# Phenotypic factor analysis

Conor V. Dolan & Abdel Abdellaoui

Biologische Psychologie, VU, Amsterdam

Boulder Workshop - March 2016

# Phenotypic factor analysis

A psychometric statistical technique to investigate the dimensionality of related variables in terms of common latent variables (a.k.a. common factors).

A data reduction statistical technique to summarize or reduce a number related variables to one or a few summary variables. Not a causal model.

# 16 depression items (with response categories 5)

- 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

**A psychometric analysis**:

Investigate the dimenionality of the item responses in terms of substantive latent variables. A causal hypothesis: the latent variable ("depression") causes the item response.

**A data reduction analyses**:

Reduce the dimensionality of the data from 16 variables to a 1 (or 2 or 3) variables, while retaining as much information as possible (principal component analysis; PCA).

The linear common factor model

as a statistical (regression) model - formal representation
as a causal – psychometric - model

- what is a common factor substantively?
- implication in terms of data summary and causal modeling
**- why is the 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

…and of many statistical models including
genetic covariance structure modeling

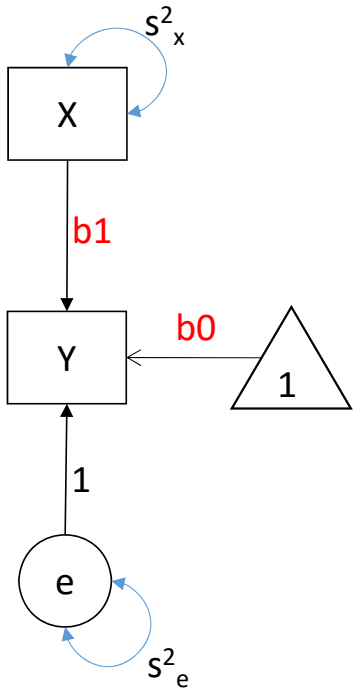
# Path diagram regression model “regression of Y on X”:

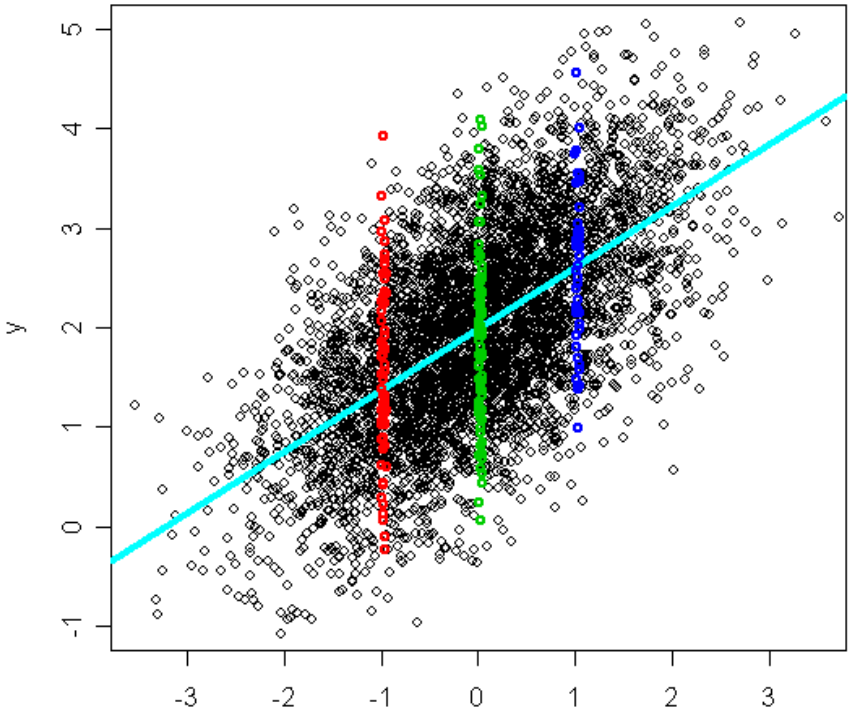
The model: $Y_i = b0 + b1*X_i + e_i$

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:

$$\begin{array}{c} \\ x \\ y \end{array} \begin{array}{c} \begin{array}{cc} x & y \end{array} \\ \begin{pmatrix} s^2_x & s_{xy} \\ s_{xy} & s^2_y \end{pmatrix} \end{array} = \begin{array}{c} \begin{array}{cc} x & y \end{array} \\ \begin{pmatrix} s^2_x & b1*s^2_x \\ b1*s^2_x & b1^2*s^2_x + s^2_e \end{pmatrix} \end{array}$$



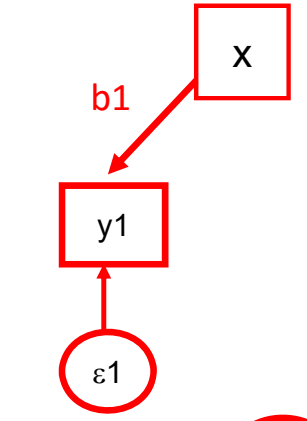



**red**: the data points of cases with x=-1
**green**: the data points of cases with x=0
**dark blue**: the data points of cases with x=1

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

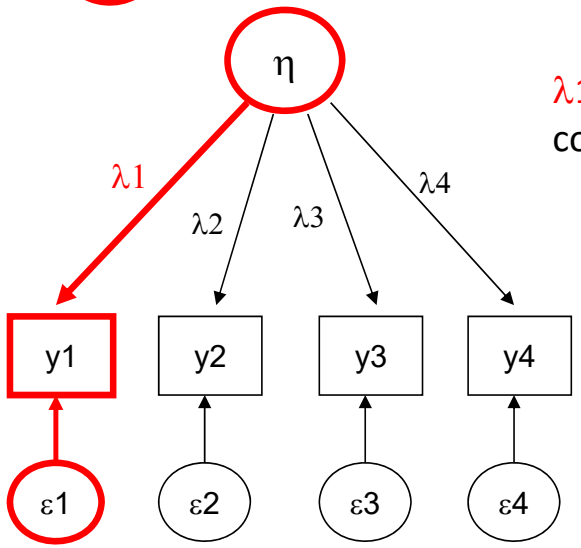
# Single common factor model: A set of linear regression equations

$Y_i = b0 + b1*X_i + e_i$



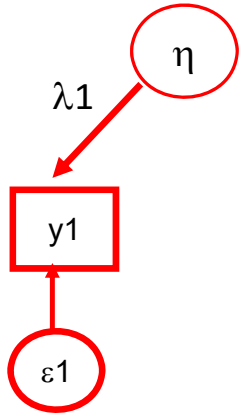

b1 is a regression coefficient (slope parameter)

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




λ1 is a factor loading (a regression coefficient by a different name!)

