# Phenotypic factor analysis

Conor Dolan, Mike Neale, & Michel Nivard

Factor analysis Part I:

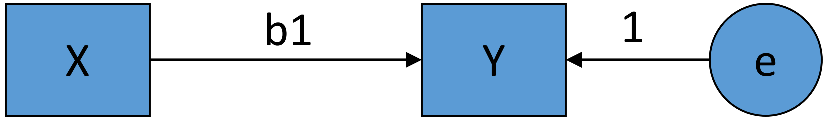
The linear factor model

as a statistical (regression) model - formal representation
as a causal – psychometric - model (vs data reduction)
- what is a common factor substantively?
- implication in terms of data summary and causal modeling
- why is the phenotypic factor model relevant to genetic modeling?
- what can we learn about the phenotypic common factors from twin data?

If you understand linear regression,
you understand a key ingredient of the linear factor model
(as a statistical model).

If you understand logistic linear regression,
you understand a key ingredient of the ordinal factor model
(as a statistical model).

# Path diagram regression model “regression of Y on X”:

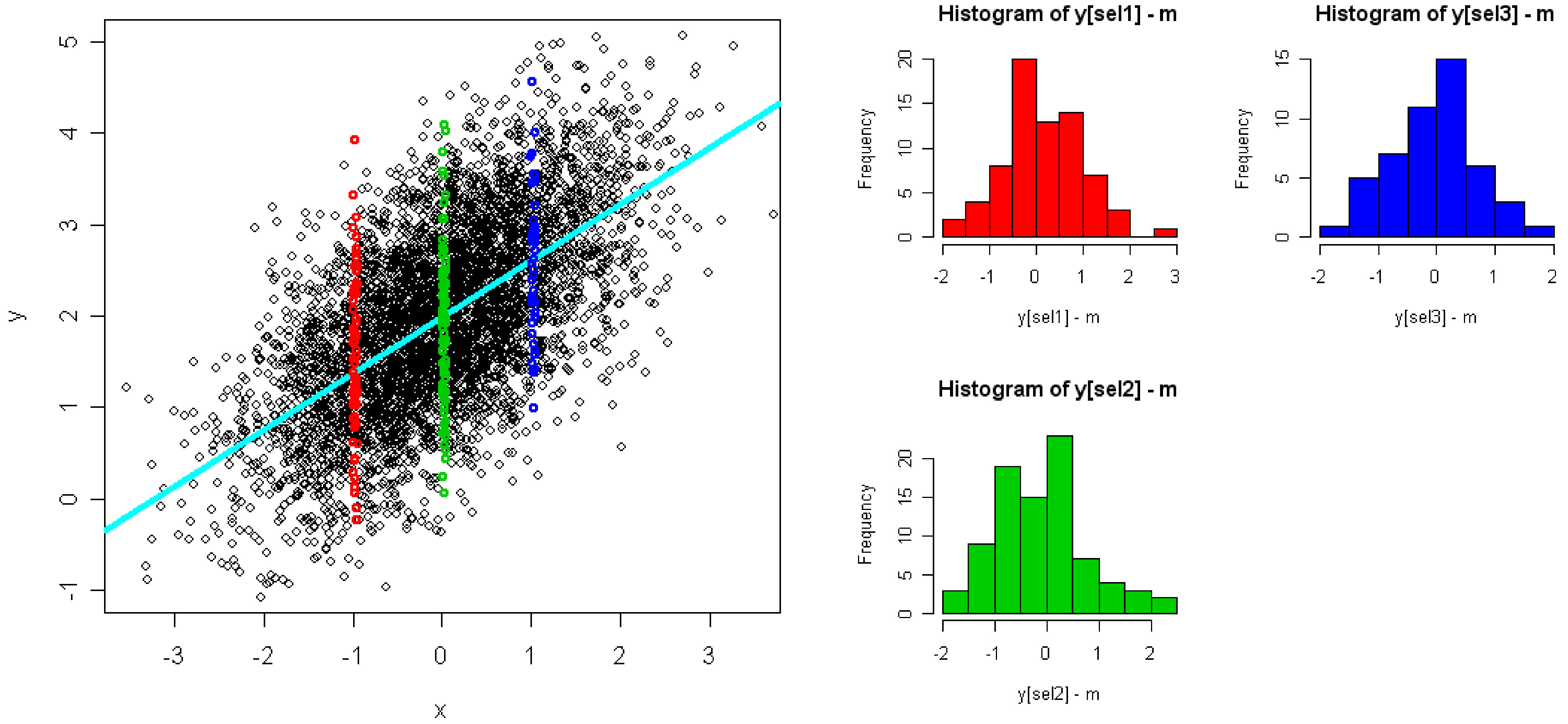


The model for the $Y_i = b0 + b1*X_i + e_i$

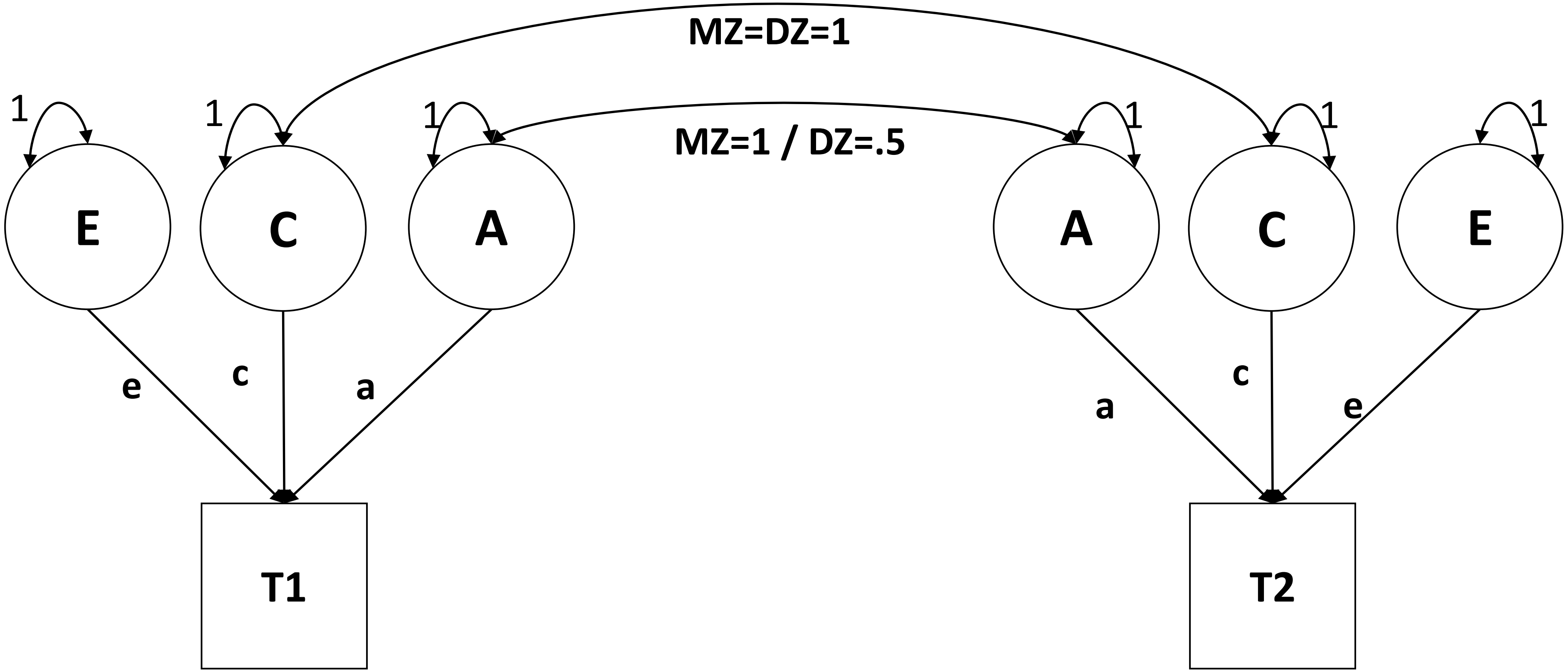
The implied model for the mean:
mean(Y) = b0 + b1*mean(X)
mean(Y) = b0 (if mean(x) = 0)

The implied model for the covariance matrix:

$$\begin{array}{ccc} & x & y \\ x & s^2x & sxy \\ y & sxy & s^2y \end{array} = \begin{array}{cc} x & y \\ s^2x & b1*s^2x \\ b1*s^2x & \boxed{b1^2*s^2x + s^2e} \end{array}$$



Distributional assumption in linear regression concerns the y's given (conditional on) a fixed value of x (x•).
Two important aspects: Linearity, Homoskasticity

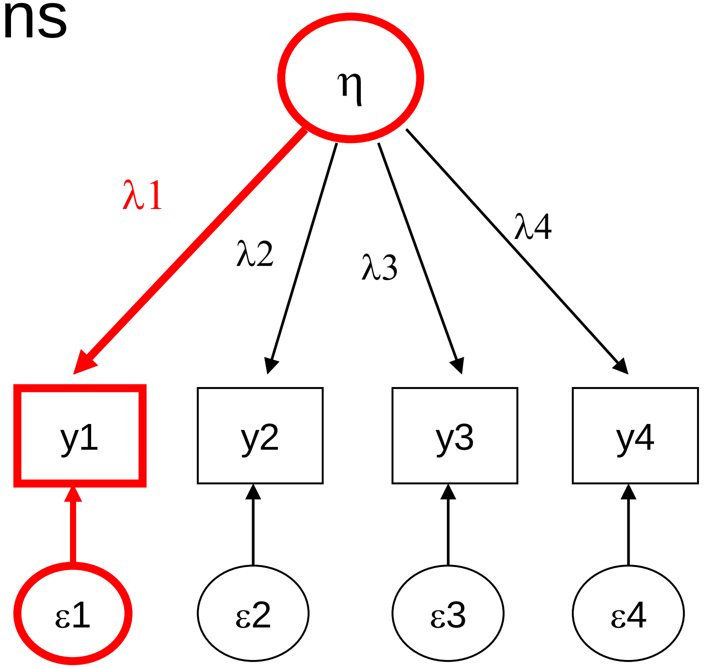


In absence of any interaction, this model is homoskedastic too!

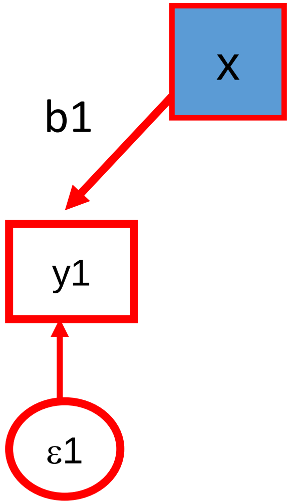
BTW: if you understand heteroskedasticity you understand an important conceptualization of GxE interaction.

# Single common factor model:
# A set of linear regression equations

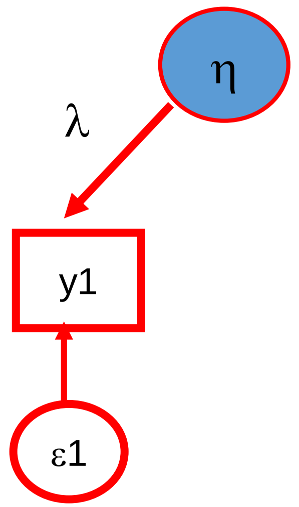
$$\left.\begin{aligned} y_{i1} &= t_1 + \lambda_1 \eta_i + \varepsilon_{i1}, \\ y_{i2} &= t_2 + \lambda_2 \eta_i + \varepsilon_{i2}, \\ y_{i3} &= t_3 + \lambda_3 \eta_i + \varepsilon_{i3}, \\ y_{i4} &= t_4 + \lambda_4 \eta_i + \varepsilon_{i4}. \end{aligned}\right\}$$



$y_{i1} = b_0 + b_1 x_i + e_i$



$$y_{i1} = \tau_1 + \lambda_1 \eta_i + e_i$$



The implied model for the covariance matrix:

| | $\eta$ | $y$ |
|---|---|---|
| $\eta$ | $\sigma^2{}_\eta$ | $\lambda_1\sigma^2{}_\eta$ |
| $y$ | $\lambda_1\sigma^2{}_\eta$ | $\lambda_1{}^2\sigma^2{}_\eta + \sigma^2{}_\varepsilon$ |

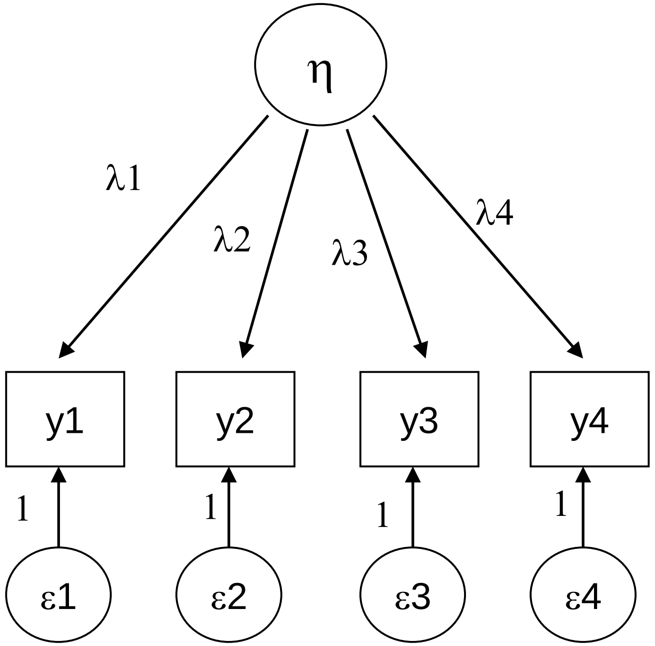
$$\text{Mean}(y_1) = \tau_1 + \lambda_1 \text{Mean}(\eta) = \tau_1$$

$$R^2 = (\lambda_1{}^2 * \sigma^2{}_\eta\ ) \,/\, (\lambda_1{}^2 * \sigma^2{}_\eta + \sigma^2{}_\varepsilon)$$

But what is the point if the common factor (**the independent variable,** $\eta$) is not observed?

