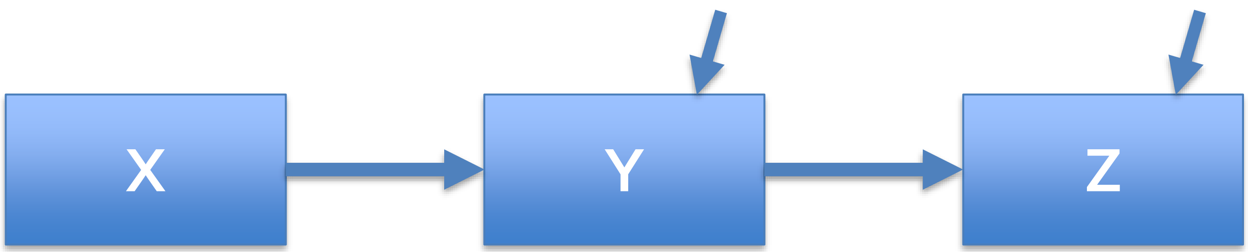
DF?

4*5 /2 =10
1+ 3+ 1+ 5 = 10
10-10 = 0



Model fit

How well does the theoretical model fit the data



Model *implies* a covariance structure: covariance between X and Z is lower than other 2

Model fit: How close is the model-implied vcov matrix to the observed vcov matrix?

Maximum likelihood estimation

- k: Number of observed variables
- $\mathbf{S}$: Sample covariance matrix
- $\mathbf{\Sigma}$: Model-implied covariance matrix

The objective function is given by:

- $F_{ML} = \log|\mathbf{\Sigma}| - \log|\mathbf{S}| + \mathrm{trace}(\mathbf{S}\mathbf{\Sigma}^{-1}) - k$

And the model Chi square:

- $\chi^2 = (N-1)\, F_{ML}$

And the df are p - q

Chi-square measure of fit

- $\chi^2 = (N-1)\, F_{ML}$
- Asymptotically chi-square with df = p – q
- Null hypothesis: $\mathbf{\Sigma}_{population} = \mathbf{\Sigma}_{model}$
- Alternative hypothesis: $\mathbf{\Sigma}_{population} \neq \mathbf{\Sigma}_{model}$
- We take **S** to be an estimator of $\mathbf{\Sigma}_{population}$
  - Rejecting the null hypothesis ($p<.05$) means our model fits the data **badly**
  - Failing to reject the null hypothesis ($p>.05$) means the model fits the data well

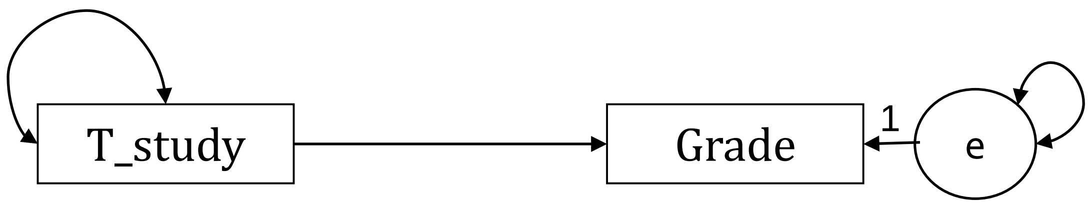
Chi-square measure of fit

- $\chi^2 = (N\text{-}1)\, F_{ML}$
- Asymptotically chi-square with df = p – q

- Null hypothesis: $\boldsymbol{\Sigma}_{population} = \boldsymbol{\Sigma}_{model}$
- Alternative hypothesis: $\boldsymbol{\Sigma}_{population} \neq \boldsymbol{\Sigma}_{model}$

- Does our model fit the data significantly worse than the **saturated model?**

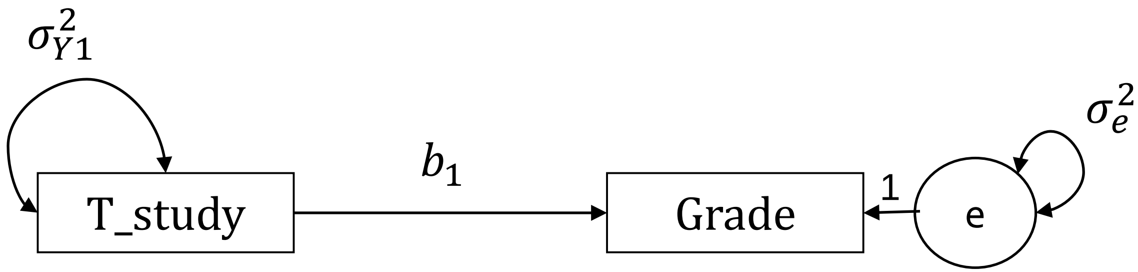
Example 1



	T_study	Grade
T_study	s_{Y1}^2	
Grade	s_{Y1Y2}	s_{Y2}^2

$\boldsymbol{S}$: Observed Covariance Matrix

Example 1

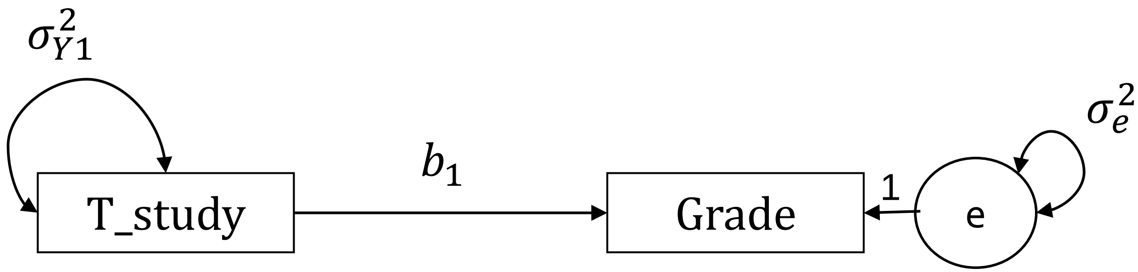


$$Grade_i = b_1 T_{study_i} + e_i \qquad e_i \sim N(0, \sigma_e^2)$$

	T_study	Grade
T_study	s^2_{Y1}	
Grade	s_{Y1Y2}	s^2_{Y2}

$\boldsymbol{S}$: Observed Covariance Matrix

Example 1



$$Grade_i = b_1 T_{study_i} + e_i \qquad e_i \sim N(0, \sigma_e^2)$$

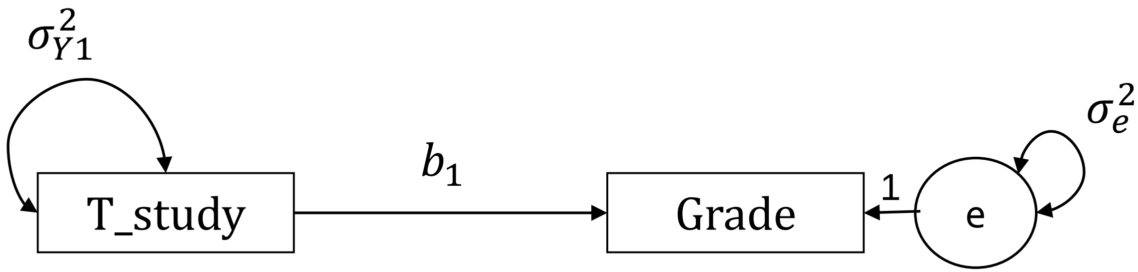
	T_study	Grade
T_study	s_{Y1}^2	
Grade	s_{Y1Y2}	s_{Y2}^2