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




The implied model for the covariance matrix:

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

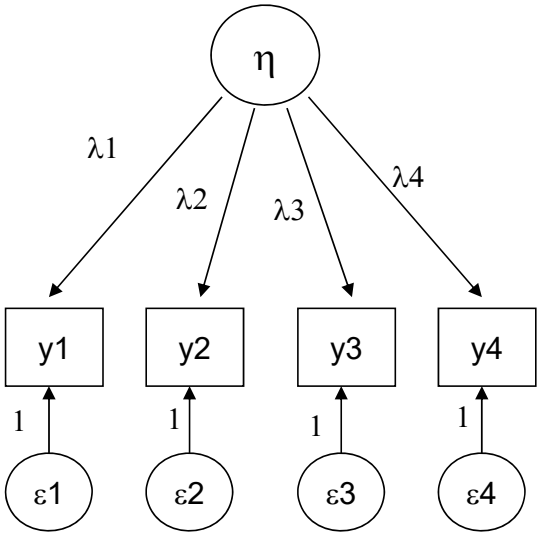
$$\text{Mean}(y_1) = \tau_1 + \lambda_1 \text{Mean}(\eta) = \tau_1$$

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

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**

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

Matrix algebra is
1) Notationally Efficient
2) Basis of Multivariate Statistics

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

ny number of variables
ne number of common factors

$$\mathbf{y}_i - \mathbf{t} = \Lambda \eta_i + \varepsilon_i$$

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

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

ny x 1 1 x 1

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

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

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

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

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

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

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

$$y_{i1} = \lambda_1 \eta_i + \varepsilon_{i1},$$
$$y_{i2} = \lambda_2 \eta_i + \varepsilon_{i2},$$
$$y_{i3} = \lambda_3 \eta_i + \varepsilon_{i3},$$
$$y_{i4} = \lambda_4 \eta_i + \varepsilon_{i4}.$$

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

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

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

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

$$y_{i1} = \lambda_1 \eta_i + \varepsilon_{i1},$$
$$y_{i2} = \lambda_2 \eta_i + \varepsilon_{i2},$$
$$y_{i3} = \lambda_3 \eta_i + \varepsilon_{i3},$$
$$y_{i4} = \lambda_4 \eta_i + \varepsilon_{i4}.$$

ny number of variables
ne number of common factors

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

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

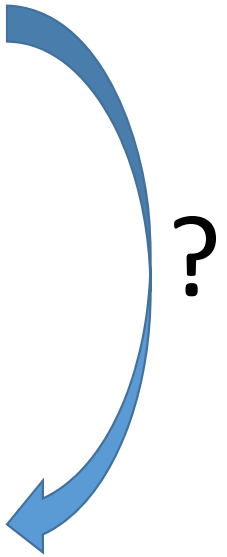
$$\Sigma_{\mathbf{y}} = E[\mathbf{y}\mathbf{y}^t] = E[(\Lambda\eta_i + \varepsilon_i)(\Lambda\eta_i + \varepsilon_i)^t] = \Lambda\Psi\Lambda^t + \Theta$$

$E[\eta_i \eta_i^t] = \Psi$ and $E[\varepsilon_i \varepsilon_i^t] = \Theta$

You can represent this model in OpenMx using matrices

| | 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 (University of Amsterdam)

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

$\Sigma_y = \Lambda E[\eta_i \eta_i^t]\Lambda^t + E[\varepsilon_i \varepsilon_i^t] = \Lambda\Psi\Lambda^t + \Theta$

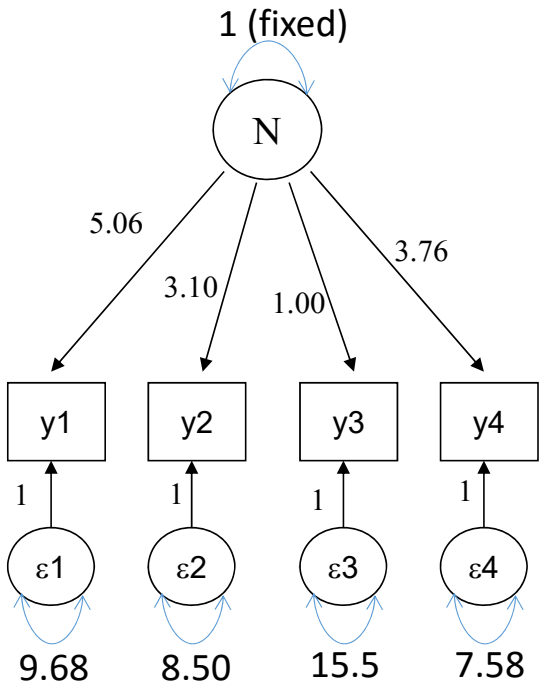
$\Lambda^t$ = [5.059 3.100 1.005 3.758]
diag($\Theta$) = [9.684 8.500 15.584 7.587]
$\Psi$ = [1] (fixed)

Observed covariance matrix

| | | | |
|---|---|---|---|
| 35.278 | | | |
| 15.763 | 18.109 | | |
| 4.942 | 2.661 | 16.594 | |
| 18.970 | 11.622 | 4.262 | 21.709 |

Expected covariance matrix ($\Sigma_y$)

| | | | |
|---|---|---|---|
| 35.278 | | | |
| 15.682 | 18.109 | | |
| 5.085 | 3.115 | 16.594 | |
| 19.011 | 11.649 | 3.777 | 21.709 |




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

var(n1) = $5.06^2$ + 9.68 =35.27
rel(n1) = $5.06^2$ / 35.27 = .725
($R^2$ in regression of y1 on N)