# Single common factor model:
# A set of linear regression equations

$$\left.\begin{aligned}
y_{i1} &= t_1 + \lambda_1 \eta_i + \varepsilon_{i1}, \\
y_{i2} &= t_2 + \lambda_2 \eta_i + \varepsilon_{i2}, \\
y_{i3} &= t_3 + \lambda_3 \eta_i + \varepsilon_{i3}, \\
y_{i4} &= t_4 + \lambda_4 \eta_i + \varepsilon_{i4}.
\end{aligned}\right\}$$



Implies a covariance matrix:

$$\Sigma = \begin{bmatrix}
\lambda_1^2 \sigma_\eta^2 + \sigma_{\varepsilon_1}^2 & & & \\
\lambda_1 \sigma_\eta^2 \, \lambda_2 & \lambda_2^2 \sigma_\eta^2 + \sigma_{\varepsilon 2}^2 & & \\
\lambda_1 \sigma_\eta^2 \lambda_3 & \lambda_2 \sigma_\eta^2 \lambda_3 & \lambda_3^2 \sigma_\eta^2 + \sigma_{\varepsilon 3}^2 & \\
\lambda_1 \sigma_\eta^2 \lambda_4 & \lambda_2 \sigma_\eta^2 \lambda_4 & \lambda_3 \sigma_\eta^2 \, \lambda_4 & \lambda_4^2 \sigma_\eta^2 + \sigma_{\varepsilon 4}^2
\end{bmatrix}$$

# A set of linear regression coefficients expressed as a single matrix equation: using matrix algebra

$$\left.\begin{array}{l} y_{i1} - t_1 = \lambda_1 \eta_i + \varepsilon_{i1}, \\ y_{i2} - t_2 = \lambda_2 \eta_i + \varepsilon_{i2}, \\ y_{i3} - t_3 = \lambda_3 \eta_i + \varepsilon_{i3}, \\ y_{i4} - t_4 = \lambda_4 \eta_i + \varepsilon_{i4}. \end{array}\right\} \quad \mathbf{y_i} - \mathbf{t} = \boldsymbol{\Lambda}\boldsymbol{\eta}_\mathbf{i} + \boldsymbol{\varepsilon}_\mathbf{i}$$

Matrix algebra is

1) Notationally Efficient

2) Basis of Multivariate Statistics (useful to know!)

$$\left.\begin{aligned} y_{i1}-t_1 &= \lambda_1\eta_i+\varepsilon_{i1}, \\ y_{i2}-t_2 &= \lambda_2\eta_i+\varepsilon_{i2}, \\ y_{i3}-t_3 &= \lambda_3\eta_i+\varepsilon_{i3}, \\ y_{i4}-t_4 &= \lambda_4\eta_i+\varepsilon_{i4}. \end{aligned}\right\}$$

ny number of variables
ne number of common factors

$$\mathbf{y}_i - \mathbf{t} = \boldsymbol{\Lambda}\boldsymbol{\eta}_i + \boldsymbol{\varepsilon}_i$$

ny x 1 ny x ne ne x 1 ny x 1

$$\mathbf{t}_i = \begin{pmatrix} t_1 \\ t_2 \\ t_3 \\ t_4 \end{pmatrix} \quad \mathbf{y}_i = \begin{pmatrix} y_{i1} \\ y_{i2} \\ y_{i3} \\ y_{i4} \end{pmatrix} \quad \boldsymbol{\varepsilon}_i = \begin{pmatrix} \varepsilon_{i1} \\ \varepsilon_{i2} \\ \varepsilon_{i3} \\ \varepsilon_{i4} \end{pmatrix} \quad \boldsymbol{\Lambda} = \begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \\ \lambda_4 \end{pmatrix} \quad \boldsymbol{\eta}_i = \begin{pmatrix} \eta_i \end{pmatrix}$$

1 x 1

ny x 1

$$\left.\begin{aligned} y_{i1} &= \lambda_1 \eta_i + \varepsilon_{i1}, \\ y_{i2} &= \lambda_2 \eta_i + \varepsilon_{i2}, \\ y_{i3} &= \lambda_3 \eta_i + \varepsilon_{i3}, \\ y_{i4} &= \lambda_4 \eta_i + \varepsilon_{i4}. \end{aligned}\right\}$$

ny number of variables
ne number of common factors

$$\mathbf{y}_i = \boldsymbol{\Lambda} \boldsymbol{\eta}_i + \boldsymbol{\varepsilon}_i$$

**Centered t = 0!**

ny x 1 ny x ne ne x 1 ny x 1

$\sigma^2_{y1} = E[y_1 y_1] = E[(\lambda_1 \eta_i + \varepsilon_i)(\lambda_1 \eta_i + \varepsilon_i)] =$

$E[(\lambda_1 \eta_i \lambda_1 \eta_i + \lambda_1 \eta_i \varepsilon_i + \varepsilon_i \lambda_1 \eta_i + \varepsilon_i \varepsilon_i)] =$

$E[\lambda_1 \eta_i \lambda_1 \eta_i] + \cancel{E[\lambda_1 \eta_i \varepsilon_i] + E[\varepsilon_i \lambda_1 \eta_i]} + E[\varepsilon_i \varepsilon_i] =$

$\lambda_1 \lambda_1 E[\eta_i \eta_i] + \cancel{\lambda_1 E[\eta_i \varepsilon_i] + \lambda_1 E[\varepsilon_i \eta_i]} + E[\varepsilon_i \varepsilon_i] =$

$\lambda_1 \lambda_1 E[\eta_i \eta_i] + E[\varepsilon_i \varepsilon_i] = \lambda_1^2 \sigma^2_\eta + \sigma^2_\varepsilon$

$$\left.\begin{aligned} y_{i1} &= \lambda_1 \eta_i + \varepsilon_{i1}, \\ y_{i2} &= \lambda_2 \eta_i + \varepsilon_{i2}, \\ y_{i3} &= \lambda_3 \eta_i + \varepsilon_{i3}, \\ y_{i4} &= \lambda_4 \eta_i + \varepsilon_{i4}. \end{aligned}\right\}$$

ny number of variables
ne number of common factors

$$\mathbf{y}_i = \Lambda \eta_i + \varepsilon_I$$

**Centered t = 0!**

ny x 1   ny x ne   ne x 1   ny x 1

$\Sigma_y = E[y{*}y^t] = E[(\Lambda\eta_i + \varepsilon_i)(\Lambda\eta_i + \varepsilon_i)^t] =$ (1)

$E[(\Lambda\eta_i + \varepsilon_i)(\eta_i^t\Lambda^t + \varepsilon_i^t)] =$ (2)

$E[\Lambda\eta_i\,\eta_i^t\Lambda^t + \Lambda\eta_i\varepsilon_i^t + \varepsilon_i\eta_i^t\Lambda^t + \varepsilon_i\varepsilon_i^t] =$ (3)

$E[\Lambda\eta_i\,\eta_i^t\Lambda^t] + E[\Lambda\eta_i\varepsilon_i^t] + E[\varepsilon_i\eta_i^t\Lambda^t] + E[\varepsilon_i\varepsilon_i^t] =$ (4)

$\Lambda E[\eta_i\,\eta_i^t]\Lambda^t + \Lambda E[\eta_i\varepsilon_i^t] + E[\varepsilon_i\eta_i^t]\Lambda^t + E[\varepsilon_i\varepsilon_i^t] =$ (5)

$\Sigma_y = \Lambda E[\eta_i\,\eta_i^t]\Lambda^t + E[\varepsilon_i\varepsilon_i^t] = \Lambda\Psi\Lambda^t + \Theta$ (6)

$$\left.\begin{aligned} y_{i1} &= \lambda_1 \eta_i + \varepsilon_{i1}, \\ y_{i2} &= \lambda_2 \eta_i + \varepsilon_{i2}, \\ y_{i3} &= \lambda_3 \eta_i + \varepsilon_{i3}, \\ y_{i4} &= \lambda_4 \eta_i + \varepsilon_{i4}. \end{aligned}\right\}$$

ny number of variables
ne number of common factors

$$\underset{\text{ny x 1}}{\mathbf{y}_i} = \underset{\text{ny x ne}}{\boldsymbol{\Lambda}}\underset{\text{ne x 1}}{\boldsymbol{\eta}_i} + \underset{\text{ny x 1}}{\boldsymbol{\varepsilon}_i}$$

$\boldsymbol{\Sigma}_{\mathbf{y}} = \mathrm{E}[\mathbf{y}\mathbf{y}^t] = \mathrm{E}[(\boldsymbol{\Lambda}\boldsymbol{\eta}_i + \boldsymbol{\varepsilon}_i)(\boldsymbol{\Lambda}\boldsymbol{\eta}_i + \boldsymbol{\varepsilon}_i)^t] = \boldsymbol{\Lambda\Psi\Lambda}^t + \boldsymbol{\Theta}$

$\mathrm{E}[\boldsymbol{\eta}_i\, \boldsymbol{\eta}_i^t] = \boldsymbol{\Psi}$ and $\mathrm{E}[\boldsymbol{\varepsilon}_i\, \boldsymbol{\varepsilon}_i^t] = \boldsymbol{\Theta}$

You can represent this model in OpenMx using matrices

**So what? What is the use?**

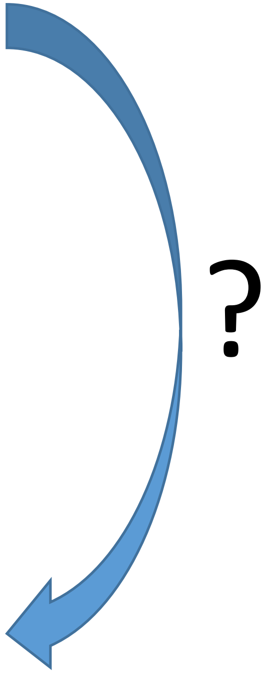
# Depression items

- I feel lonely
- I feel confused or in a fog
- I cry a lot
- I worry about my future.
- I am afraid I might think or do something bad
- I feel that I have to be perfect
- I feel that no one loves me
- I feel worthless or inferior
- I am nervous or tense
- I lack self confidence I am too fearful or anxious
- I feel too guilty
- I am self-conscious or easily embarrassed
- I am unhappy, sad or depressed
- I worry a lot
- I am too concerned about how I look
- I worry about my relations with the opposite sex

Are the data consistent with the presence of a latent construct that underlies the observed variables (items) and that accounts for the inter-correlations between variables?