S: Observed Covariance Matrix

	T_study	Grade
T_study	σ_{Y1}^2	
Grade	b_1	$b_1^2\sigma_{Y1}^2 + \sigma_e^2$

Σ: Modelled Covariance Matrix

Example 1



$$Grade_i = b_1 T_{study_i} + e_i \qquad e_i \sim N(0, \sigma_e^2)$$

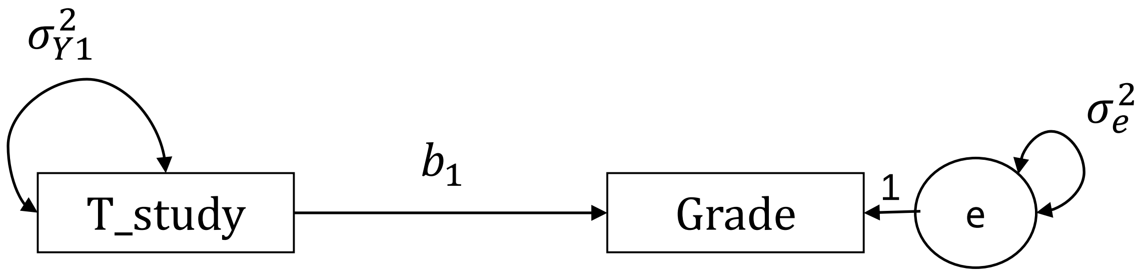
	T_study	Grade
T_study	1	
Grade	.3	1.5

S: Observed Covariance Matrix

	T_study	Grade
T_study	σ^2_{Y1}	
Grade	b_1	$b_1^2\sigma^2_{Y1} + \sigma_e^2$

Σ: Modelled Covariance Matrix

Example 1



$\sigma^2_{Y1} = 1$
$b_1 = .3$
$\sigma^2_e = 1.41$

$$Grade_i = b_1 T_{study_i} + e_i \qquad e_i \sim N(0, \sigma^2_e)$$

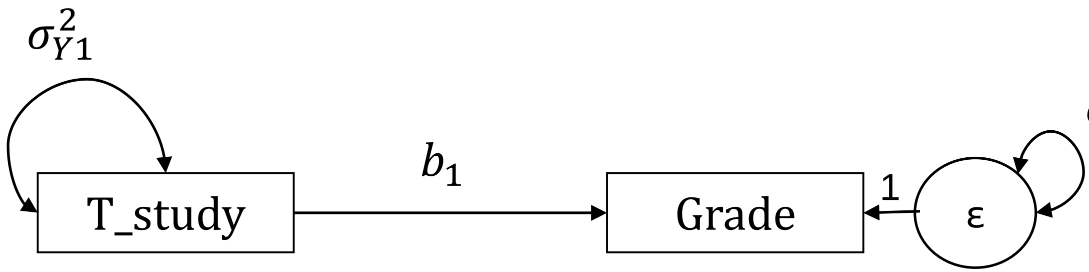
	T_study	Grade
T_study	1	
Grade	.3	1.5

S: Observed Covariance Matrix

	T_study	Grade
T_study	σ^2_{Y1}	
Grade	b_1	$b_1^2 \sigma^2_{Y1} + \sigma^2_e$

Σ: Modelled Covariance Matrix

Example 1



$\sigma^2_{Y1} = 1$
$b_1 = .3$
$\sigma^2_e = 1.41$

$$Grade_i = b_1 T_{study_i} + e_i \qquad e_i \sim N(0, \sigma^2_e)$$

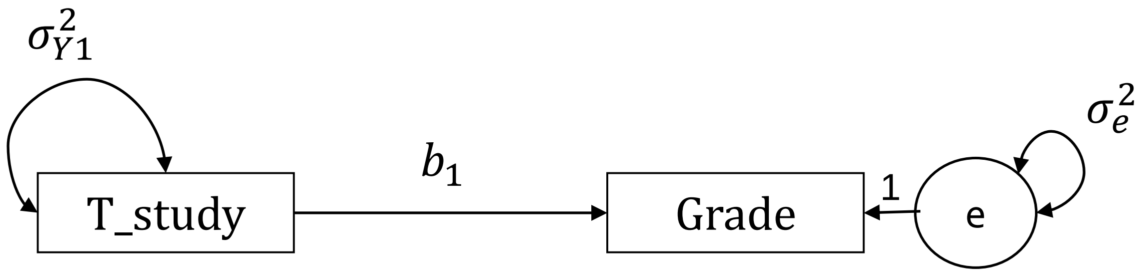
	T_study	Grade
T_study	1	
Grade	.3	1.5

S: Observed Covariance Matrix

	T_study	Grade
T_study	1	
Grade	.3	.09 + 1.41

Σ: Modelled Covariance Matrix

Example 1



$\sigma^2_{Y1} = 1$
$b_1 = .3$
$\sigma^2_e = 1.41$

$$Grade_i = b_1 T_{study_i} + e_i \qquad e_i \sim N(0, \sigma^2_e)$$

	T_study	Grade
T_study	1	
Grade	.3	1.5

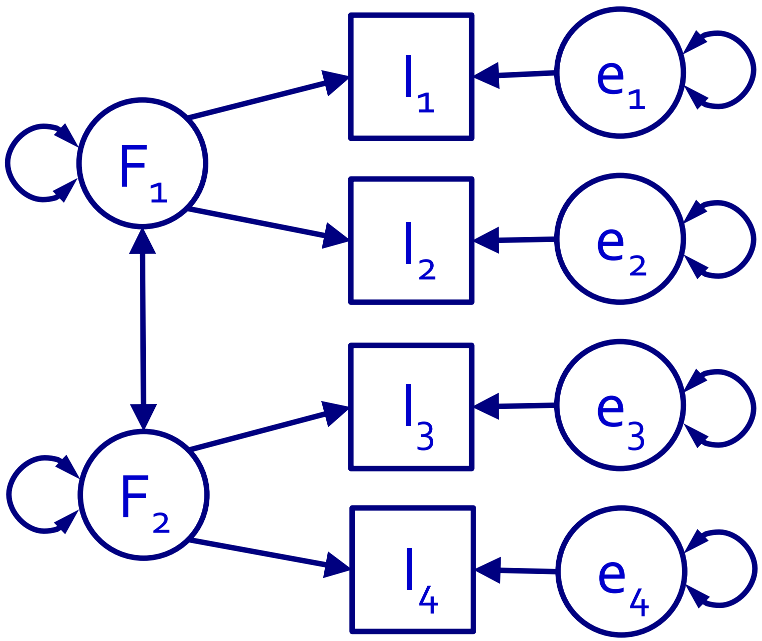
S: Observed Covariance Matrix

	T_study	Grade
T_study	1	
Grade	.3	.09 + 1.41

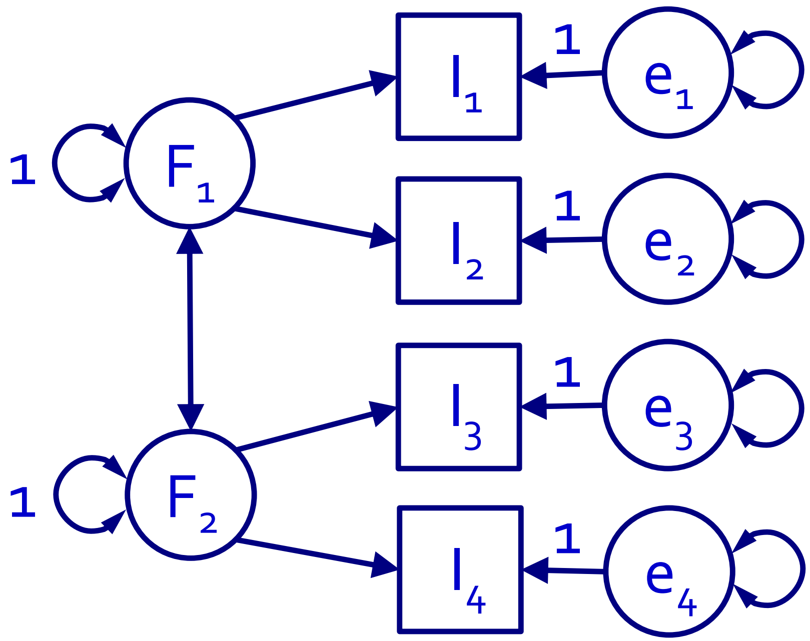
Σ: Modelled Covariance Matrix

Perfect fit, so 0 degrees of freedom!