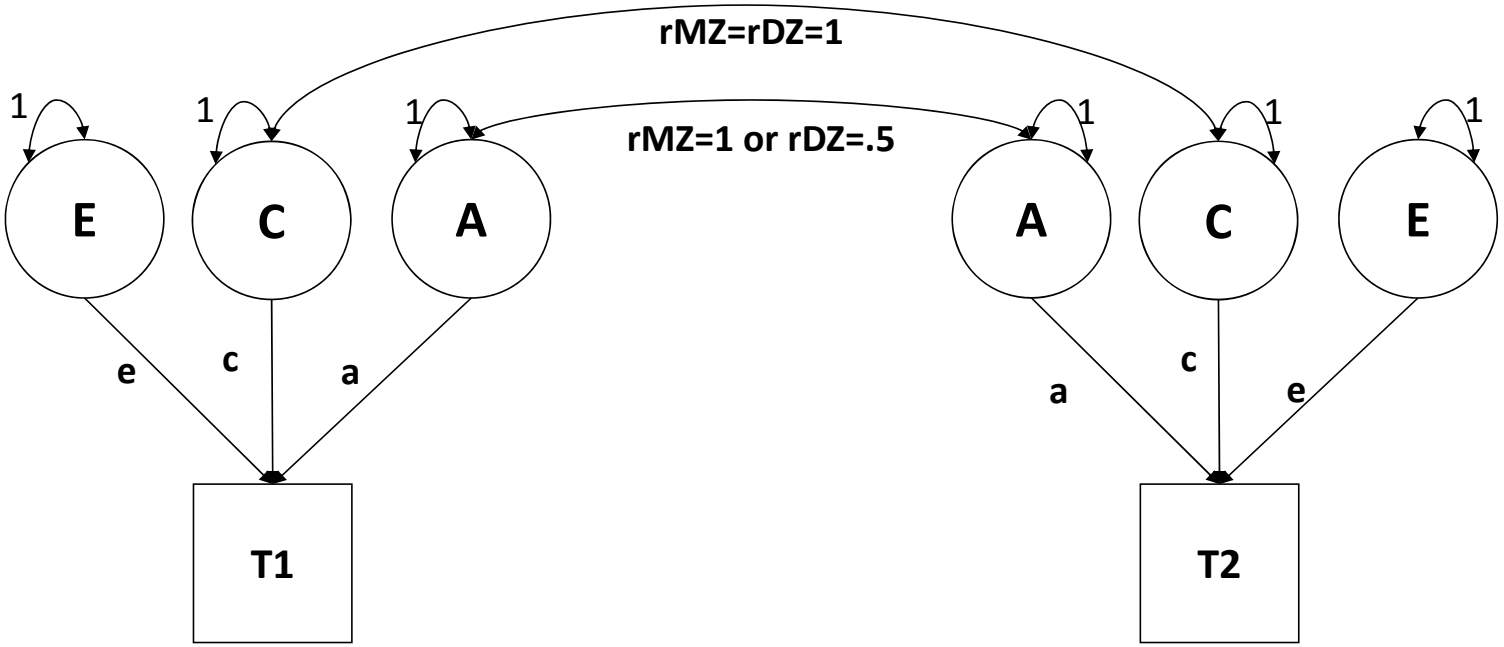
What about y3?

## 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.

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

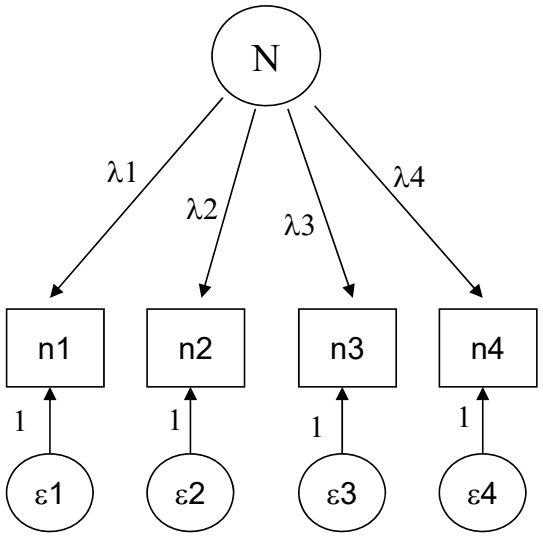
Or making it dependent on an indicator by fixing a factor loading to a "sensible value", i.e., 1 ($\lambda_1$=1)

In summary:

The single common factor model can be used to test the hypothesis that the observed covariance or correlation matrix (say 4 neuroticism items) is consistent with a single common source of variance, the common factor (i.e., neuroticism).

This is established by estimating the parameters of the singel factor model and testing whether the observed covariance matrix equals the covariance matrix based on the estimated parameters (the “implied” or “expected” covariance matrix)

The parameters can also be used to evaluated the quality of the items or indicators ($R^2$: now much of the indicator variance is explained by the common factor?)



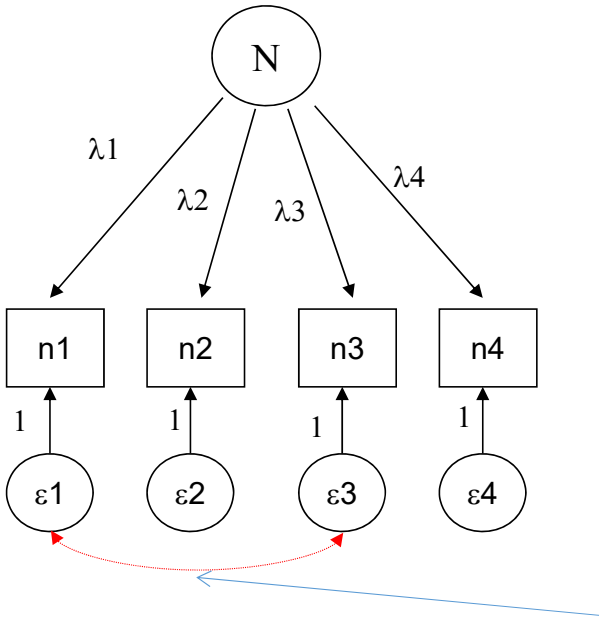

**Reflective indicators**:
They reflect the causal action of the latent variable N

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

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.

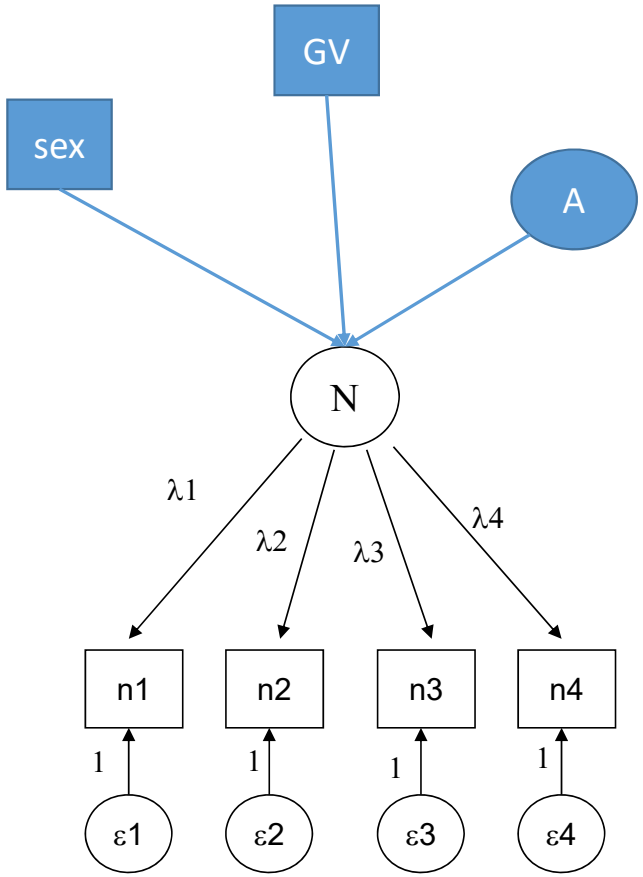
**Causal - part I**: **The position of N determines causally the response to the items. N is the only direct cause of systematic variation in the items**. I.e., if you condition on N, then the correlations among the items are zero: local independence (ε1,ε2,ε3,ε4 are uncorrelated).