The use is to test the hypothesis that the linear relations among my 4 indicators of neuroticism, as expressed in the correlation or covariance matrix, are consistent with the presence of a single common influence, the latent variable “neuroticism”.

|  | n3 | n4 | n5 | n6 |
|---|---|---|---|---|
| n3 | 35.376 | 0.624 | 0.204 | 0.685 |
| n4 | 15.807 | 18.159 | 0.154 | 0.586 |
| n5 | 4.956 | 2.668 | 16.640 | 0.225 |
| n6 | 19.023 | 11.654 | 4.274 | 21.769 |



$$\Sigma = \begin{bmatrix} \lambda_1^2\sigma_\eta^2 + \sigma_{\varepsilon_1}^2 & & & \\ \lambda_1\sigma_\eta^2\,\lambda_2 & \lambda_2^2\sigma_\eta^2 + \sigma_{\varepsilon 2}^2 & & \\ \lambda_1\sigma_\eta^2\lambda_3 & \lambda_2\sigma_\eta^2\lambda_3 & \lambda_3^2\sigma_\eta^2 + \sigma_{\varepsilon 3}^2 & \\ \lambda_1\sigma_\eta^2\lambda_4 & \lambda_2\sigma_\eta^2\lambda_4 & \lambda_3\sigma_\eta^2\,\lambda_4 & \lambda_4^2\sigma_\eta^2 + \sigma_{\varepsilon 4}^2 \end{bmatrix}$$

N=361, Female 1st year psychology students

```
#
datf=read.table('rdataf')
colnames(datf)=c('sex',
 'n1', 'n2', 'n3', 'n4', 'n5', 'n6',
 'e1', 'e2', 'e3', 'e4', 'e5', 'e6',
 'o1', 'o2', 'o3', 'o4', 'o5', 'o6',
 'a1', 'a2', 'a3', 'a4', 'a5', 'a6',
 'c1', 'c2', 'c3', 'c4', 'c5', 'c6')
datf_n4=datf[,4:7]
Sn4=cov(datf_n4)
round(Sn4,3)
fa1=factanal(covmat=Sn4,n.obs=361,factors=1)
#
library(OpenMx)
datf_n4=as.data.frame(datf_n4)
ny=4
selvars=colnames(datf_n4)
#
Rn4=mxMatrix(type='Stand',nrow=ny,ncol=ny,free=TRUE,value=.5,
                          lbound=-.9,ubound=.9,name='corn4')
Sdsn4=mxMatrix(type='Diag',nrow=ny,ncol=ny,free=TRUE,value=5,name='sdsn4')
Meann4=mxMatrix(type='Full',nrow=1,ncol=ny,free=TRUE,value=20,name='men4')
MkS=mxAlgebra(expression=sdsn4%*%corn4%*%sdsn4,name='Ssatn4')
modelp1=mxModel('part1',Rn4, Sdsn4, Meann4,MkS)
#
N4model0=mxModel("N4sat",
   mxData( observed=datf, type="raw"),   # the data
   mxFIMLObjective( covariance="part1.Ssatn4", means="part1.men4",
   dimnames=selvars)  # the fit function
                                    )
Model1 <-  mxModel("model1", modelp1, N4model0,
                 mxAlgebra(N4sat.objective, name="minus2loglikelihood"),
                 mxAlgebraObjective("minus2loglikelihood"))

# fit the model
Model1_o <- mxRun(Model1)
#
Ly=mxMatrix(type='Full',nrow=ny,ncol=1,free=TRUE,value=1,name='Ly')
Te=mxMatrix(type='Diag',nrow=ny,ncol=ny,free=TRUE,value=5,name='Te')
Ps=mxMatrix(type='Symm',nrow=1,ncol=1,free=FALSE,value=1,name='Ps')
Meann4=mxMatrix(type='Full',nrow=1,ncol=ny,free=TRUE,value=20,name='men4')
MkS=mxAlgebra(expression=Ly%*%Ps%*%t(Ly)+Te,name='S1fn4')
#
modelp1=mxModel('part1',Ly, Te, Ps, Meann4, MkS)
#
N4model2=mxModel("N4f1",
   mxData( observed=datf, type="raw"),   # the data
   mxFIMLObjective( covariance="part1.S1fn4", means="part1.men4",
   dimnames=selvars)  # the fit function
#                                    )
Model2 <-  mxModel("model1", modelp1, N4model2,
                 mxAlgebra(N4f1.objective, name="minus2loglikelihood"),
                 mxAlgebraObjective("minus2loglikelihood"))
# fit the model
Model2_o <- mxRun(Model2)
MxCompare(Model1_o,Model2_o)
```

The chi2 goodness of fit test ($\chi2$=1.36, df=2) suggest that the model fits well. The observed covariance structure is consistent with my theory.

Does this prove the presence of the latent variable?
**necessary but not sufficient….**

Why df=2?

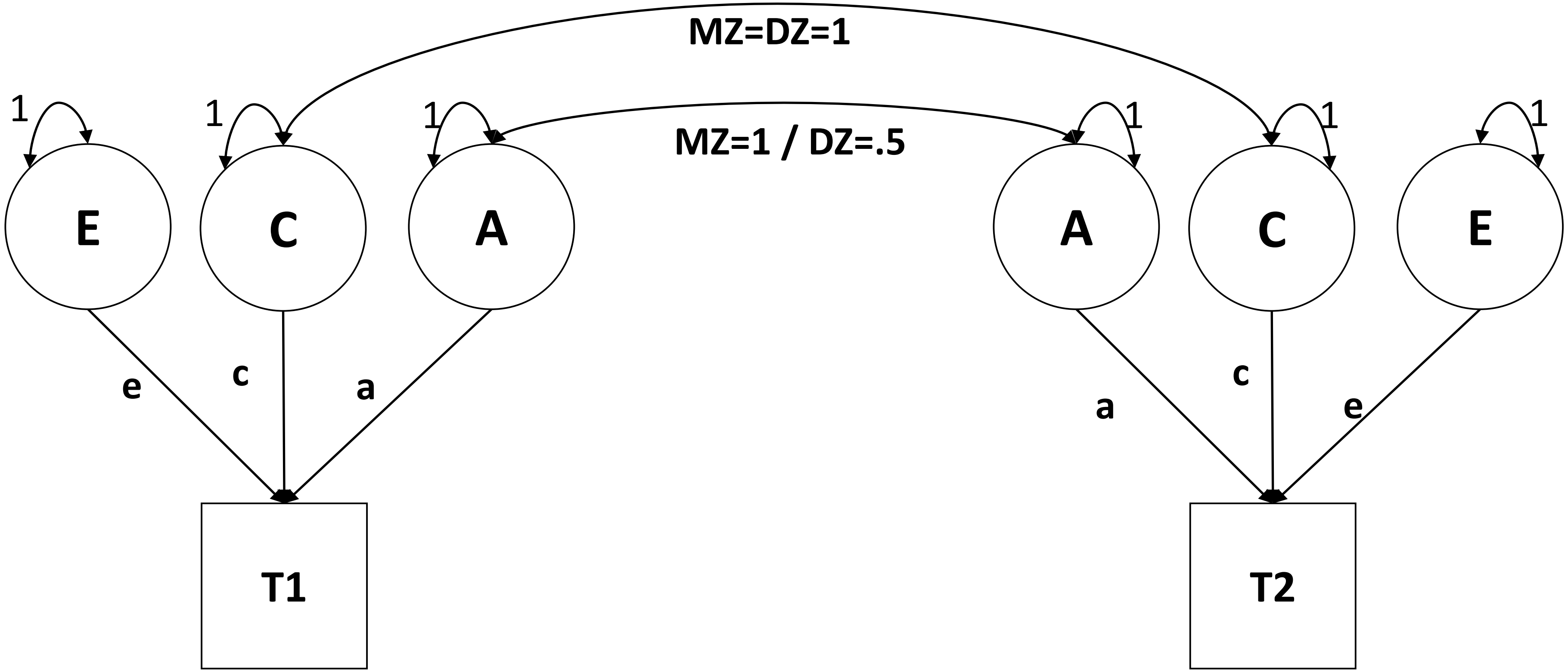
Count the observed statistics (S), and the estimated parameters (P): df = S-P.

# A technical aspect of the common factor model: scaling.

$$\Sigma = \begin{bmatrix} \lambda_1^2\sigma_\eta^2 + \sigma_{\varepsilon_1}^2 & & & \\ \lambda_1\sigma_\eta^2\,\lambda_2 & \lambda_2^2\sigma_\eta^2 + \sigma_{\varepsilon 2}^2 & & \\ \lambda_1\sigma_\eta^2\lambda_3 & \lambda_2\sigma_\eta^2\lambda_3 & \lambda_3^2\sigma_\eta^2 + \sigma_{\varepsilon 3}^2 & \\ \lambda_1\sigma_\eta^2\lambda_4 & \lambda_2\sigma_\eta^2\lambda_4 & \lambda_3\sigma_\eta^2\,\lambda_4 & \lambda_4^2\sigma_\eta^2 + \sigma_{\varepsilon 4}^2 \end{bmatrix}$$

The mean and variance of the common factor? The common factor is latent!
Scale by setting the mean to zero. ($\mu_\eta = 0$)
Scale by fixing variance to “sensible value” ($\sigma_\eta^2 = 1$)
Scale by making it dependent on an indicator by fixing a factor loading to 1 ($\lambda_1$=1)

But we know about scaling, because this model uses the same scaling (var(E)=var(C)=var(A) = 1; mean(A)=mean(C)=mean(E)=0)



# A technical aspect of the common factor model: scaling.

$$\Sigma = \begin{bmatrix} \lambda_1^2 + \sigma_{\varepsilon_1}^2 & & & \\ \lambda_1\lambda_2 & \lambda_2^2 + \sigma_{\varepsilon 2}^2 & & \\ \lambda_1\lambda_3 & \lambda_2\lambda_3 & \lambda_3^2 + \sigma_{\varepsilon 3}^2 & \\ \lambda_1\lambda_4 & \lambda_2\lambda_4 & \lambda_3\lambda_4 & \lambda_4^2 + \sigma_{\varepsilon 4}^2 \end{bmatrix}$$

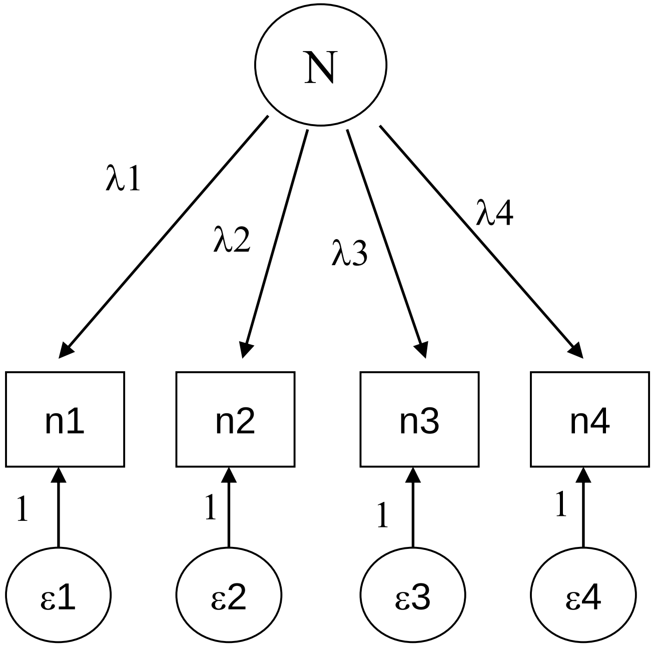
Scale the common factor by fixing to “sensible value” ($\sigma_\eta^2 = 1$)

# A technical aspect of the common factor model: scaling.

$$\Sigma = \begin{bmatrix} \sigma_\eta^2 + \sigma_{\varepsilon_1}^2 & & & \\ \sigma_\eta^2 \lambda_2 & \lambda_2^2 \sigma_\eta^2 + \sigma_{\varepsilon 2}^2 & & \\ \sigma_\eta^2 \lambda_3 & \lambda_2 \sigma_\eta^2 \lambda_3 & \lambda_3^2 \sigma_\eta^2 + \sigma_{\varepsilon 3}^2 & \\ \sigma_\eta^2 \lambda_4 & \lambda_2 \sigma_\eta^2 \lambda_4 & \lambda_3 \sigma_\eta^2 \lambda_4 & \lambda_4^2 \sigma_\eta^2 + \sigma_{\varepsilon 4}^2 \end{bmatrix}$$

Or making it dependent on an indicator by fixing a factor loading to 1 ($\lambda_1$=1)

# A substantive aspect of the common factor model: interpretation (you bring to the model!)

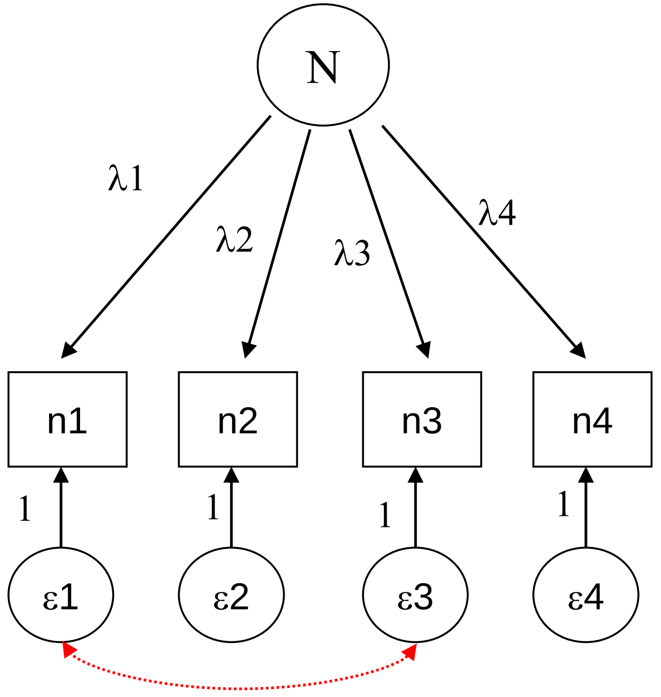


Reflective indicators:
They reflect the causal action of the latent variable N

Strong realistic view of the latent variable N:

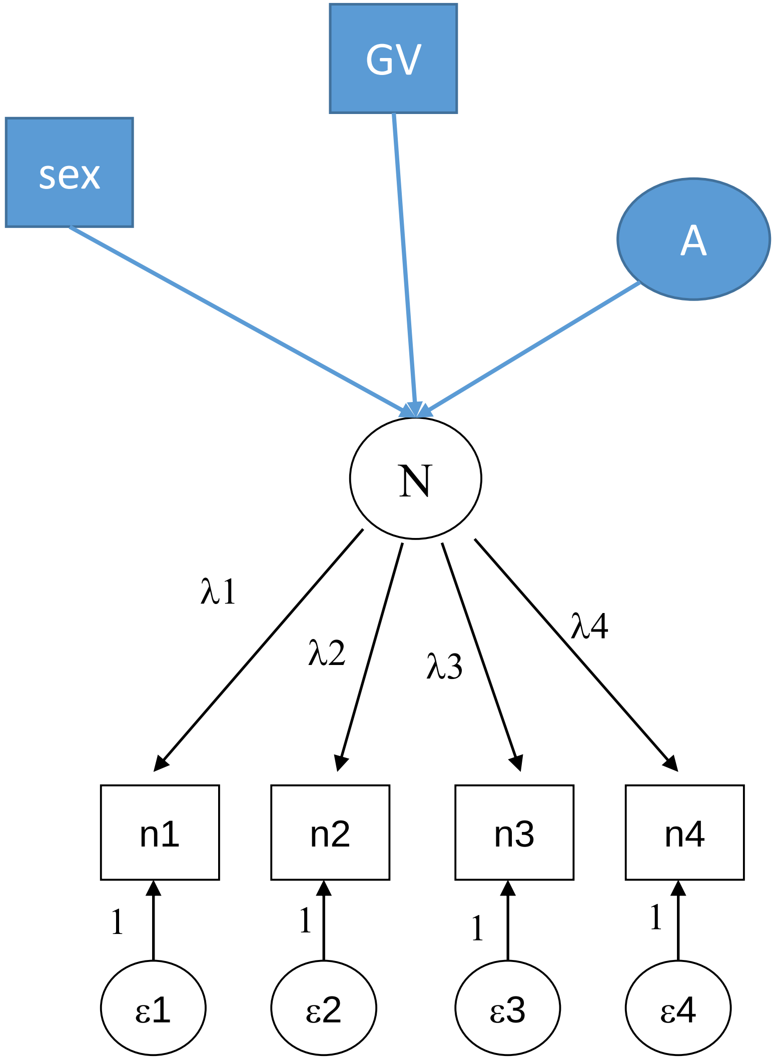
N is **a real, causal, unidimensional** source of individual differences. It **exists beyond the realm of the indicator set**, and is not dependent on any given indicator set.