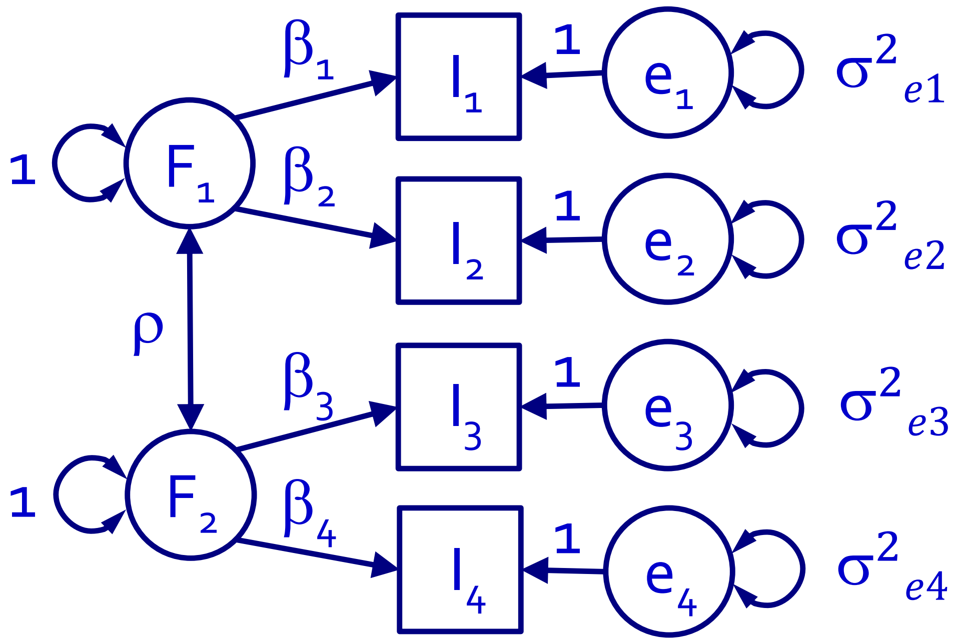
Example 2



Example 2



Example 2



nvar = 4 p = nvar(nvar + 1)/2 = 10 q = 9 df = p – q = 1

Example 2

Sample covariance matrix:

$$S_N = \begin{array}{c} \\ I_1 \\ I_2 \\ I_3 \\ I_4 \end{array} \begin{array}{c} \begin{array}{cccc} I_1 & I_2 & I_3 & I_4 \end{array} \\ \begin{bmatrix} 3.474 & 2.826 & 0.984 & 0.741 \\ 2.826 & 3.745 & 0.971 & 0.817 \\ 0.984 & 0.971 & 3.136 & 2.199 \\ 0.741 & 0.817 & 2.199 & 2.005 \end{bmatrix} \end{array}$$

$N = 543$

Example 2

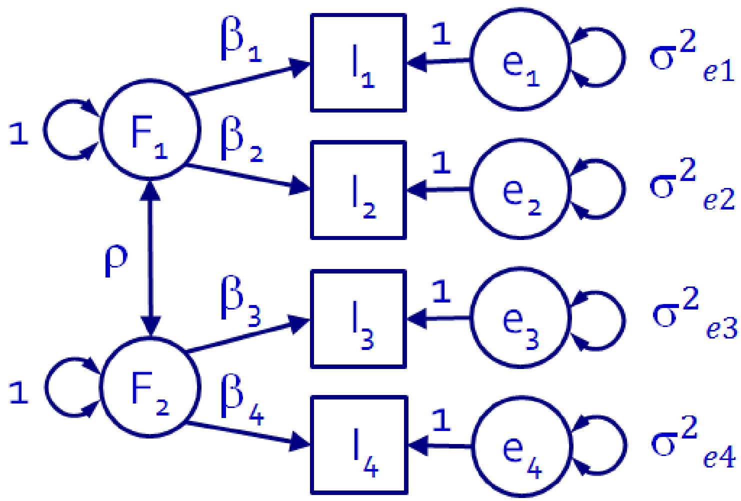
Model implied covariance matrix:

Algorithm ('tracing rules') based on path model which can be used to obtain expression for Σ:

http://ibgwww.colorado.edu/twins2002/cdrom/HTML/BOOK/node78.htm

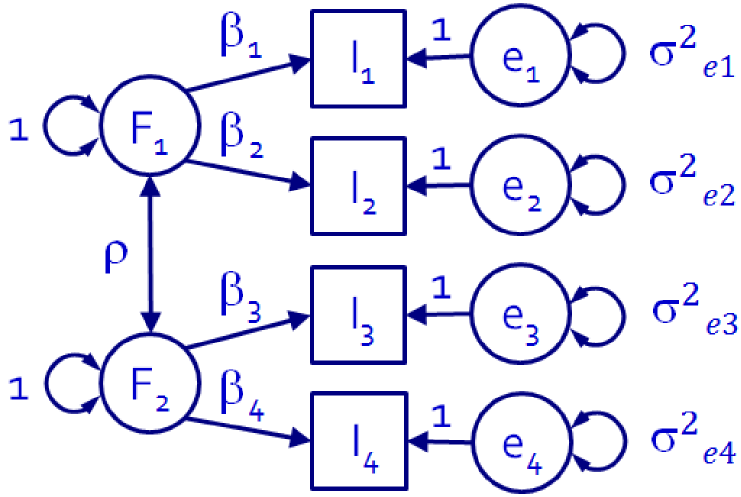
$$
\Sigma_{\text{model}} = \begin{array}{c c}
 & \begin{array}{cccc} I_1 & I_2 & I_3 & I_4 \end{array} \\
\begin{array}{c} I_1 \\ I_2 \\ I_3 \\ I_4 \end{array} &
\begin{bmatrix}
\beta_1\beta_1 + \sigma^2{}_{e1} & \beta_1\beta_2 & \beta_1\beta_3\rho & \beta_1\beta_4\rho \\
\beta_2\beta_1 & \beta_2\beta_2 + \sigma^2{}_{e2} & \beta_2\beta_3\rho & \beta_2\beta_4\rho \\
\beta_3\beta_1\rho & \beta_3\beta_2\rho & \beta_3\beta_3 + \sigma^2{}_{e3} & \beta_3\beta_4 \\
\beta_4\beta_1\rho & \beta_4\beta_2\rho & \beta_4\beta_3 & \beta_4\beta_4 + \sigma^2{}_{e4}
\end{bmatrix}
\end{array}
$$

Example 2



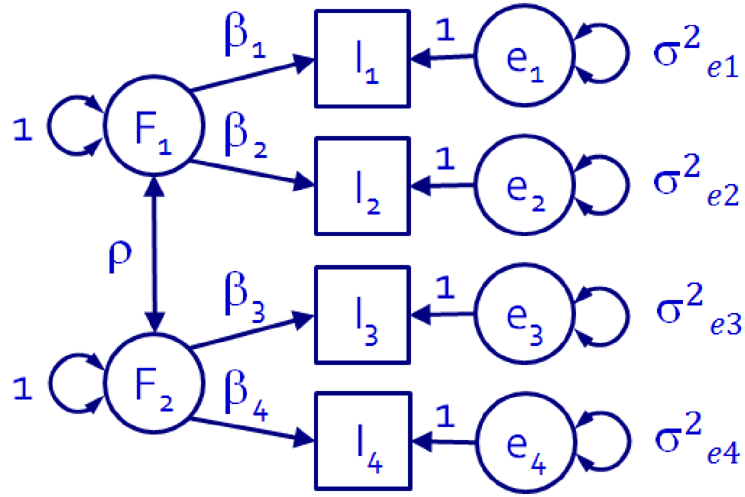
$$\Sigma_{model} = \begin{array}{c} \\ I_1 \\ I_2 \\ I_3 \\ I_4 \end{array} \begin{array}{c} \begin{array}{cccc} I_1 & I_2 & I_3 & I_4 \end{array} \\ \begin{bmatrix} \beta_1\beta_1 + \sigma^2_{e1} & \beta_1\beta_2 & \beta_1\beta_3\rho & \beta_1\beta_4\rho \\ \beta_2\beta_1 & \beta_2\beta_2 + \sigma^2_{e2} & \beta_2\beta_3\rho & \beta_2\beta_4\rho \\ \beta_3\beta_1\rho & \beta_3\beta_2\rho & \beta_3\beta_3 + \sigma^2_{e3} & \beta_3\beta_4 \\ \beta_4\beta_1\rho & \beta_4\beta_2\rho & \beta_4\beta_3 & \beta_4\beta_4 + \sigma^2_{e4} \end{bmatrix} \end{array}$$

Example 2



$$\Sigma_{\text{model}} = \begin{array}{c} \\ I_1 \\ I_2 \\ I_3 \\ I_4 \end{array} \begin{array}{c} \begin{array}{cccc} I_1 & I_2 & I_3 & I_4 \end{array} \\ \begin{bmatrix} \beta_1\beta_1 + \sigma^2{}_{e1} & & & \\ \beta_2\beta_1 & \beta_2\beta_2 + \sigma^2{}_{e2} & & \\ \beta_3\beta_1\rho & \beta_3\beta_2\rho & \beta_3\beta_3 + \sigma^2{}_{e3} & \\ \beta_4\beta_1\rho & \beta_4\beta_2\rho & \beta_4\beta_3 & \beta_4\beta_4 + \sigma^2{}_{e4} \end{bmatrix} \end{array}$$