**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).

violation of local independence (correlated residuals 1 & 3)

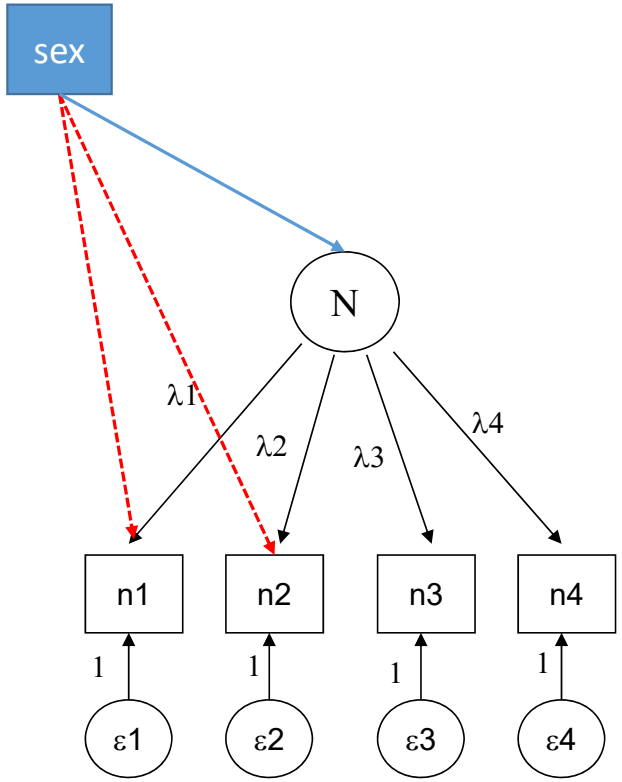
**Reflective indicators**: They **reflect** the causal action of the latent variable N




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

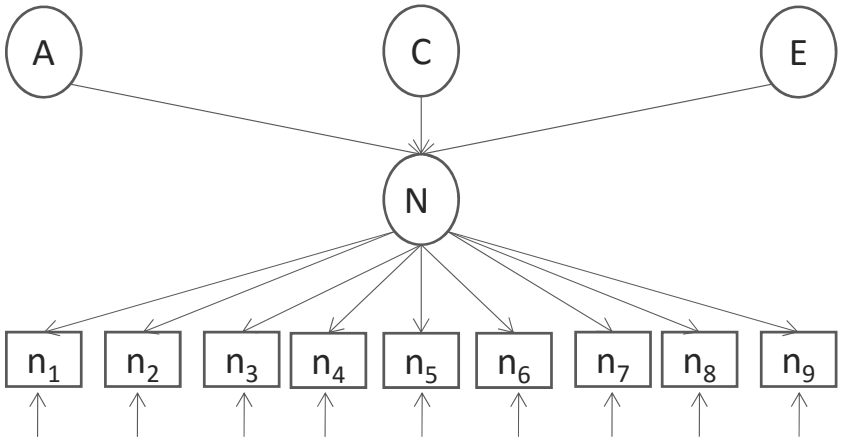
**Causal part II**: The relationship between any external variable (latent or observed) and the indicators is **mediated by the common factor N**: essence of "measurement invariance".

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



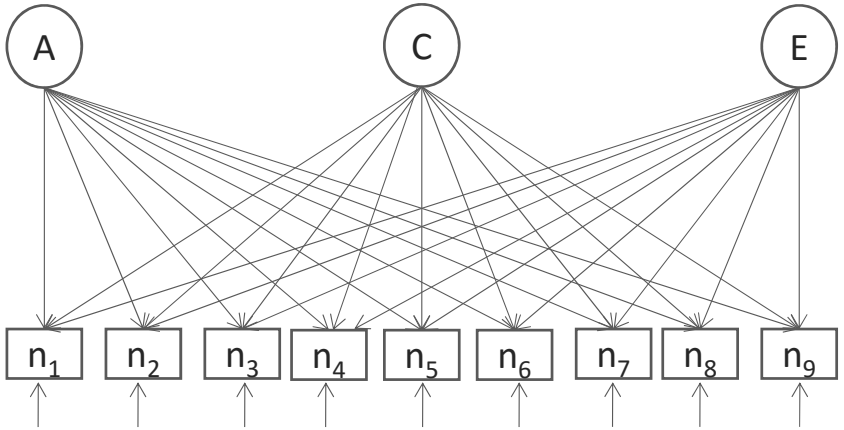

Direct relationships are supposed to be absent. (these destroy unidimensionality….)

Interestingly, the classical 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 or Biometric model

Implies phenotypic multidimensionality…..

What about N in the phenotypic analysis? The phenotypic (1 factor) model was incorrect!

Psychological Methods


1082-989X/13/$12.00 DOI: 10.1037/a0032755

# Can Genetics Help Psychometrics? Improving Dimensionality Assessment Through Genetic Factor Modeling

Sanja Franić
Vrije Universiteit Amsterdam

Conor V. Dolan and Denny Borsboom
University of Amsterdam

James J. Hudziak
University of Vermont

Catherina E. M. van Beijsterveldt and Dorret I. Boomsma
Vrije Universiteit Amsterdam