Causal - part I: **The position of N determines causally the response to the items. N is the only direct cause of systematic variation in the items**. I.e., if you condition on N, then the correlations among the items are zero: local independence.



Causal - part I: **The position of N determines causally the response to the items. N is the only direct cause of systematic variation in the items**. I.e., if you condition on N, then the correlations among the items are zero: local independence (as it is called in psychometrics).

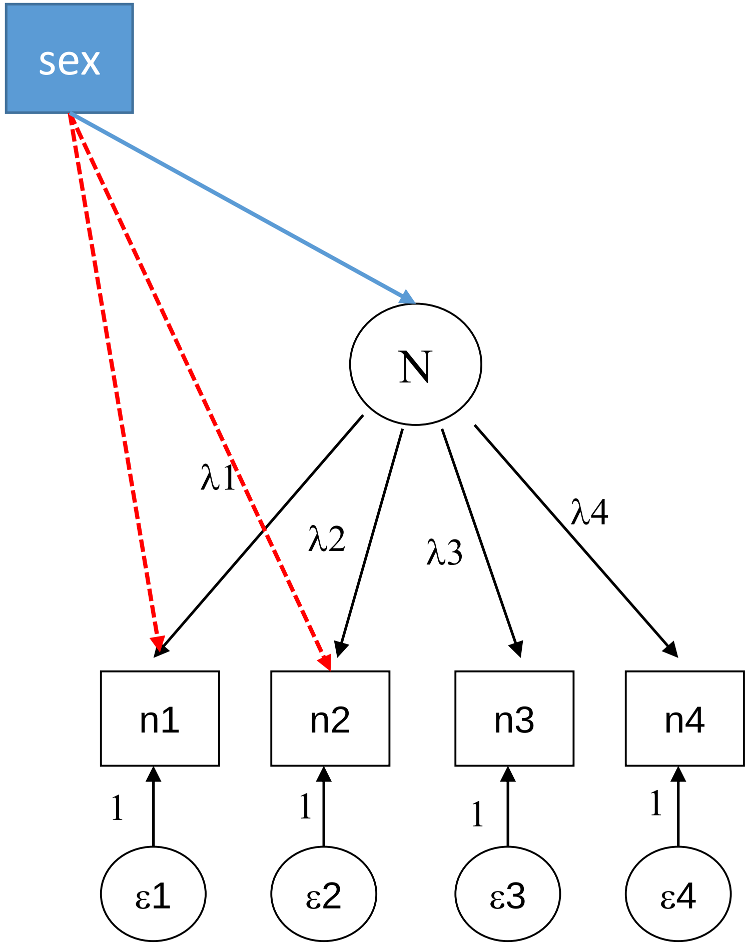
Reflective indicators:
They reflect the causal action of the latent variable N



A substantive aspect of the common factor model: interpretation (you bring to the model).

Causal part II: The relationship between any external variable (latent or observed) and the indicators is **mediated by the common factor N**: essence of “measurement invariance”.

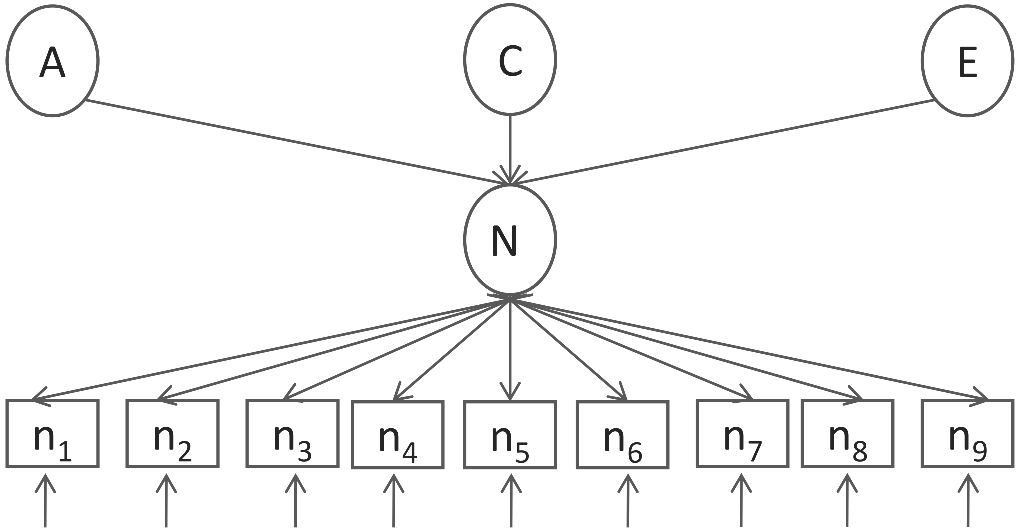
If you condition on N, then the correlation between the external variables and the indicators is zero.



Direct relationships are supposed to be absent.

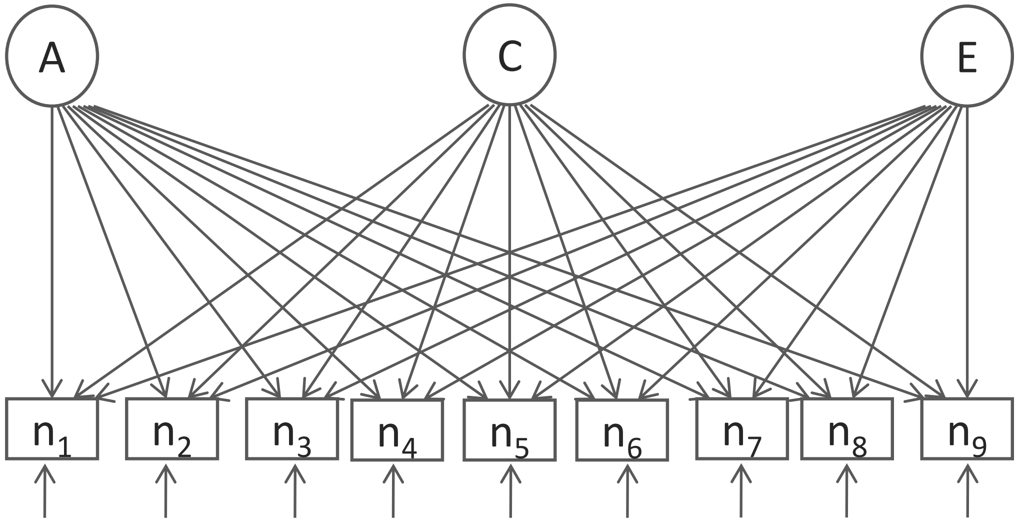
(these destroy unidimensionality….)

Twin design affords an omnibus test of the mediatory role of N



Common pathway model
Psychometric model

Phenotypic unidimensionality
N mediates all external sources of individual differences

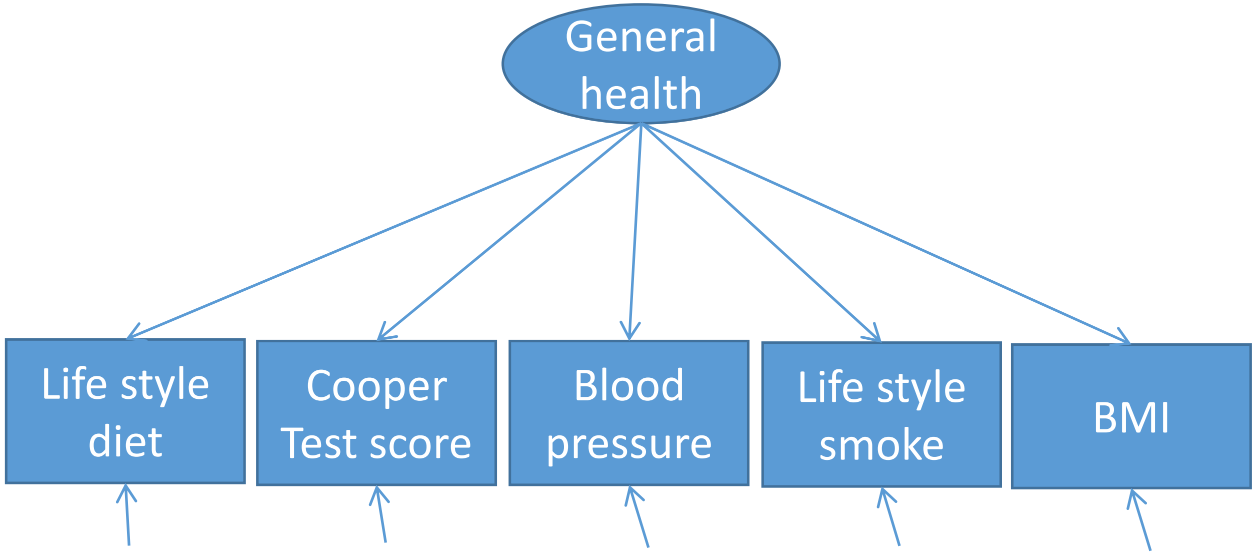


Independent pathway model
Biometric model

Phenotypic multidimensionality…..

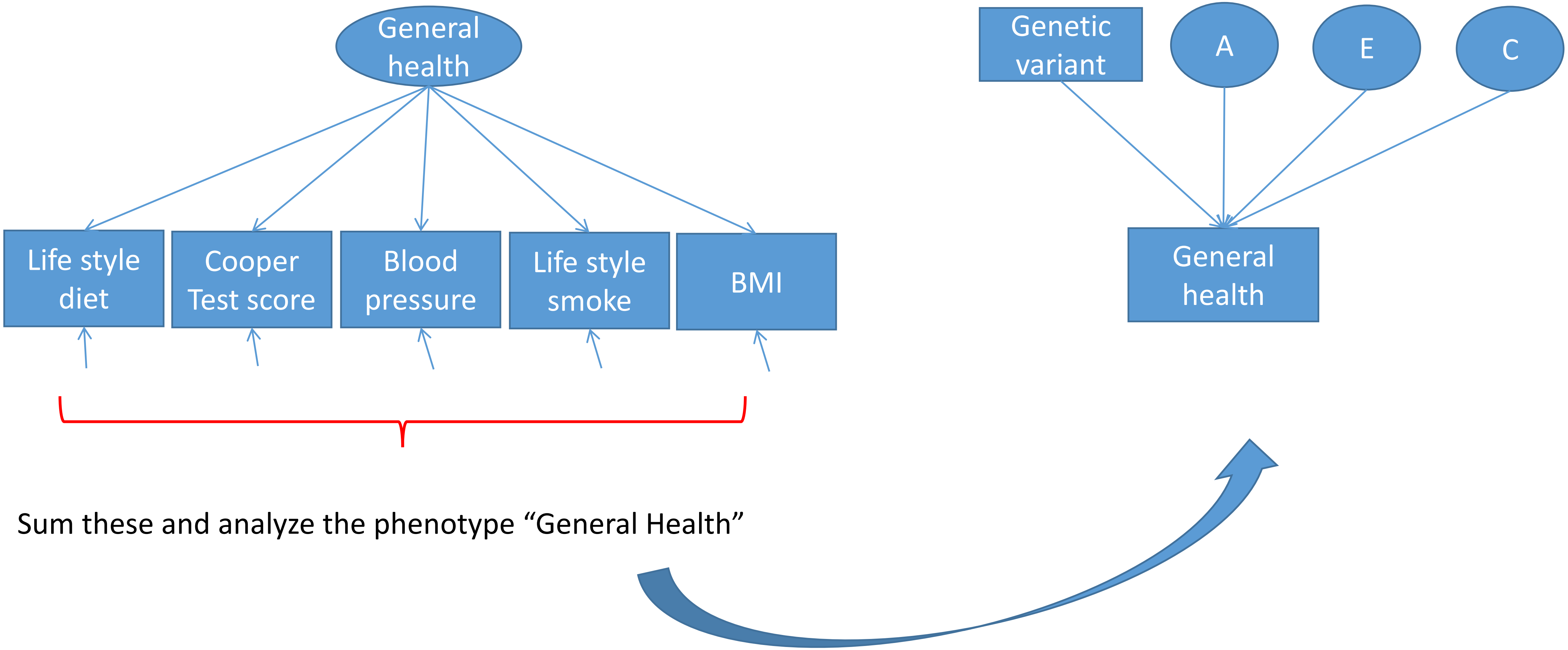
What about N in the phenotypic analysis?
The phenotypic model was incorrect!