Example 2



$$\Sigma_{model} = \begin{array}{c} \\ I_1 \\ I_2 \\ I_3 \\ I_4 \end{array} \begin{array}{c} \begin{array}{cccc} I_1 & I_2 & I_3 & I_4 \end{array} \\ \begin{bmatrix} \beta_1\beta_1 + \sigma^2_{e1} & & & \\ \boldsymbol{\beta_2\beta_1} & \beta_2\beta_2 + \sigma^2_{e2} & & \\ \beta_3\beta_1\boldsymbol{\rho} & \beta_3\beta_2\boldsymbol{\rho} & \beta_3\beta_3 + \sigma^2_{e3} & \\ \beta_4\beta_1\boldsymbol{\rho} & \beta_4\beta_2\boldsymbol{\rho} & \boldsymbol{\beta_4\beta_3} & \beta_4\beta_4 + \sigma^2_{e4} \end{bmatrix} \end{array}$$

Example

Sample and model implied covariance matrices:

$$\mathbf{S}_N = \begin{bmatrix} 3.474 & & & \\ 2.826 & 3.745 & & \\ 0.984 & 0.969 & 3.130 & \\ 0.740 & 0.815 & 2.195 & 2.001 \end{bmatrix}$$

$N = 543$

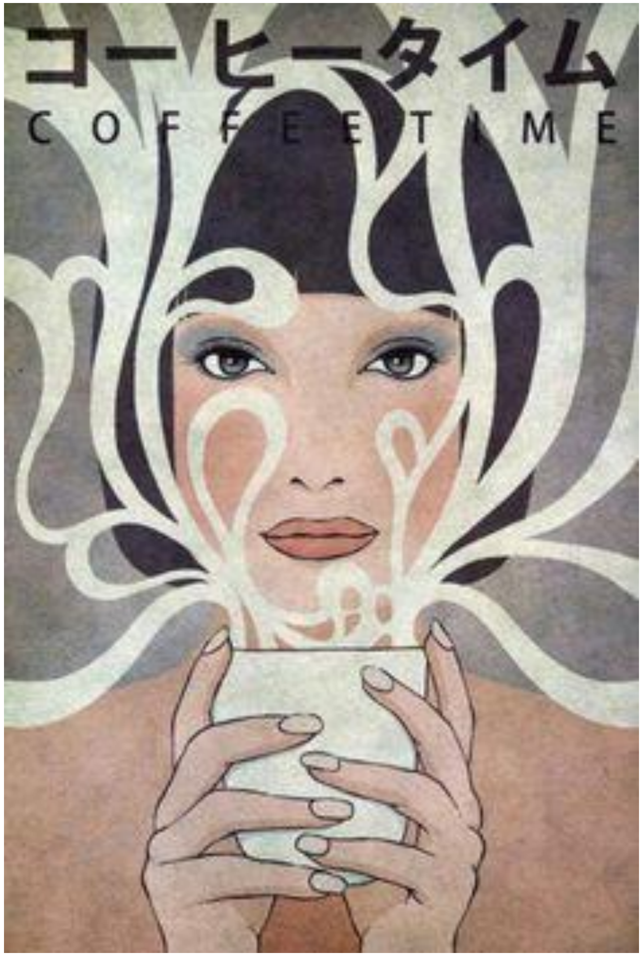
$$\hat{\Sigma}_{model} = \begin{bmatrix} 3.474 & & & \\ 2.826 & 3.745 & & \\ 0.957 & 0.995 & 3.130 & \\ 0.762 & 0.792 & 2.195 & 2.001 \end{bmatrix}$$

$X^2 = 5.281$

$df = 1$

$p = 0.022$

Break



Today

- Intro to Confirmatory Factor Analysis
  - EFA vs CFA
  - Giving Latent Variables a scale
- Model Fit 1
  - Complexity and Degrees of Freedom
  - The Chi-Square Test $\chi 2$
- Model Fit 2
  - Alternative Fit Measures
- Extensions
  - Second order factors
  - Means and intercepts

Problem with chi-square

- Large N $\rightarrow$ high power to detect small discrepancies → "always" significant
- Small N $\rightarrow$ low power to detect large discrepancies → "usually" not significant

Always report the chi-square, df and p, but consider other fit indices as well

Approximate fit

- Root mean squared error of approximation
  - RMSEA = $\sqrt{\frac{\chi 2 - df}{df\,(N-1)}}$ Steiger & Lindt (1980)
    - RMSEA < .05 close fit
    - RMSEA < .08 satisfactory fit
    - RMSEA > .10 bad fit

Unreliable with small N and small df

Incremental fit

Comparative fit index (CFI)

- Chi-square comparison to baseline model
- $0 \leq \text{CFI} \leq 1$
- Rules of thumb: <.90 bad fit, >.95 good fit
- Too low when the correlations between observed variables are low

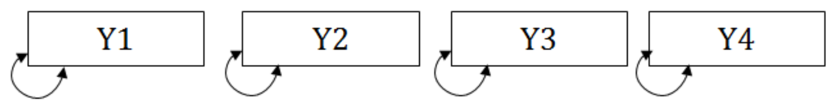
Model fit from lavaan

```
> summary(fit_trust_model_3f, fit.measures = TRUE)
lavaan 0.6-6 ended normally after 45 iterations

  Estimator                                         ML
  Optimization method                           NLMINB
  Number of free parameters                         27

                                                  Used       Total
  Number of observations                         15448       18187

Model Test User Model:

  Test statistic                              9188.922
  Degrees of freedom                                51
  P-value (Chi-square)                           0.000

Model Test Baseline Model:

  Test statistic                             75675.049
  Degrees of freedom                                66
  P-value                                        0.000
```

Chi square of your model

Chi square of a rudimentary default model

What is this baseline model?



Independence model:
Only variances, all covariances fixed @0

Model fit from lavaan

<- This emphasizes that we are looking at relative fit indices, comparing two models

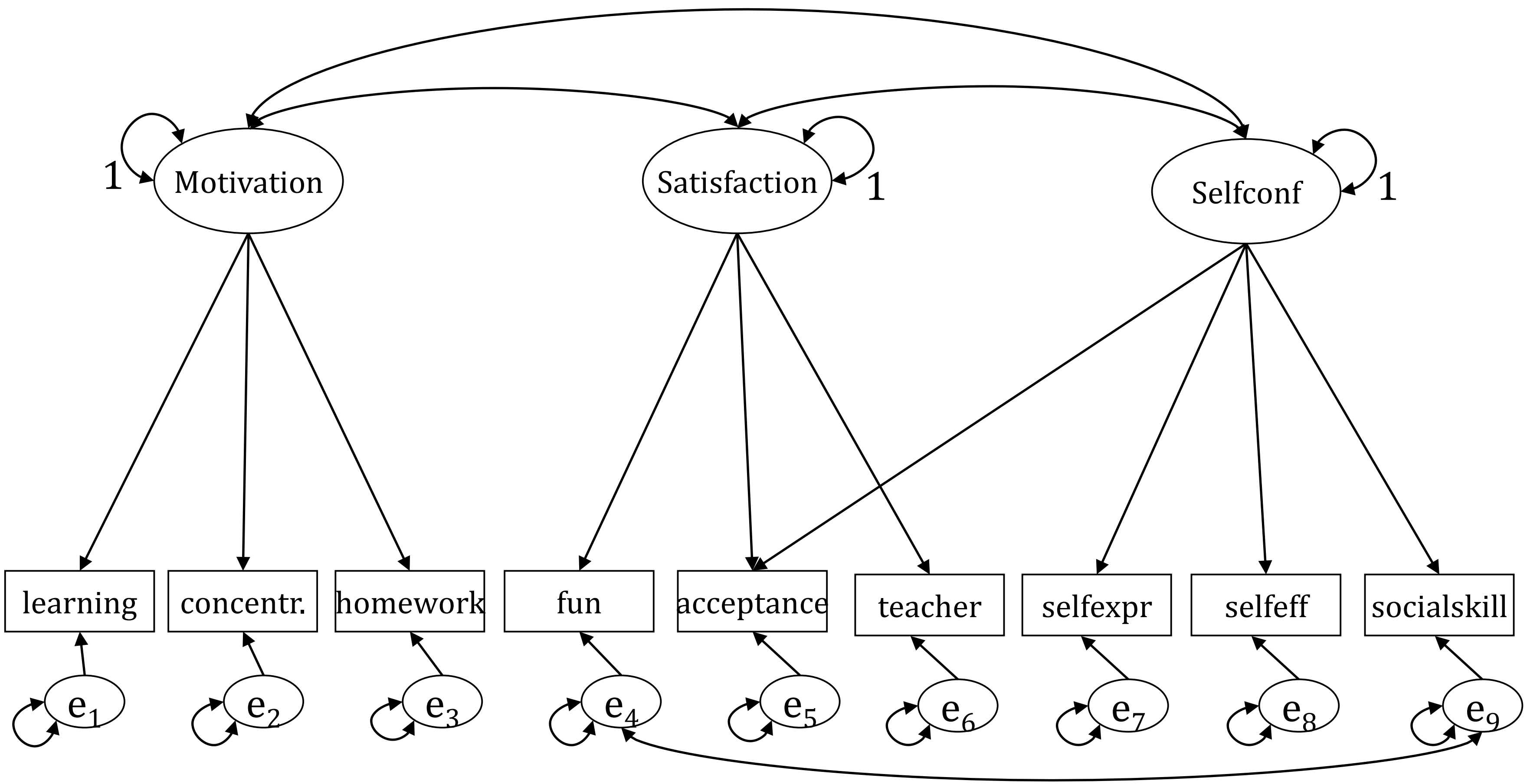
```
User Model versus Baseline Model:

  Comparative Fit Index (CFI)                    0.879
  Tucker-Lewis Index (TLI)                       0.844

Loglikelihood and Information Criteria:

  Loglikelihood user model (H0)            -357923.209
  Loglikelihood unrestricted model (H1)    -353328.748

  Akaike (AIC)                              715900.419
  Bayesian (BIC)                            716106.840
  Sample-size adjusted Bayesian (BIC)       716021.036

Root Mean Square Error of Approximation:

  RMSEA                                          0.108
  90 Percent confidence interval - lower         0.106
  90 Percent confidence interval - upper         0.110
  P-value RMSEA <= 0.05                          0.000

Standardized Root Mean Square Residual:

  SRMR                                           0.058
```