Behav Genet
DOI 10.1007/s10519-013-9628-4

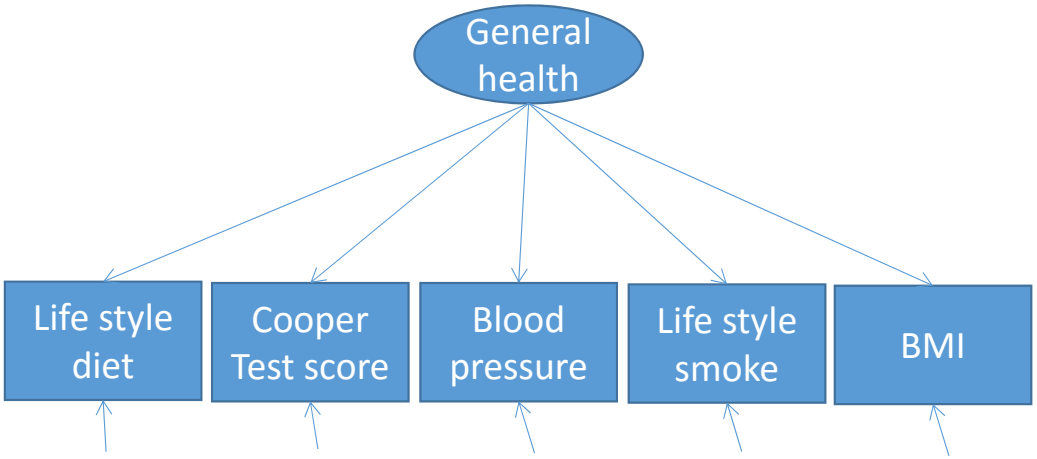
ORIGINAL RESEARCH

# Three-and-a-Half-Factor Model? The Genetic and Environmental Structure of the CBCL/6–18 Internalizing Grouping

**Sanja Franić · Conor V. Dolan · Denny Borsboom · Catherina E. M. van Beijsterveldt · Dorret I. Boomsma**

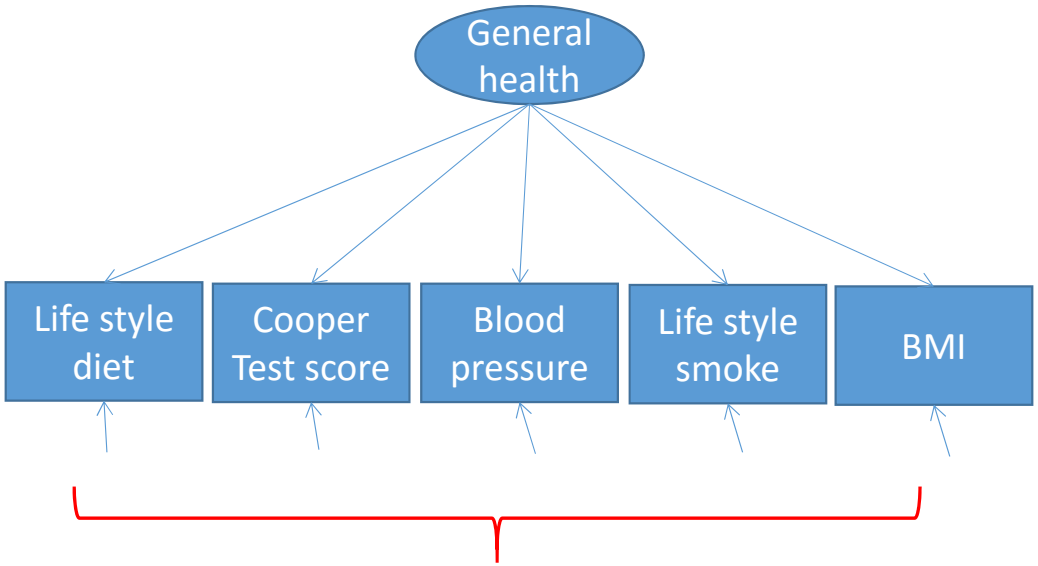
Common pathway vs Independent pathway model.

A different interpretation: factor analysis as a data summary.
Just a way to reduce multiple phenotypes into a single index.

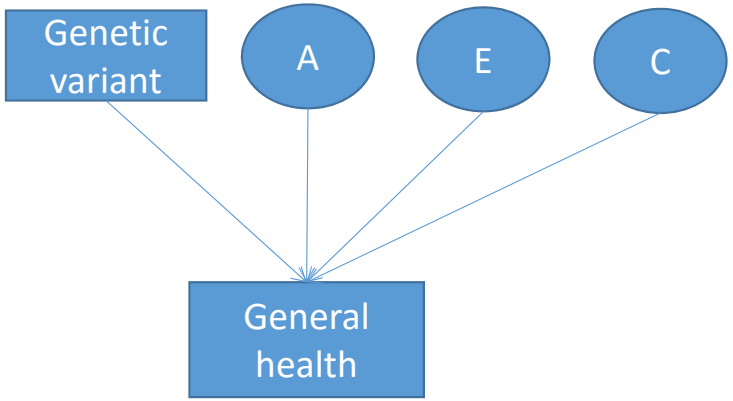



Formative indicators. No causal interpretation: General Health does not cause smoking!

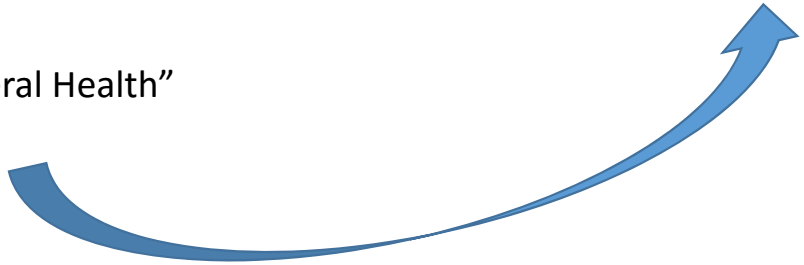
# When to use a sum score?



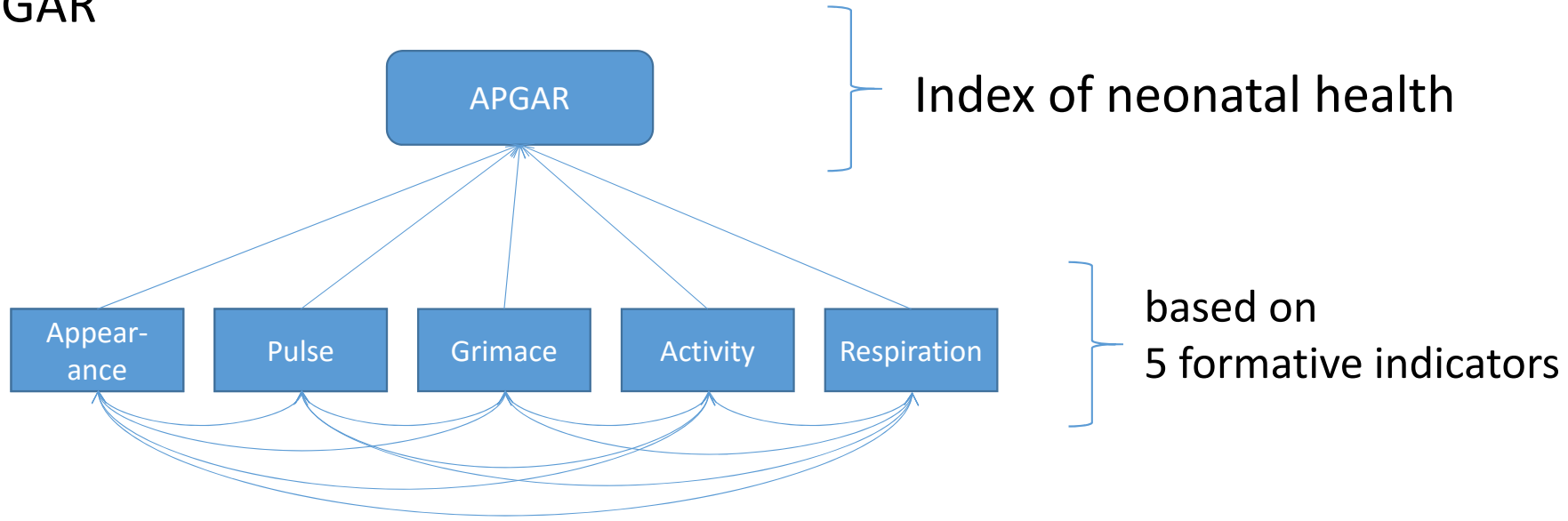

requires careful interpretation




Sum these and analyze the phenotype “General Health”



APGAR




Items are **formative**: itemscores form the APGAR score
Index variable = defined by formative items.
The APGAR is dependent on the formative items.
APGAR does not determine or cause the scores on the APGAR items
(e.g., APGAR does not cause poor respiration or blue appearance,
obviously: poor respiration causes blue appearance).