A different interpretation: factor analysis as a data summary
Just a way to reduce multiple phenotypes into a single index.
(Alternative statistical technique used to this end: principal component analysis; PCA)



Formative variable. No causal interpretation: General Health does not cause smoking! Common pathway model is not going to fit the data!

# When to use a sum score?



General health

Life style diet

Cooper Test score

Blood pressure

Life style smoke

BMI

Sum these and analyze the phenotype “General Health”

Genetic variant

A

E

C

General health

Back to the common factor model !

Multiple common factors, CFA vs. EFA with rotation

EFA (always more than one common factor).

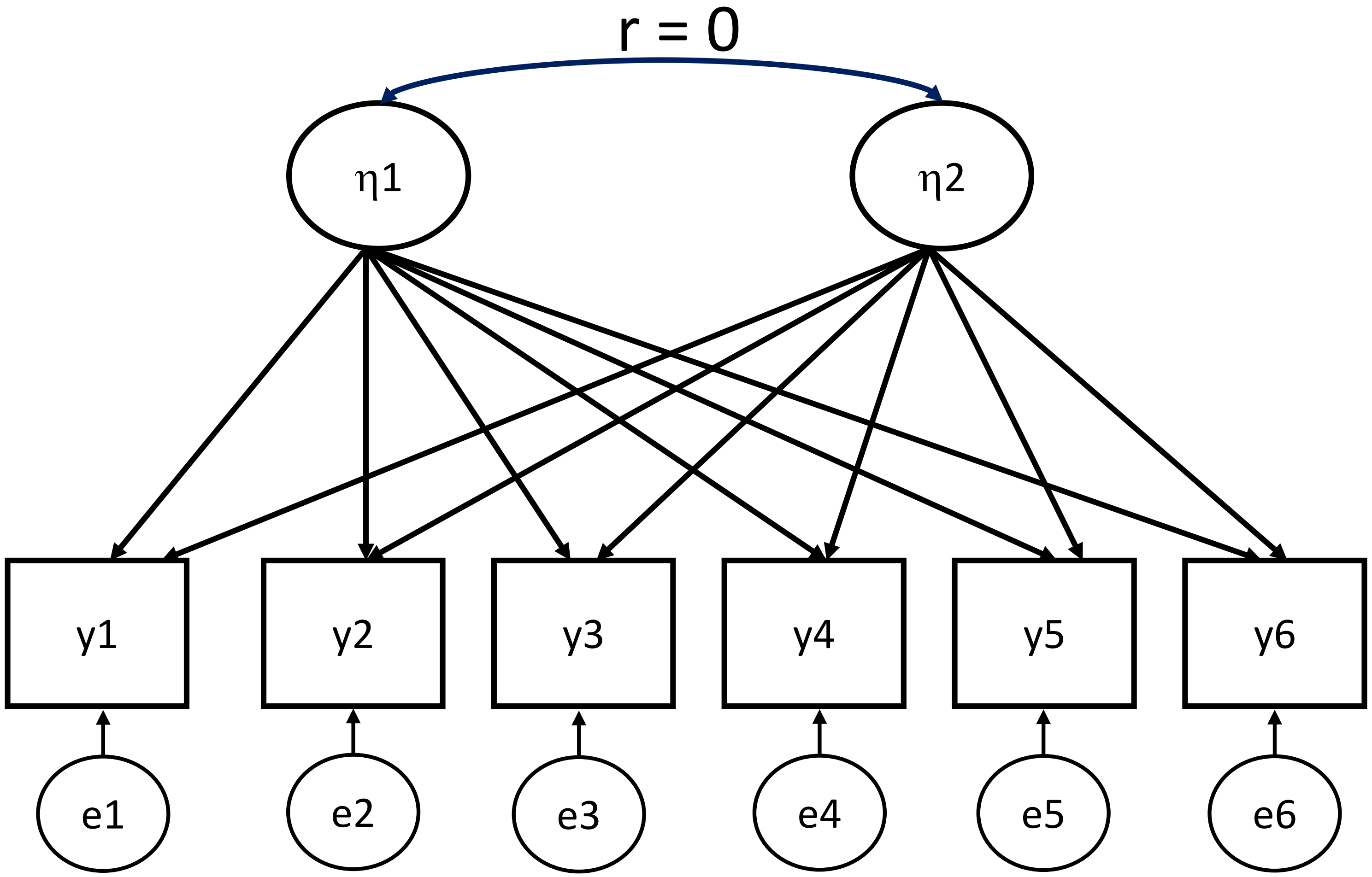
**Aim**: determine dimensionality and derive meaning of factors from loadings

Exploratory approach: How many latent variables? What is the pattern of factor loadings? Low on prior theory, but still involves choices.

How many latent variables: Screeplot, Eigenvalue > 1 rule, Goodness of fit measures ($\chi^2$, RMSEA, NNFI), info criteria (BIC, AIC).

Pattern of factor loadings: Type of rotation (varimax, oblimin, many choices!).

# EFA (two) factor model as it is fitted in standard programs



$$\left.\begin{array}{l}
y_1 = \lambda_{11}\,\eta_1 + \lambda_{12}\,\eta_2 + \varepsilon_1 \\
y_2 = \lambda_{21}\,\eta_1 + \lambda_{22}\,\eta_2 + \varepsilon_2 \\
y_3 = \lambda_{31}\,\eta_1 + \lambda_{32}\,\eta_2 + \varepsilon_3 \\
y_4 = \lambda_{41}\,\eta_1 + \lambda_{42}\,\eta_2 + \varepsilon_4 \\
y_5 = \lambda_{51}\,\eta_1 + \lambda_{52}\,\eta_2 + \varepsilon_5 \\
y_6 = \lambda_{61}\,\eta_1 + \lambda_{62}\,\eta_2 + \varepsilon_6
\end{array}\right\}$$

$$\underbrace{\mathbf{y}_i}_{ny \times 1} = \underbrace{\boldsymbol{\Lambda}}_{ny \times ne}\underbrace{\boldsymbol{\eta}_i}_{ne \times 1} + \underbrace{\boldsymbol{\varepsilon}_i}_{ny \times 1}$$

$$\boldsymbol{\eta}^t = [\eta_1\ \eta_2]$$

$$\boldsymbol{\Lambda} = \begin{array}{cc}
\lambda_{11} & \lambda_{12} \\
\lambda_{21} & \lambda_{22} \\
\dots & \dots \\
\lambda_{51} & \lambda_{52} \\
\lambda_{61} & \lambda_{62}
\end{array}$$

$$\begin{array}{ccccc}
\boldsymbol{\Sigma}_{\mathbf{y}} = & \boldsymbol{\Lambda} & \boldsymbol{\Psi} & \boldsymbol{\Lambda}^t & + \ \boldsymbol{\Theta} \\
(ny \times ny) & (ny \times ne) & (ne \times ne) & (ne \times ny) & + \ (ny \times ny)
\end{array}$$

$$\boldsymbol{\Psi} = \mathbf{I} = \begin{array}{cc} 1 & \\ 0 & 1 \end{array}$$

$$\boldsymbol{\Theta} = \mathrm{diag}(\sigma^2_{\varepsilon 1}\ \ \sigma^2_{\varepsilon 2}\ \ \sigma^2_{\varepsilon 3}\ \ \sigma^2_{\varepsilon 4}\ \ \sigma^2_{\varepsilon 5}\ \ \sigma^2_{\varepsilon 6})$$

$$\left.\begin{array}{l}
y_1 = \lambda_{11}\ \eta_1 + \lambda_{12}\ \eta_2 + \varepsilon_1 \\
y_2 = \lambda_{21}\ \eta_1 + \lambda_{22}\ \eta_2 + \varepsilon_2 \\
y_3 = \lambda_{31}\ \eta_1 + \lambda_{32}\ \eta_2 + \varepsilon_3 \\
y_4 = \lambda_{41}\ \eta_1 + \lambda_{42}\ \eta_2 + \varepsilon_4 \\
y_5 = \lambda_{51}\ \eta_1 + \lambda_{52}\ \eta_2 + \varepsilon_5 \\
y_6 = \lambda_{61}\ \eta_1 + \lambda_{62}\ \eta_2 + \varepsilon_6
\end{array}\right\}$$

Meaning of the common factors?
Based on these factor loadings? No!

$\boldsymbol{\Lambda}^t\boldsymbol{\Theta}^{-1}\boldsymbol{\Lambda}$ is diagonal (identifying constraint)

Associated with the identifying constraint: unique values of $\boldsymbol{\Lambda}$, but rotatable.

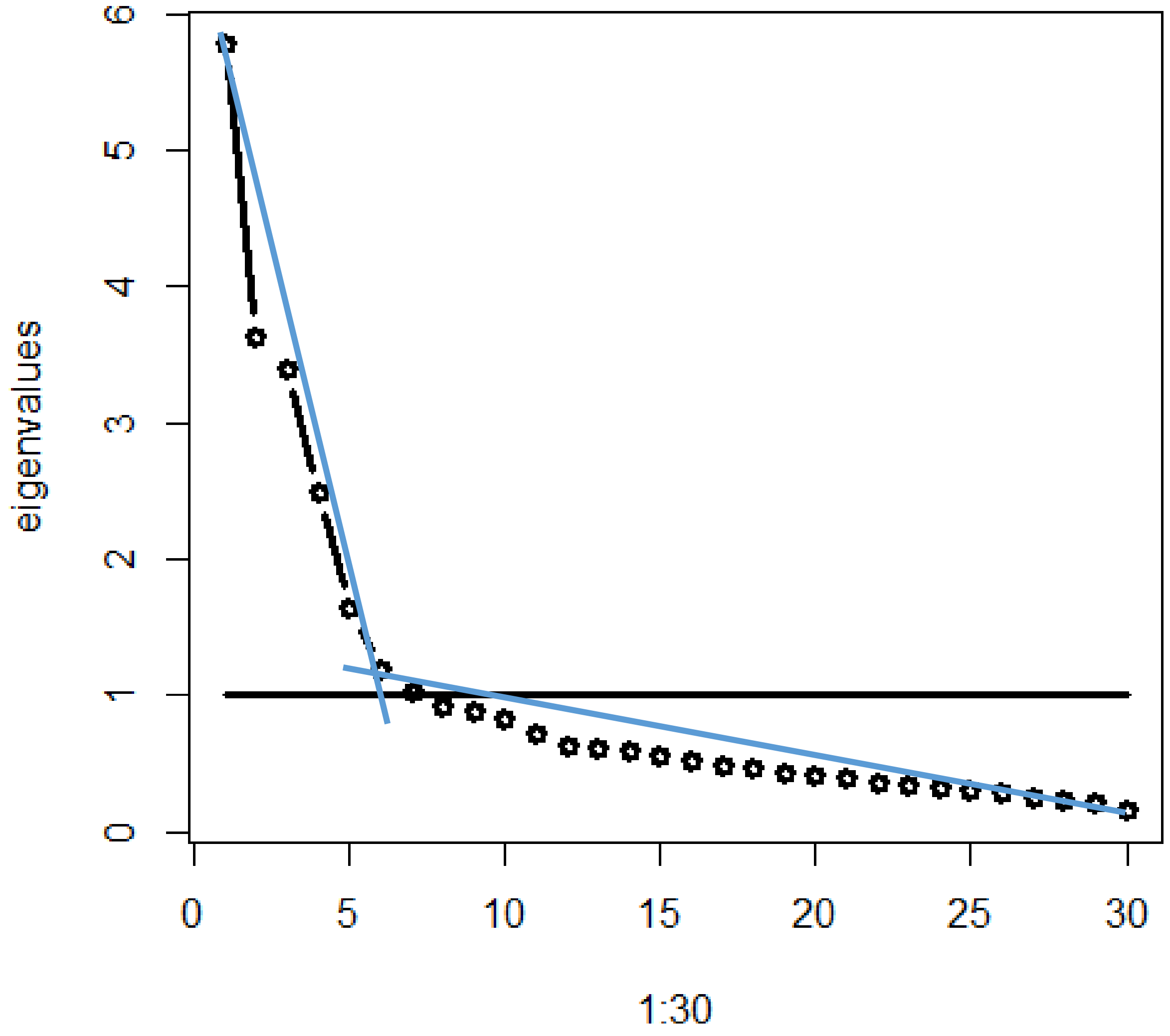
$\boldsymbol{\Lambda}\mathbf{M} = \boldsymbol{\Lambda}^*$, $\mathbf{M}\mathbf{M}^t = \mathbf{I}$, so that $\boldsymbol{\Sigma}_{\mathbf{y}} = \boldsymbol{\Lambda}\ \mathbf{M}\mathbf{M}^t\boldsymbol{\Lambda}^t + \boldsymbol{\Theta} = \boldsymbol{\Lambda}\mathbf{I}\boldsymbol{\Lambda}^t + \boldsymbol{\Theta}$

$\mathbf{\Lambda M} = \mathbf{\Lambda}^*$

$\mathbf{M}$: Rotation matrix is calculated by maximizing a rotation criterion. These minimize of maximize loadings to improve interpretability.