What if the model doesn't fit?

- Do not interpret the parameter estimates

- Revisit theory
- Or modify model

Factor model with cross loading and residual correlation

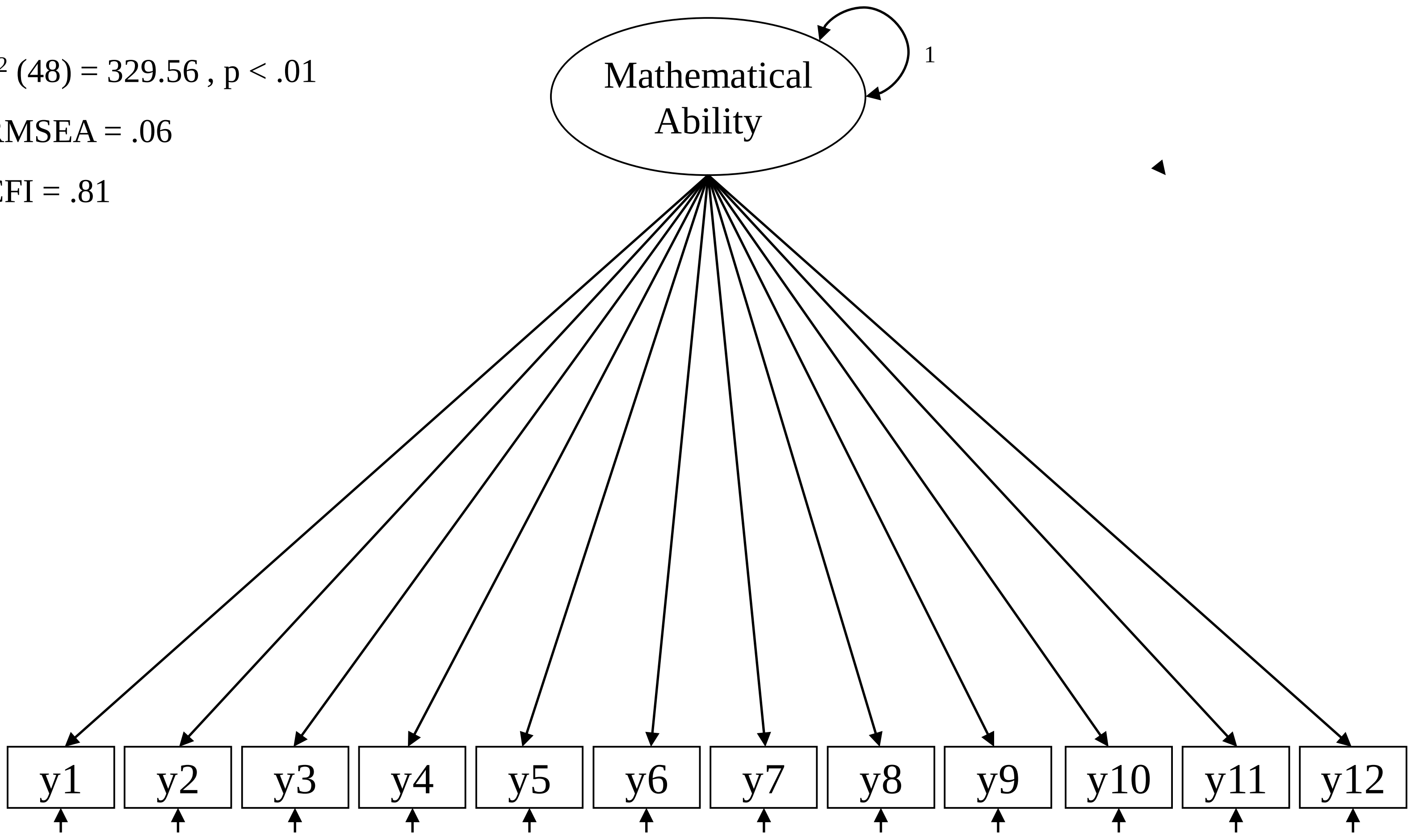


Example

χ^2 (48) = 329.56 , $p < .01$

RMSEA = .06

CFI = .81

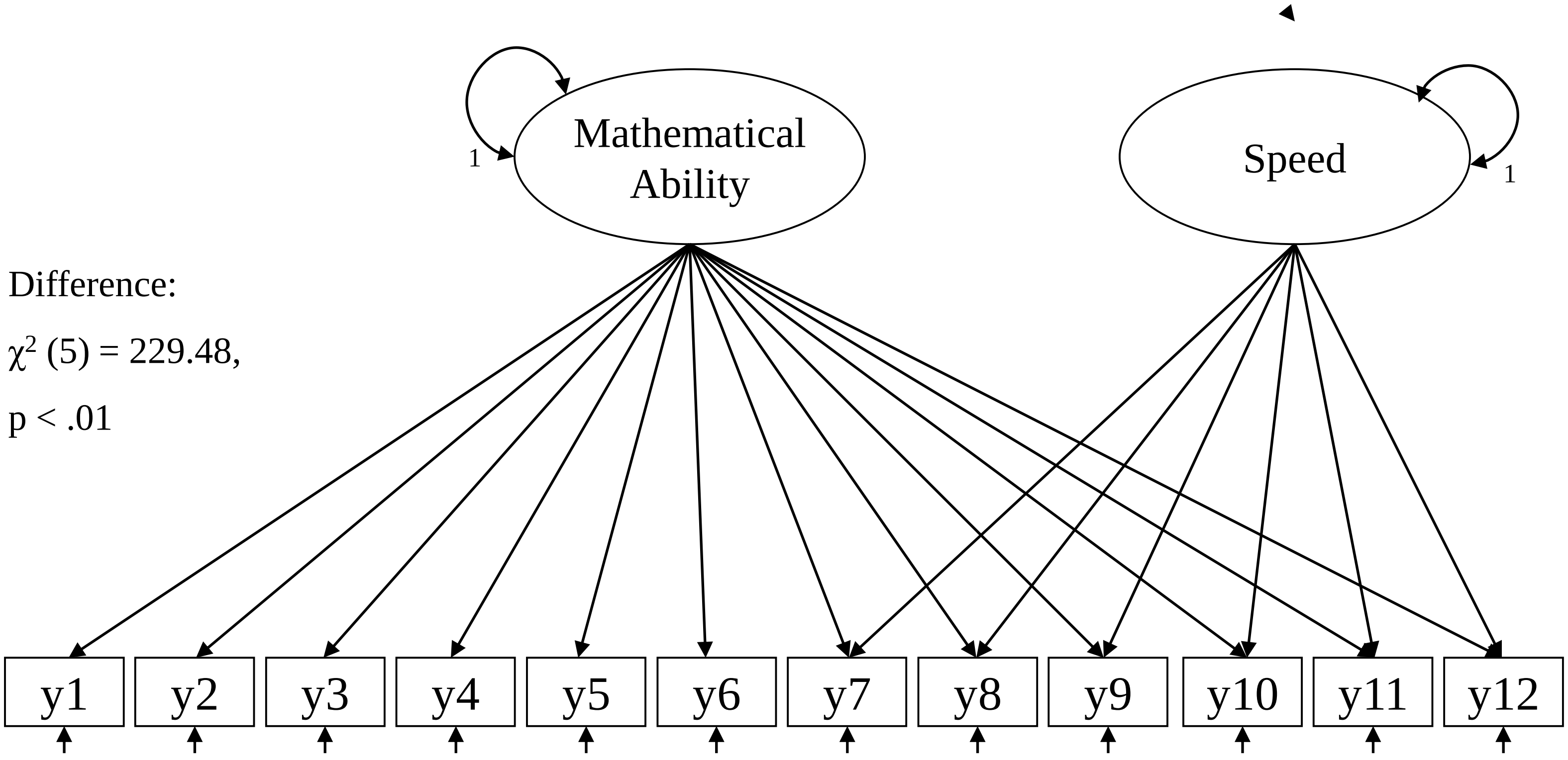


Example

$\chi^2(43) = 63.50$, $p = .023$

RMSEA = .017

CFI = .98



Difference:

χ^2 (5) = 229.48,

$p < .01$

Example

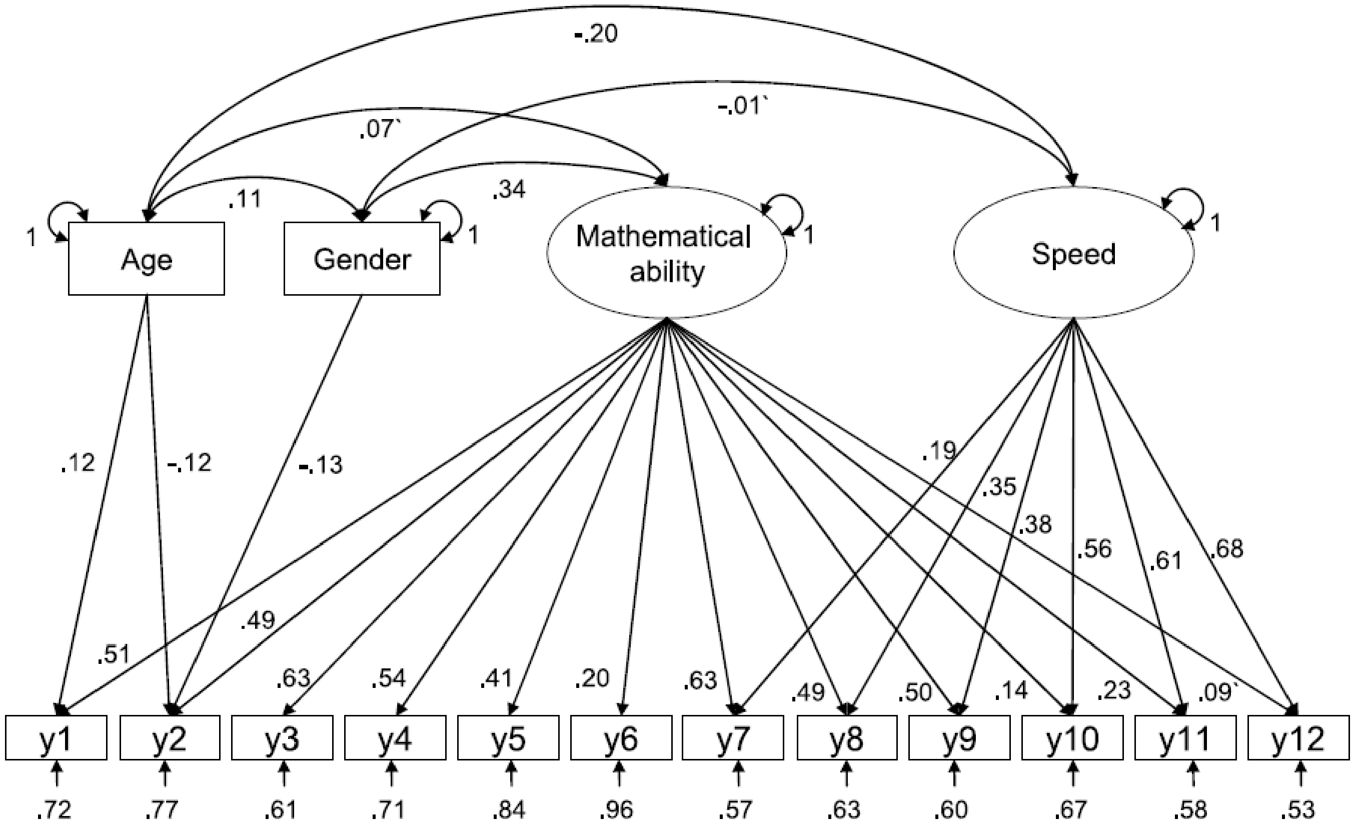
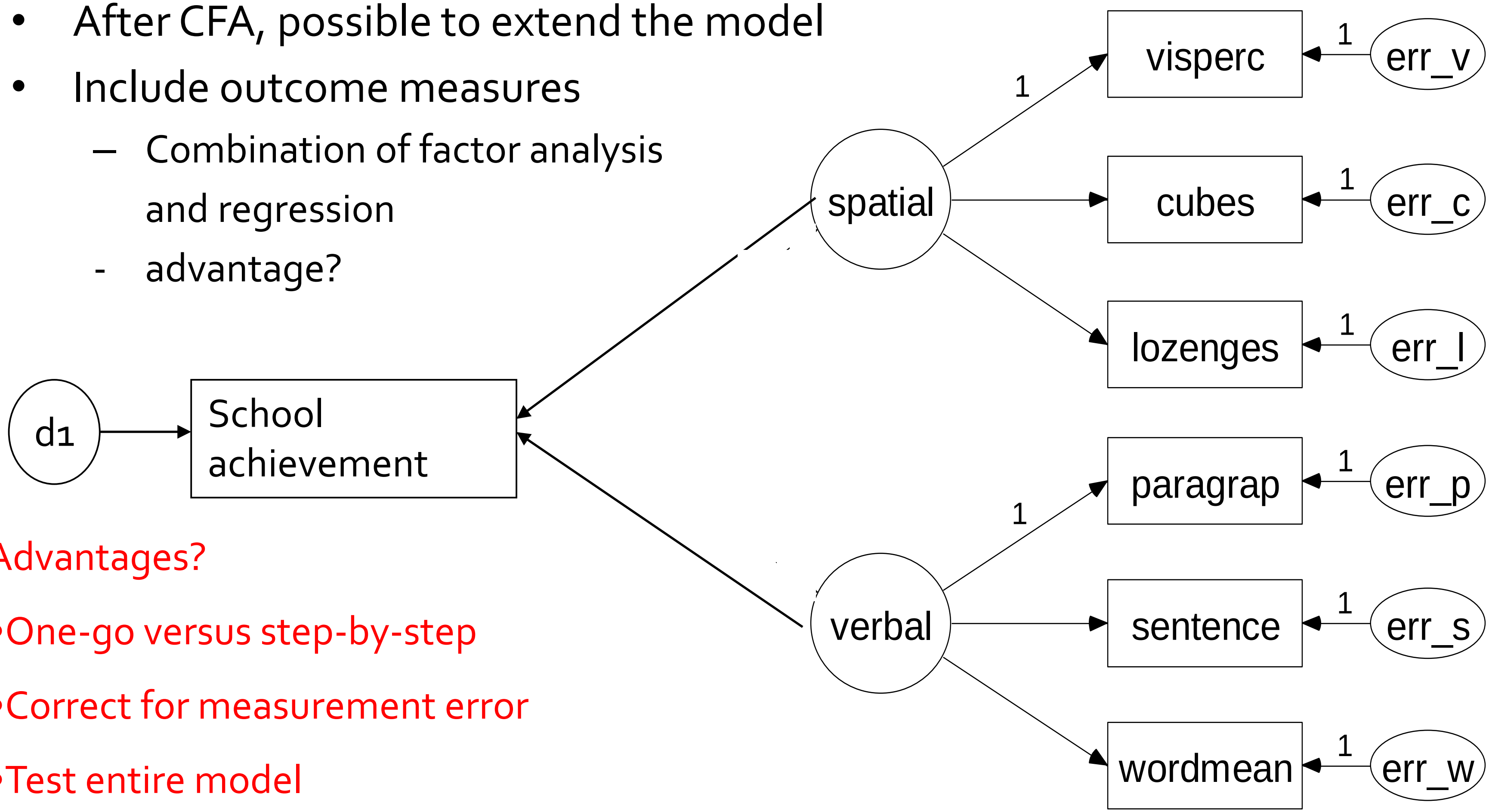