Back to the common factor model !

Multiple common factors, CFA vs. EFA with rotation

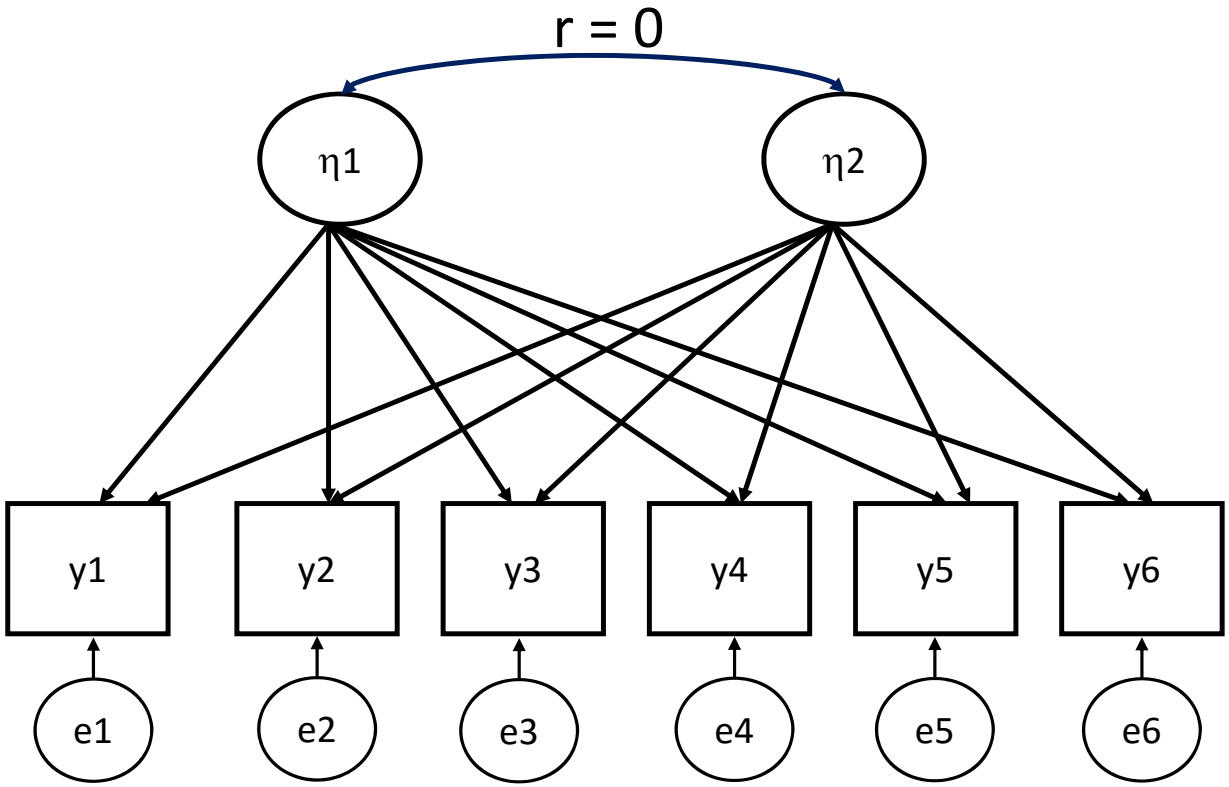
EFA (always more than one common factor).
**Aim**: determine dimensionality and derive meaning of factors from factor loadings

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:
all indicators load on all common factors....



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

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

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

$$\eta^t = [\eta_1 \ \eta_2]$$

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

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

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

$$\Psi = \mathbf{I} = \begin{matrix} 1 & 0 \\ 0 & 1 \end{matrix}$$

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

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

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

unique values of $\Lambda$, but rotatable or transformable

$\Lambda\mathbf{M} = \Lambda^*$, $\mathbf{M}\mathbf{M}^t = \mathbf{I}$, so that $\Sigma_y = \Lambda\,\mathbf{M}\mathbf{M}^t\Lambda^t + \Theta = \Lambda\mathbf{I}\Lambda^t + \Theta$

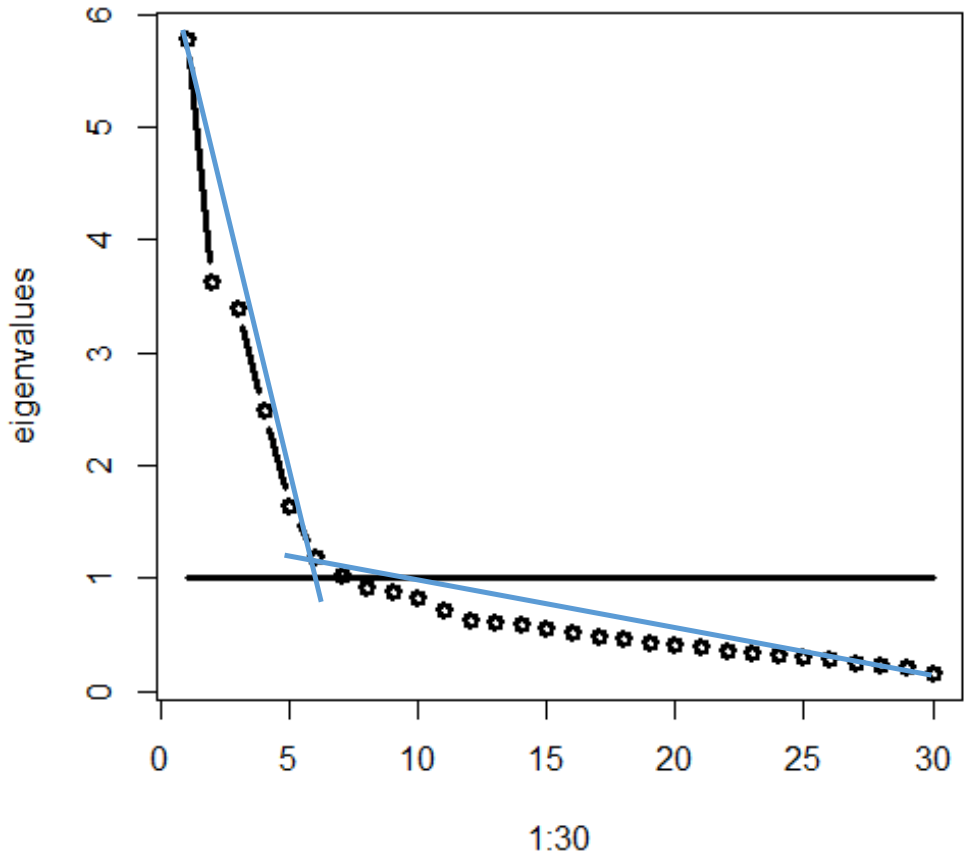
M is called a rotation matrix … to obtain interpretable $\Lambda^*$