Orthogonal rotation leaves common factors uncorrelated
Oblique rotation allows for correlation.

Rotation is just a transformation of results (no testing!).
E.g., test whether factor correlations are 0 is not possible.

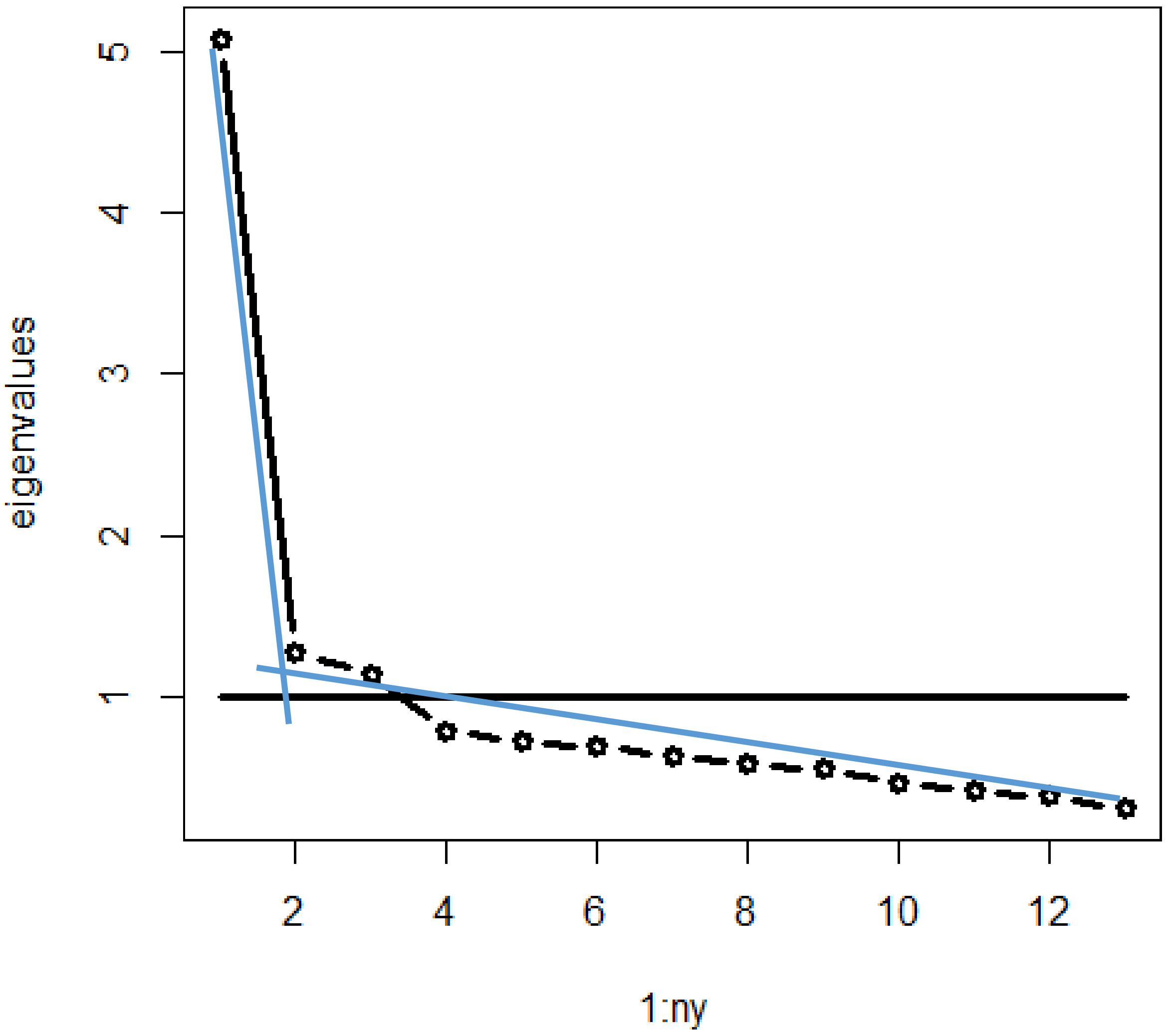


BIG 5 data 361 females students

Screeplot locate the “elbow joint” (5)

Eigenvalues > 1 rule (6?)

5 EFA factor model: Chi2(295) = 822.0



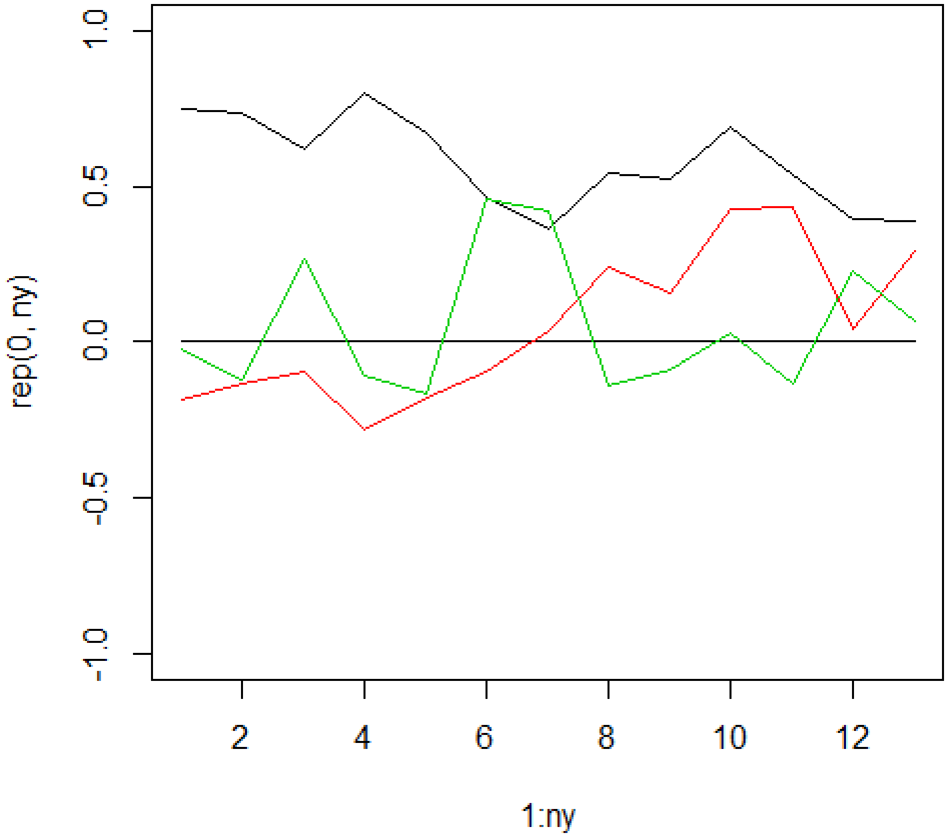
WAIS-III

1868 US whites

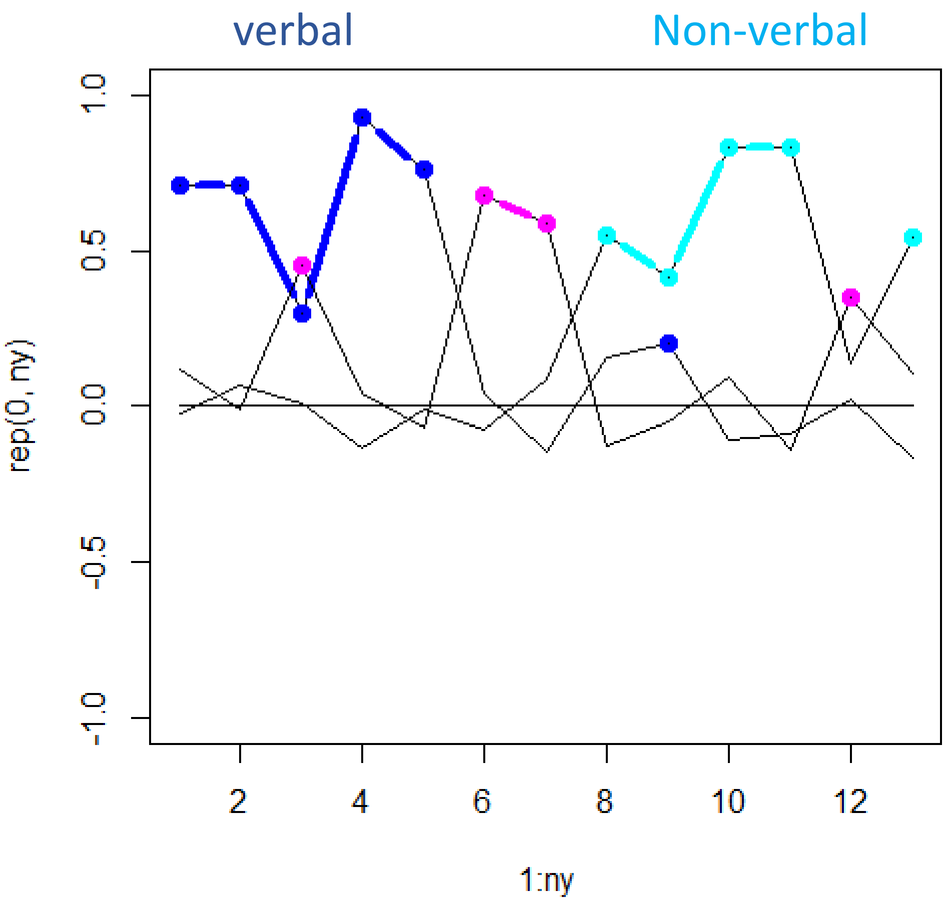
Screeplot locate the “elbow joint” (1)

Eigenvalues > 1 rule (3)

3 EFA factor model: Chi2(42) = 111.9

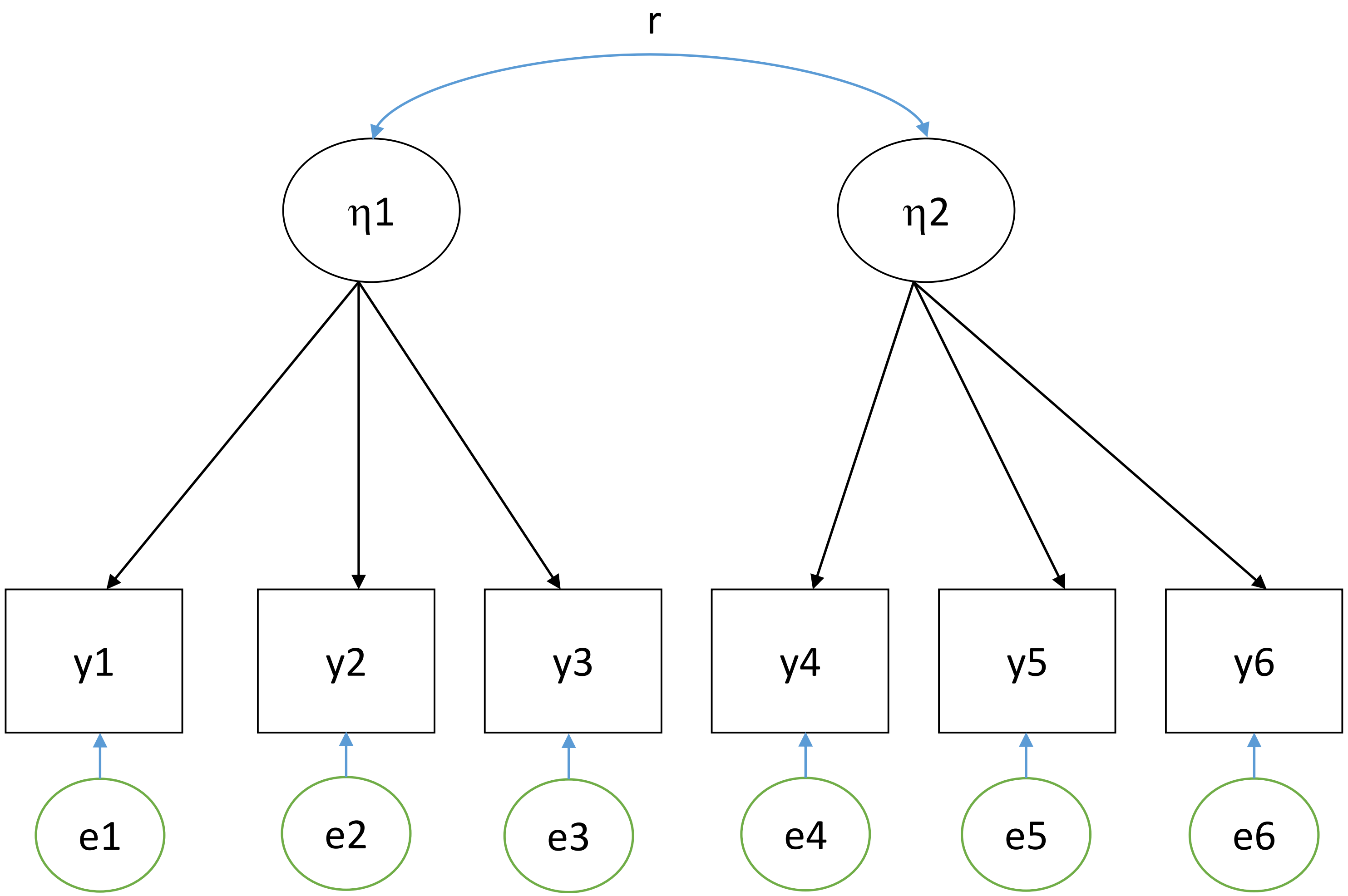


Unrotated
Chi2(42) = 111.9



Promax rotated (oblique)
Chi2(42) = 111.9

# CFA (two) factor model: impose a pattern of loadings based on theory , define the common factors based on prior knowledge.