Fig. 1 Mathematical ability measured by worded problems. Notes: All figures denote standardized parameter estimates; apostrophes indicate non-significance; $N = 1617$; model fit: $\chi^2 = 103.79$, $df = 58$, $p < 0.01$, RMSEA = 0.022 [90% CI: 0.015, 0.029], ECVI = 0.122 [90% CI: 0.108, 0.143]

Looking ahead: hybrid models

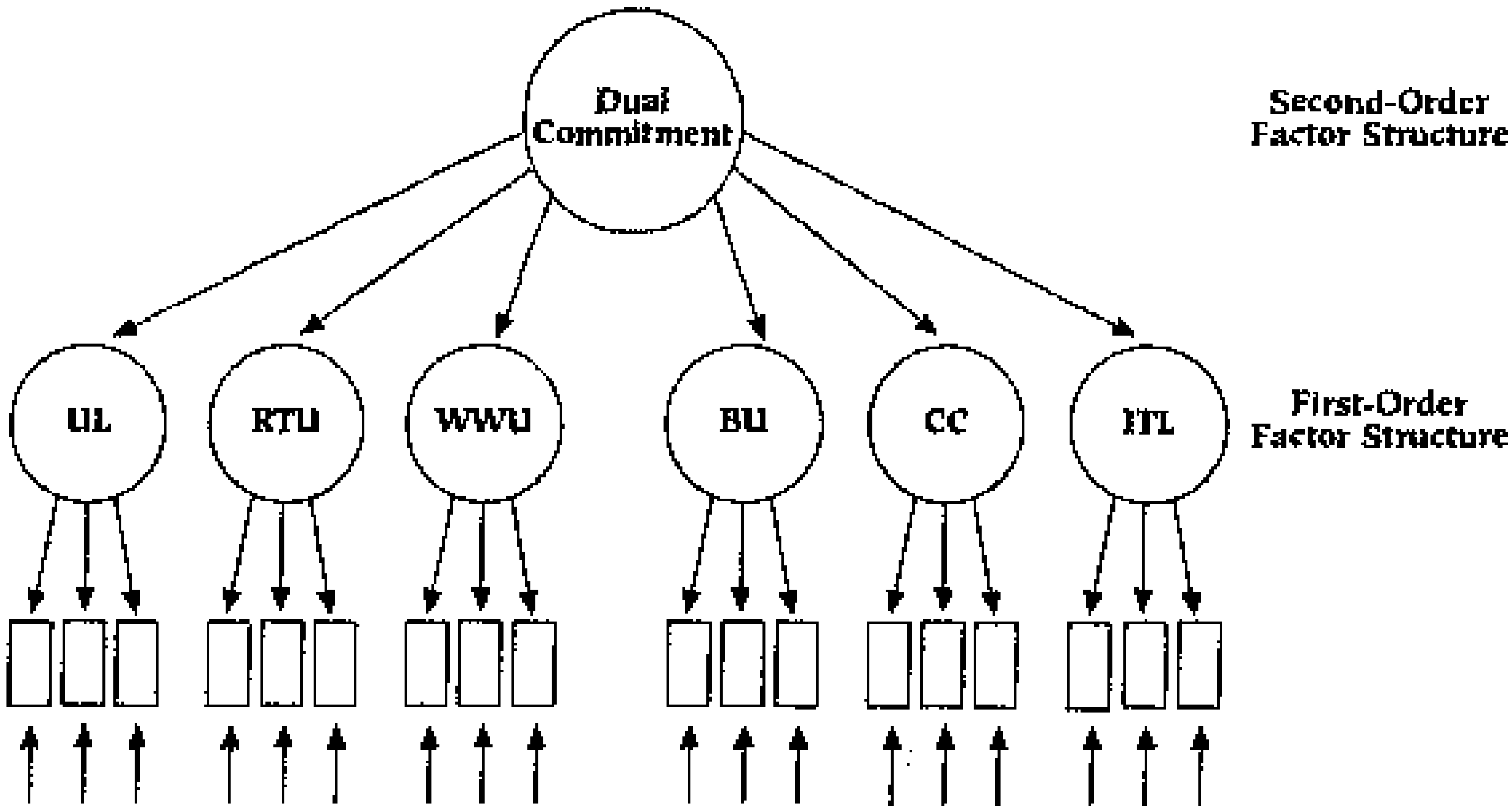
- After CFA, possible to extend the model
- Include outcome measures
  - Combination of factor analysis and regression
  - advantage?



Advantages?

- One-go versus step-by-step
- Correct for measurement error
- Test entire model

Second order CFA



[a]UL = union loyalty, RTU = responsibility to the union, WWU = willingness to work for the union, BU = belief in unionism, CC = company commitment, and ITL = intent to leave.

Second order CFA

In what circumstances can a second order CFA be useful?

When there are multiple factors which can be explained by some common theoretical latent construct (e.g. IQ tests)

- Ideally more than 2 correlated factors, for model identification

Critical thoughts

Theory should come first

Second order CFA is more complex model, so fit will be better

Bad fit means that your model does not describe reality well

Good fit does **not** mean that a second order factor exists "in reality"

Instead of a second order CFA, you could just allow the factors to correlate

A researcher has developed a new questionnaire that should measure someone depression and wants to know how many factors there are. Which technique would you use?

A. PCA
B. EFA
C. CFA

CFA vs EFA vs PCA

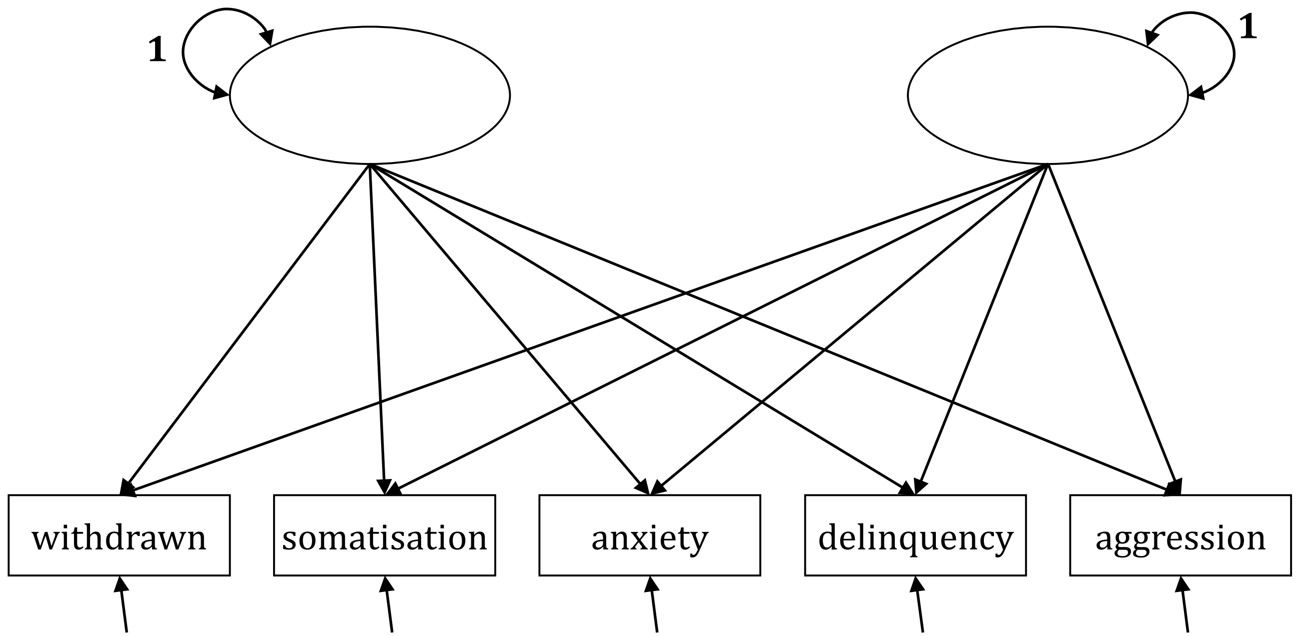
- PCA – summary of variance of items
- EFA – given the data, how many factors are there?
- CFA – is my theoretical model supported by my data?