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

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

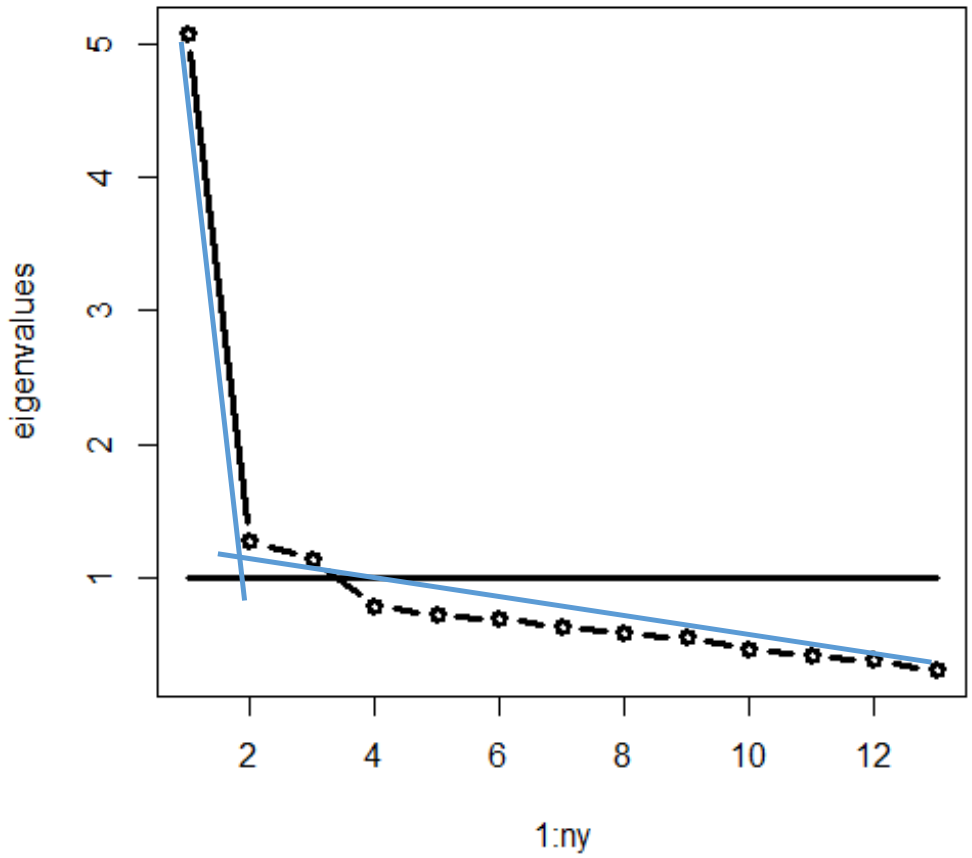
Orthogonal rotation leaves common factors uncorrelated
Oblique rotation allows for correlation.

Rotation is just a transformation of results (no testing!).




BIG 5 data 361 females students
30 indicators: expected 5 factors

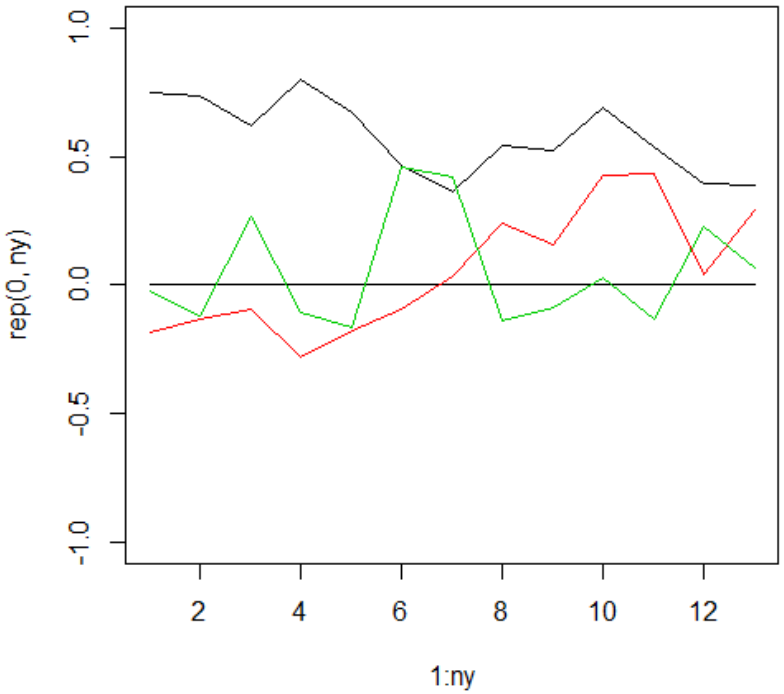
Screeplot locate the “elbow joint” (5)
Eigenvalues > 1 rule (6?)



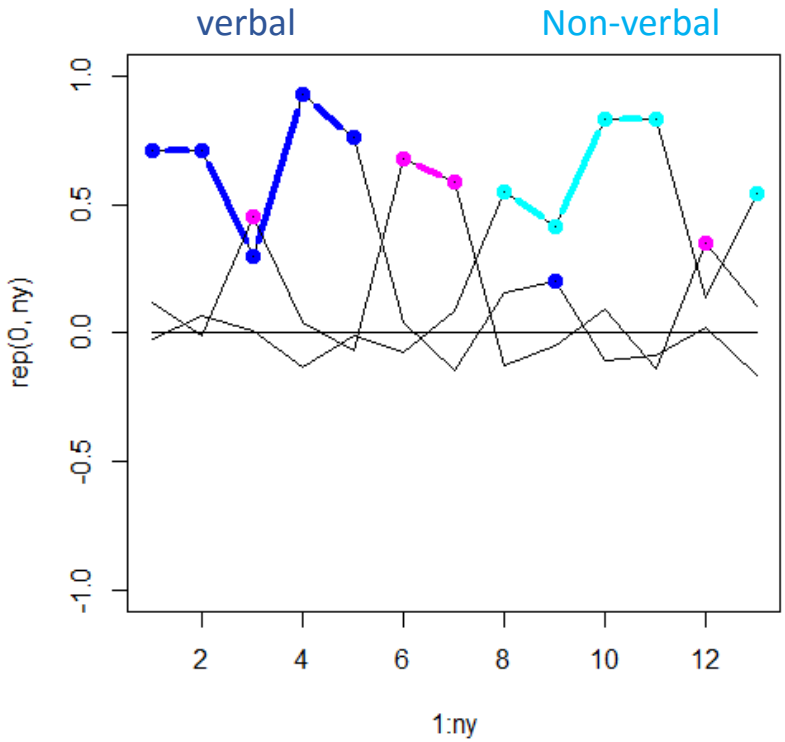

WAIS-III 1868 US adults: 13 subtests expected 3 factors.