$$\left.\begin{array}{lcl}
y_1 &=& \lambda_{11}\,\eta_1 + 0\,\eta_2 + \varepsilon_1 \\
y_2 &=& \lambda_{21}\,\eta_1 + 0\,\eta_2 + \varepsilon_2 \\
y_3 &=& \lambda_{31}\,\eta_1 + 0\,\eta_2 + \varepsilon_3 \\
y_4 &=& 0\,\eta_1 + \lambda_{42}\,\eta_2 + \varepsilon_4 \\
y_5 &=& 0\,\eta_1 + \lambda_{52}\,\eta_2 + \varepsilon_5 \\
y_6 &=& 0\,\eta_1 + \lambda_{62}\,\eta_2 + \varepsilon_6
\end{array}\right\}$$

$$\underset{ny \times 1}{\mathbf{y}_i} = \underset{ny \times ne}{\boldsymbol{\Lambda}}\,\underset{ne \times 1}{\boldsymbol{\eta}_i} + \underset{ny \times 1}{\boldsymbol{\varepsilon}_i}$$

$$\boldsymbol{\eta}^t = [\eta_1\ \eta_2]$$

$$\boldsymbol{\Lambda} = \begin{array}{cc}
\lambda_{11} & 0 \\
\lambda_{21} & 0 \\
\dots & \dots \\
0 & \lambda_{52} \\
0 & \lambda_{62}
\end{array}$$

$$\underset{(ny \times ny)}{\boldsymbol{\Sigma}_{\mathbf{y}}} = \underset{(ny \times ne)}{\boldsymbol{\Lambda}}\,\underset{(ne \times ne)}{\boldsymbol{\Psi}}\,\underset{(ne \times ny)}{\boldsymbol{\Lambda}^t} + \underset{(ny \times ny)}{\boldsymbol{\Theta}}$$

$$\boldsymbol{\Psi} = \begin{array}{cc}
1 & \\
\rho & 1
\end{array}$$

$$\boldsymbol{\Theta} = \mathrm{diag}(\sigma^2_{\varepsilon 1}\ \ \sigma^2_{\varepsilon 2}\ \ \sigma^2_{\varepsilon 3}\ \ \sigma^2_{\varepsilon 4}\ \ \sigma^2_{\varepsilon 5}\ \ \sigma^2_{\varepsilon 6})$$

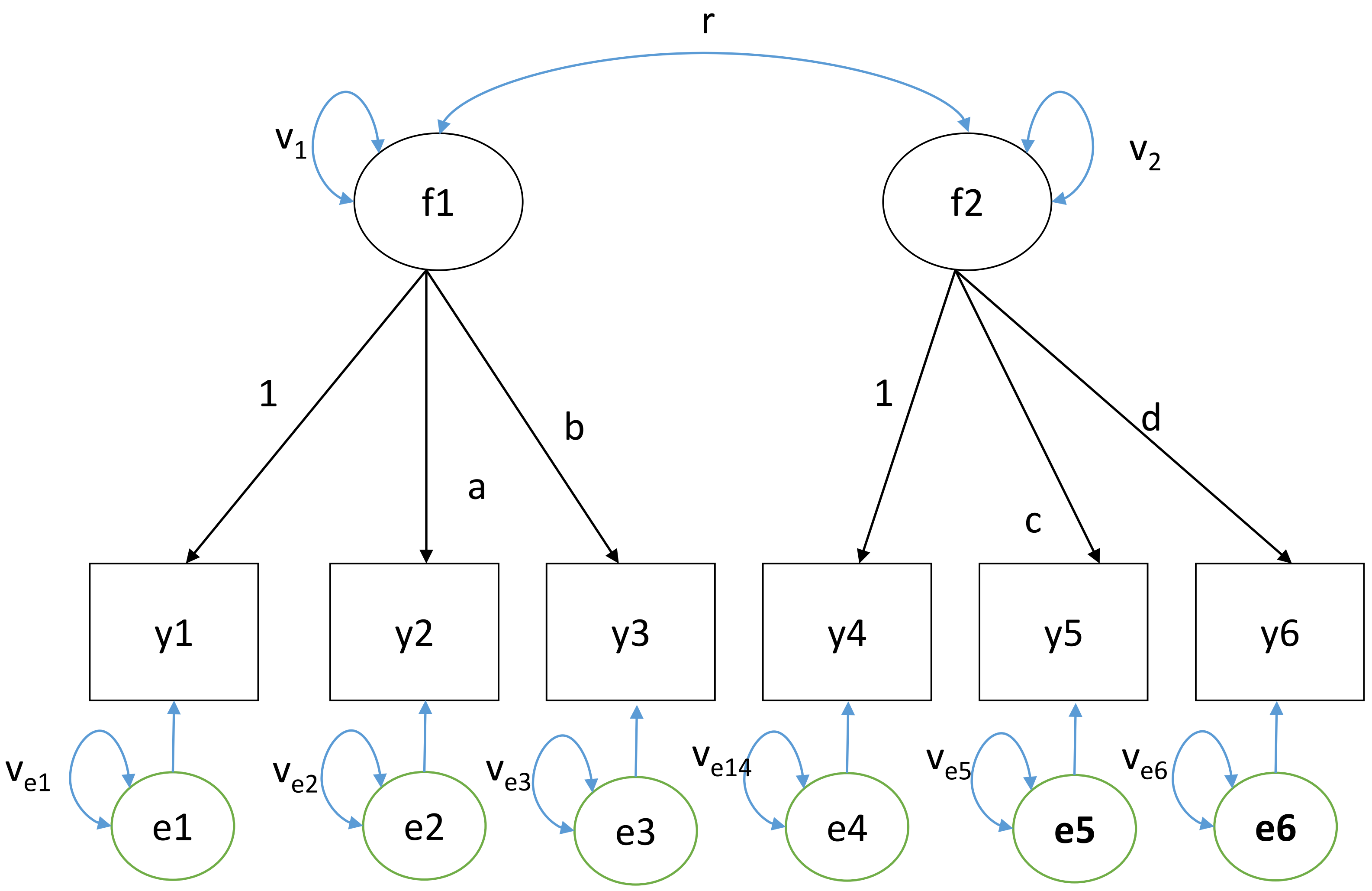
$$\Sigma_{\mathbf{y}} = \quad \Lambda \quad \Psi \quad \Lambda^t \quad + \quad \Theta$$

(ny x ny) (ny x ne)(ne x ne)(ne x ny) + (ny x ny)

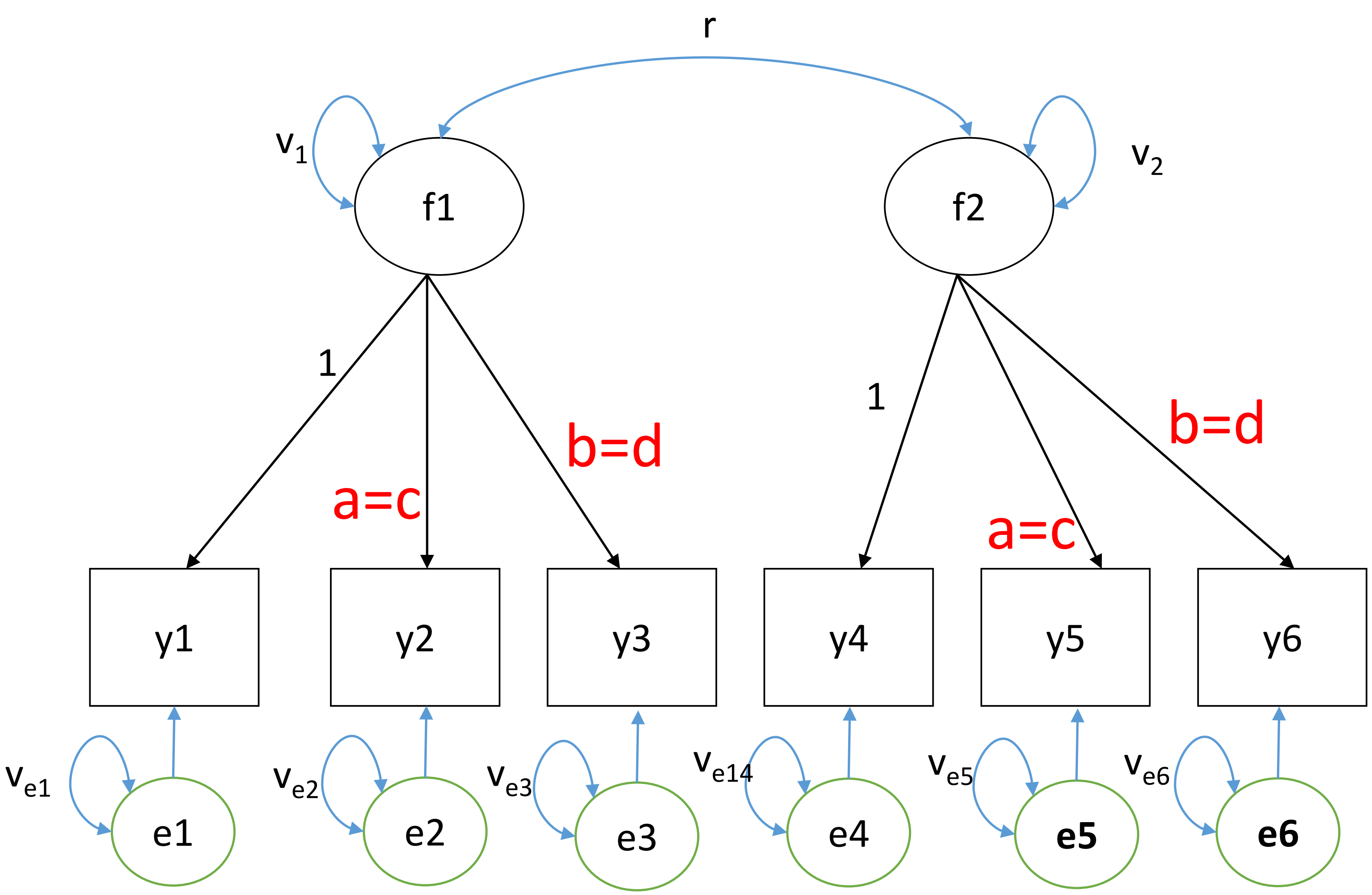
In CFA, in contrast to EFA, you can impose all kinds of constraints on the parameters

In CFA, in constrast to EFA, you can estimate off-diagonal elements in the cov matrix of the residuals $\Theta$