Learning check

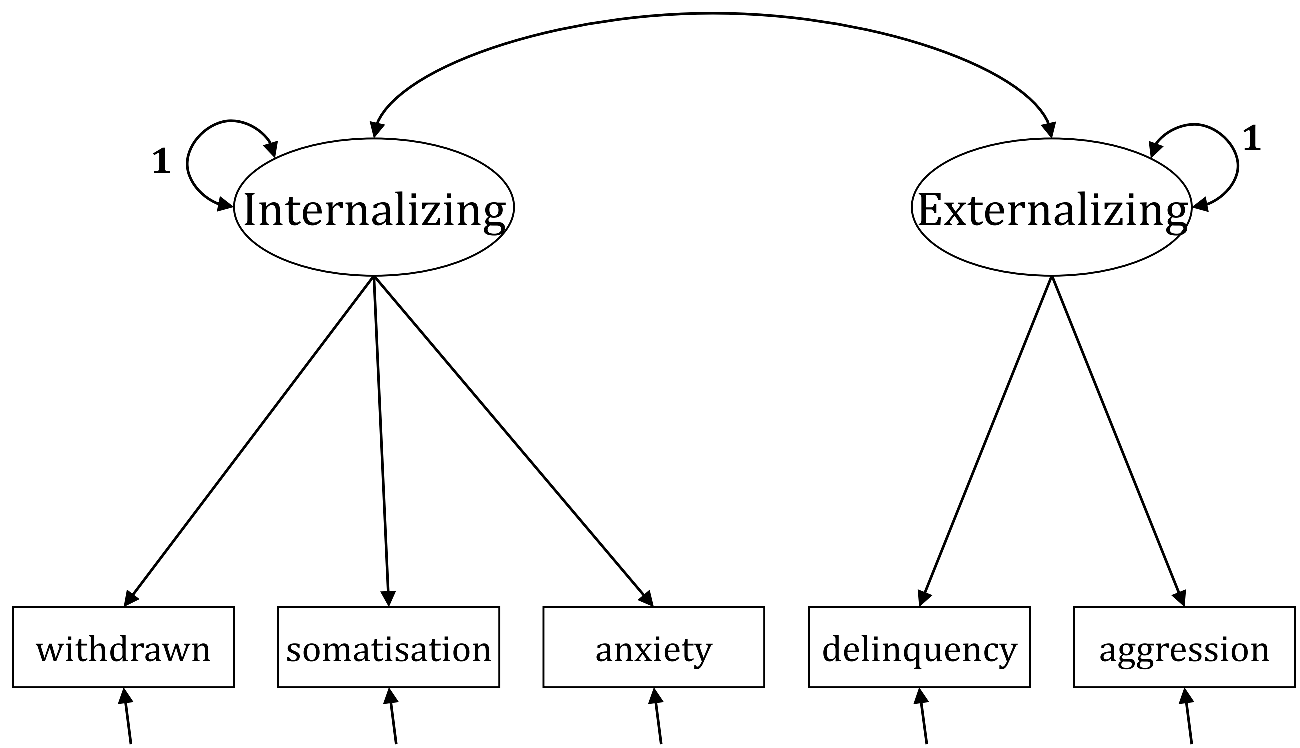
TRUE or FALSE?

Generally, more factor loadings are estimated in a EFA model than in a CFA model

Exploratory factor model



Confirmatory factor model

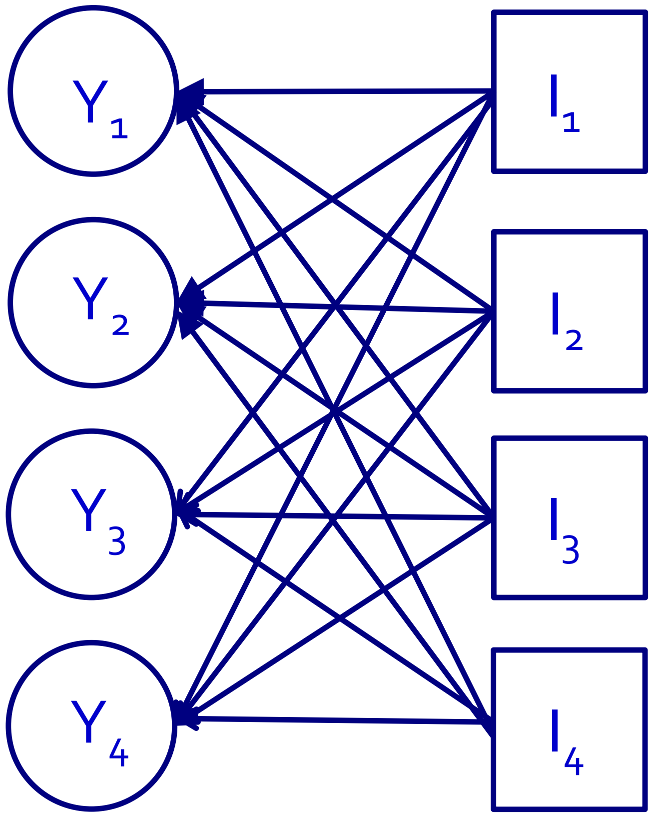


Learning check

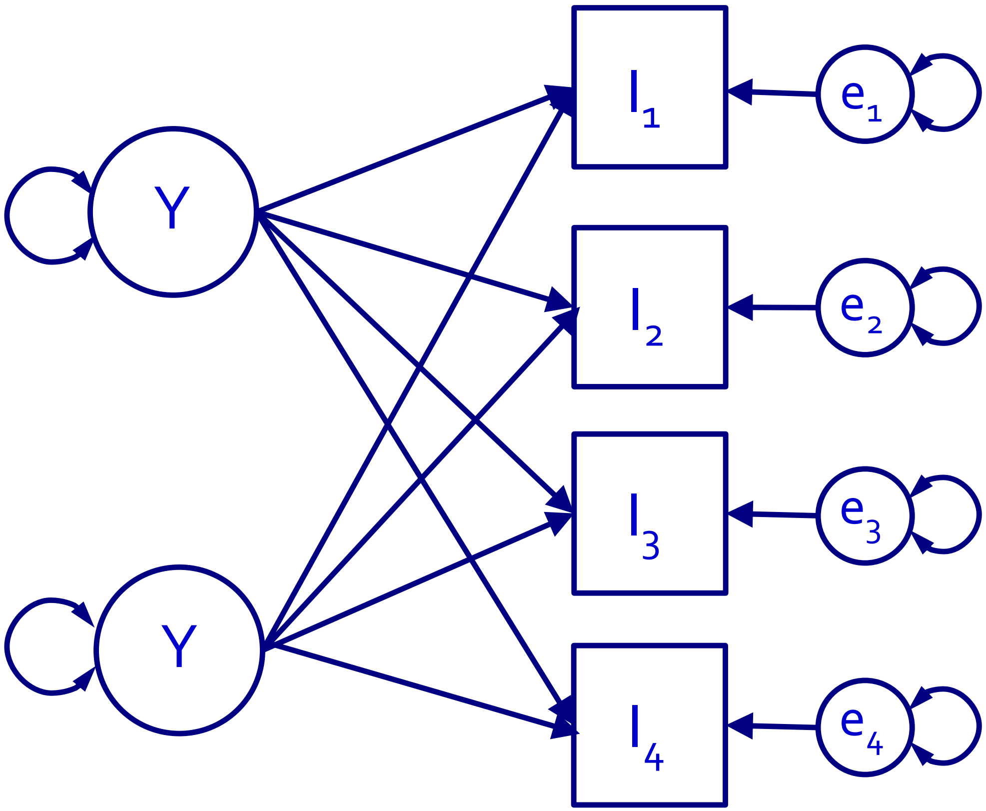
TRUE or FALSE:

PCA and EFA both assume that indicator variables do not have measurement error

PCA



EFA

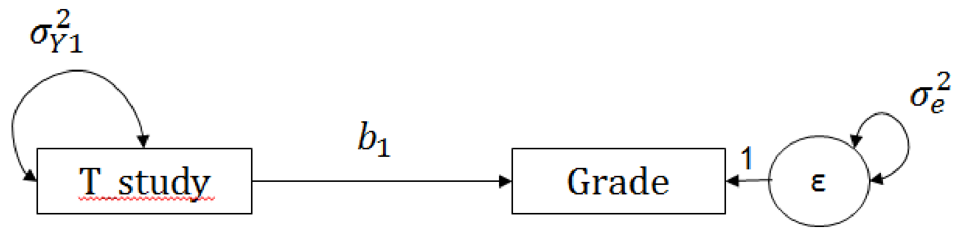


Means and Intercepts

- We have modeled only variances and covariances

We have ignored:

1. Means
2. Intercepts



$$T_{study_i} = b_1\ Grade_i + e_i$$

- Every observed variable has a mean
- We can estimate intercepts and latent means
- This will be covered in more detail in the coming weeks – GLM, multi-group models

Means and Intercepts

- In SEM, you can choose to estimate means and intercepts or not
- If you have missing data, you have to estimate means and intercepts
- Doing this will result in a different number of estimated parameters
- But it will not change the **degrees of freedom**
  - We add z observed means, and estimate z means or intercepts