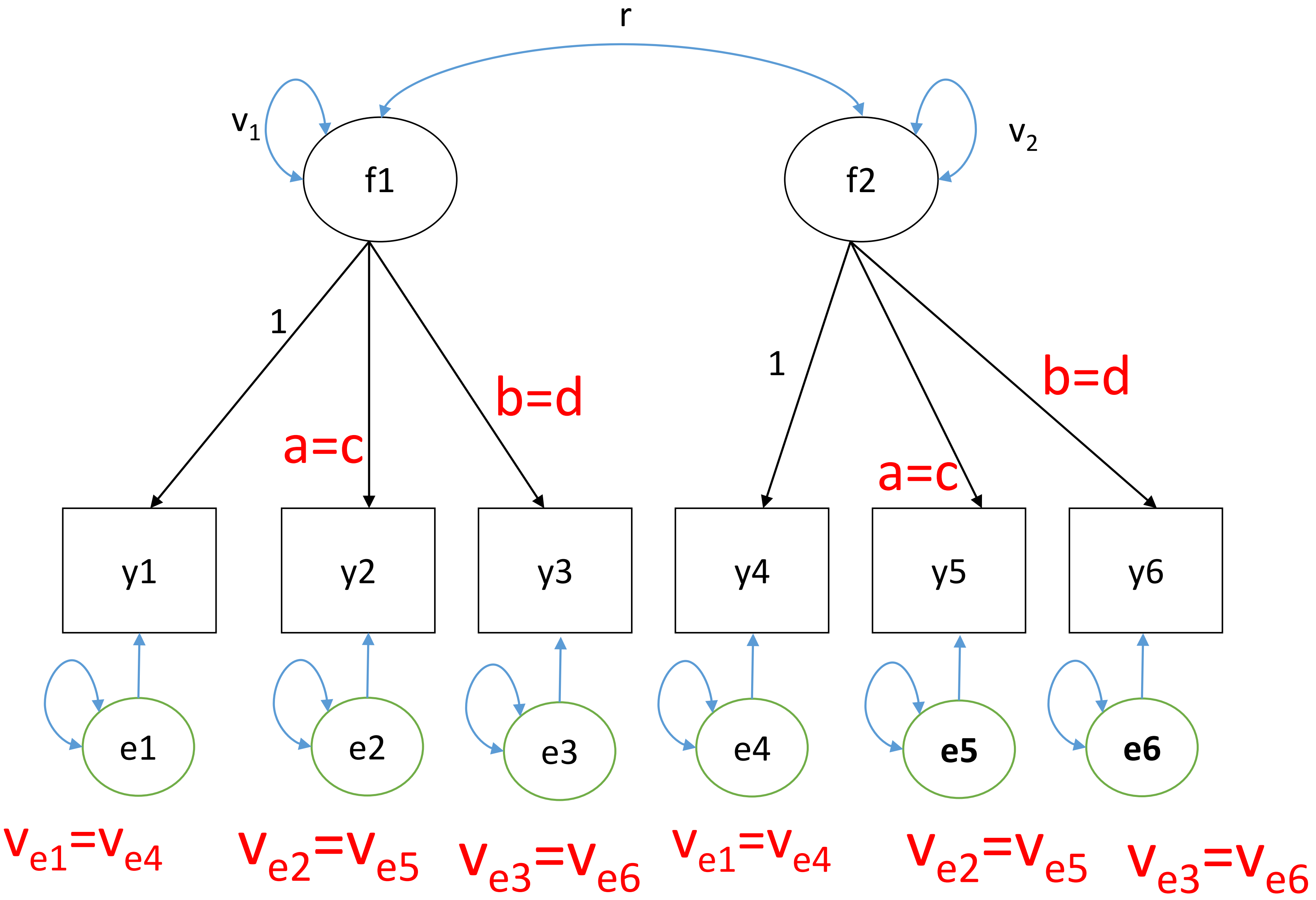
# Suppose 3 indicators at 2 time points



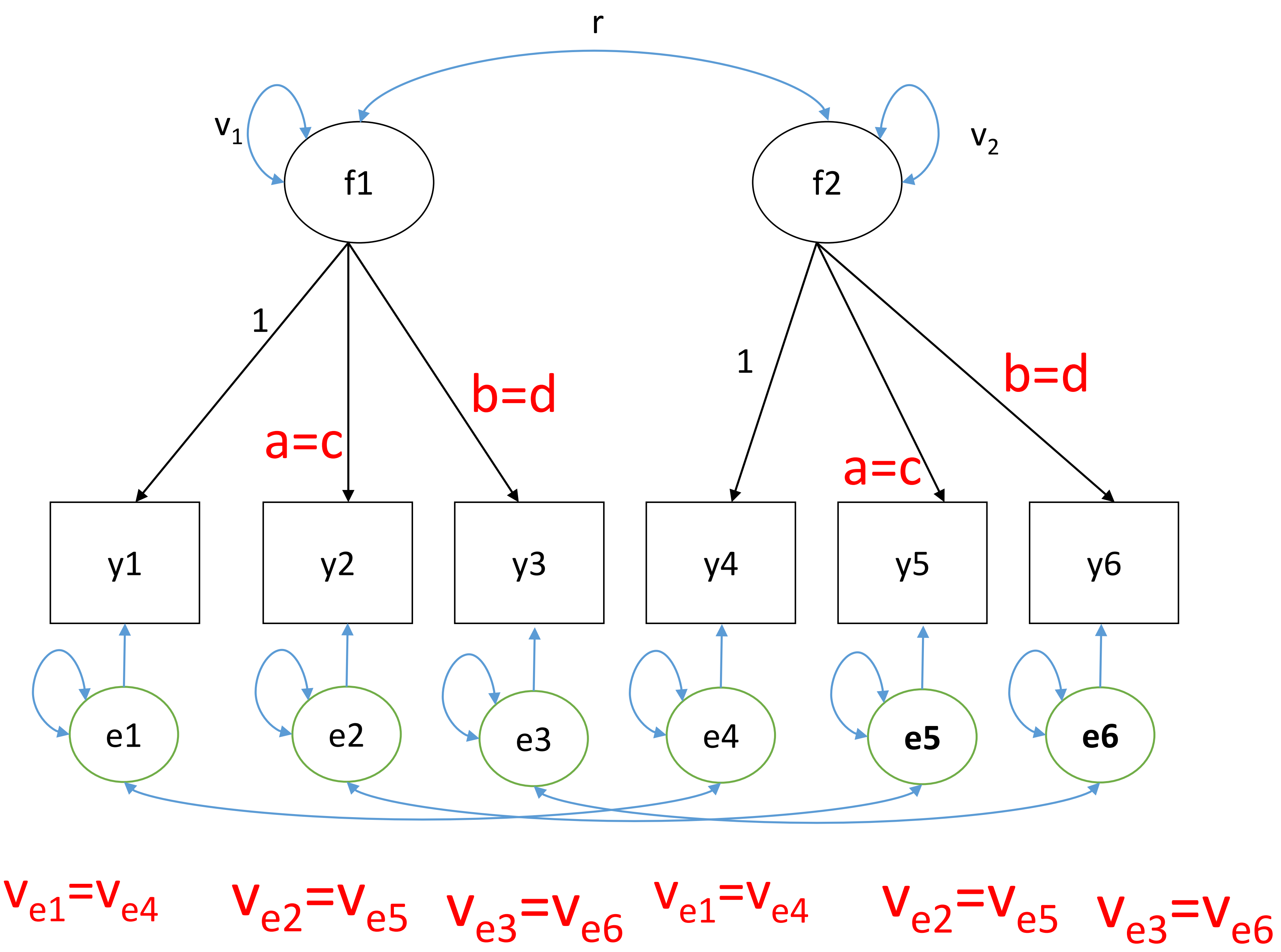
# Suppose 3 indicators at 2 time points

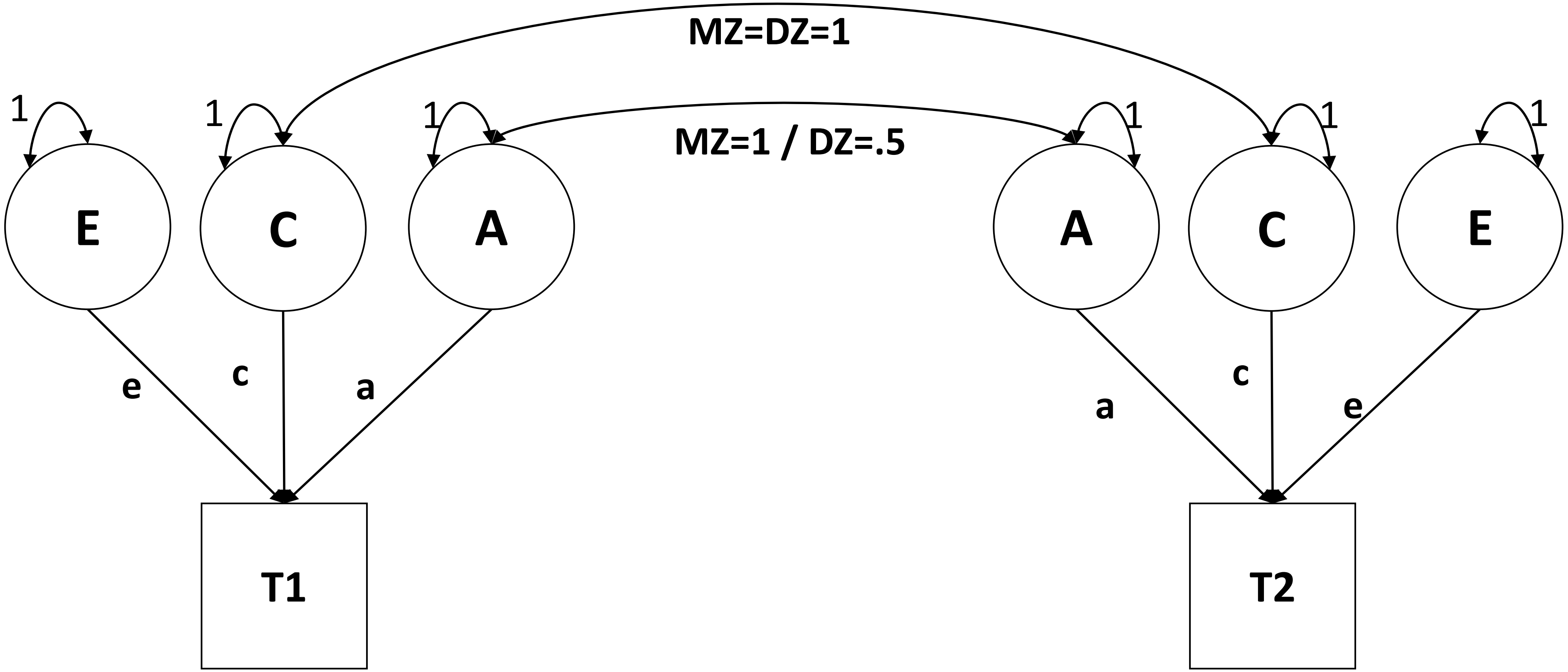


# Suppose 3 indicators at 2 time points



# Suppose 3 indicators at 2 time points





Equality constraints are nothing new!

The linear common factor model: “continuous” indicators (7 point Likert scale is “continuous”)

What about ordinal or binary indicators?

Linear regression model is key ingredient in the linear factor model

Logistic or probit regression is a key ingredient in ordinal or discrete factor analysis.

The model for the $Y_i = b0 + b1*X_i + e_i$

$E[Y|X=x^{\circ}] = b0 + b1*x^{\circ}$

Logit:

$E[Z|X=x^{\circ}] = \text{Prob}(Z=0|X=x^{\circ}) = \exp(b0 + b1*x^{\circ}) / \{1 + \exp(b0 + b1*x^{\circ})\}$

Probit:

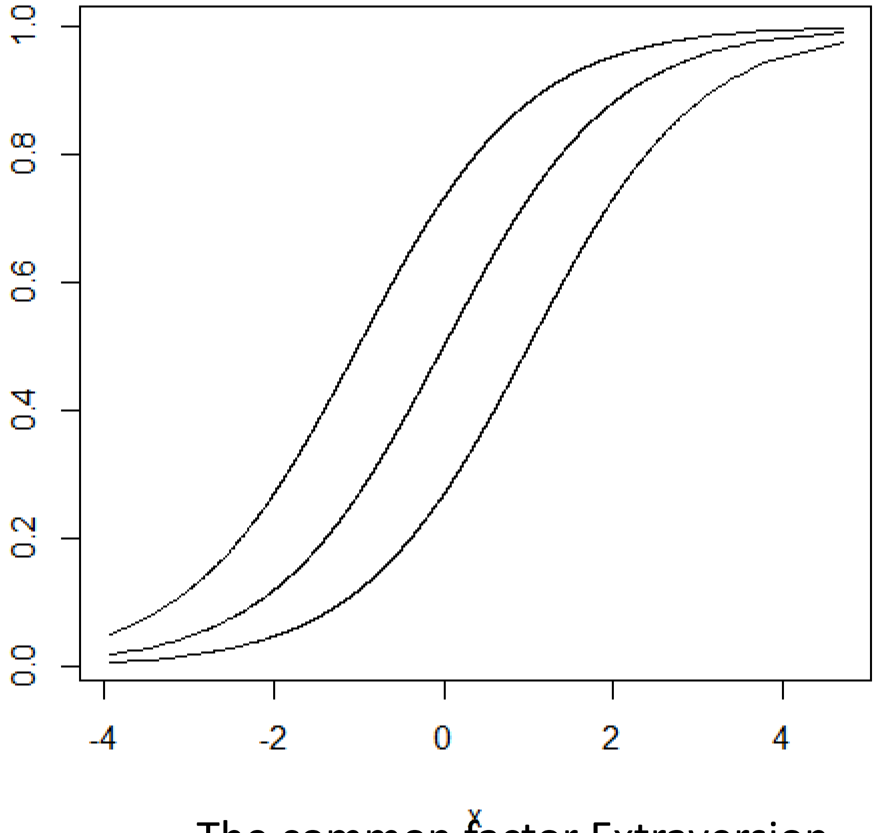
$E[Z|X=x^{\circ}] = \text{Prob}(Z=0|X=x^{\circ}) = \Phi(b0 + b1*x^{\circ})$, $\Phi(.)$ cumulative st. normal distribution

Replace X, observed predictor, but $\eta$, the common factor.

Do you like to go to parties? (y/n)

Prob(yes|X=x˚)

The probability of endorsement



The common factor Extraversion