Screeplot locate the “elbow joint” (1)
Eigenvalues > 1 rule (3)

3 EFA factor model: Chi2(42) = 111.9

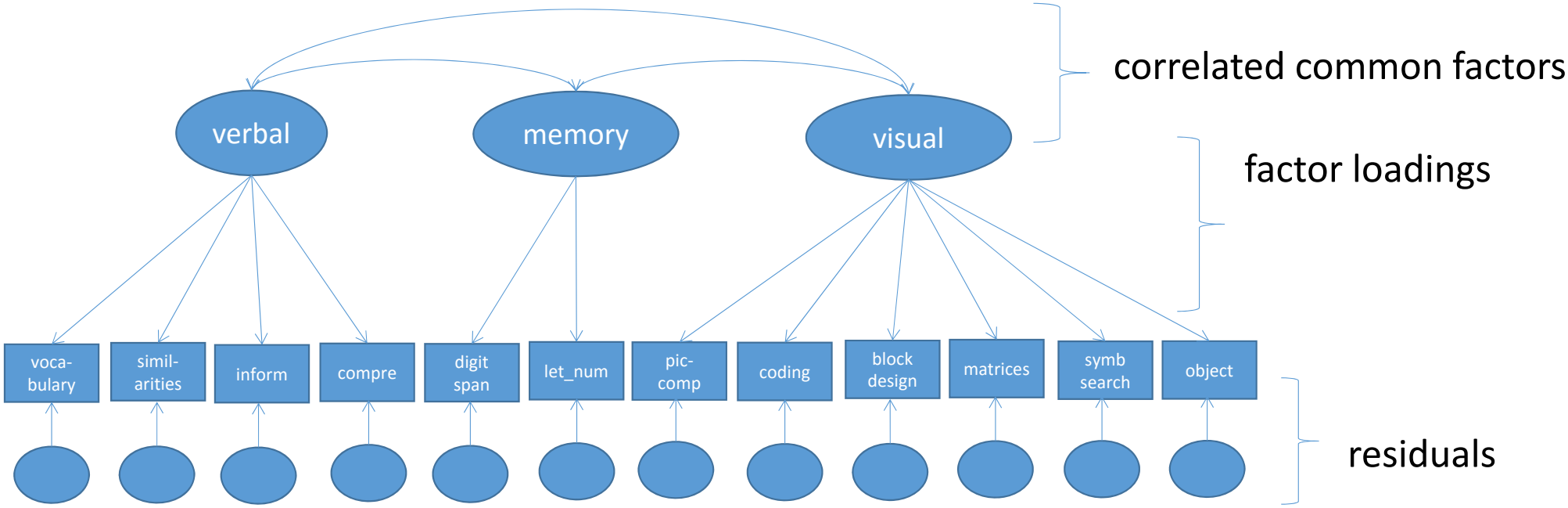



Plot of unrotated factor loadings
Chi2(42) = 111.9

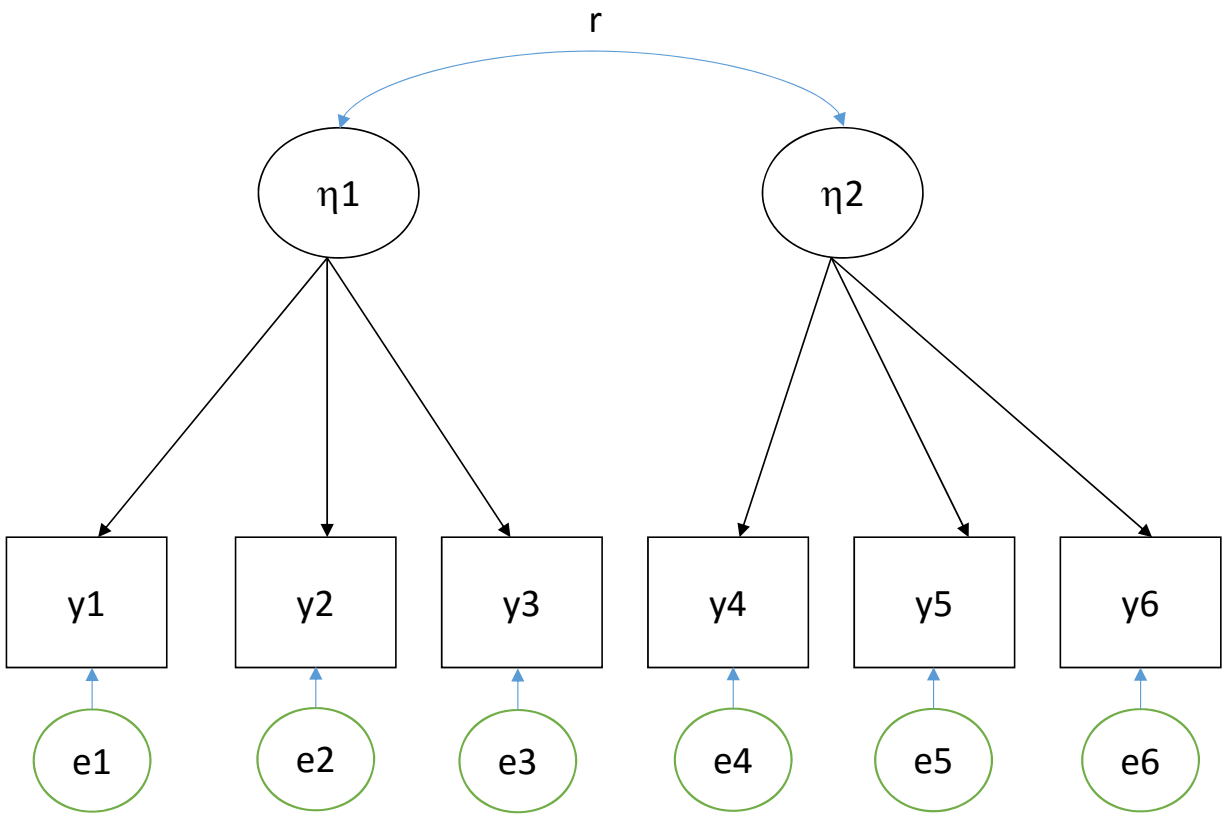



Promax rotated loadings (oblique)
Chi2(42) = 111.9

# Expected model....but if we expect this, why use EFA?



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



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

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

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

$$\eta^t = [\eta_1 \;\; \eta_2]$$

$$
\Lambda = \begin{matrix} \lambda_{11} & 0 \\ \lambda_{21} & 0 \\ \cdots & \cdots \\ 0 & \lambda_{52} \\ 0 & \lambda_{62} \end{matrix}
$$

$$\Sigma_y = \Lambda \Psi \Lambda^t + \Theta$$

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

$$
\Psi = \begin{matrix} 1 & \\ \rho & 1 \end{matrix}
\qquad
\Theta = \mathrm{diag}(\sigma^2_{\varepsilon 1} \;\; \sigma^2_{\varepsilon 2} \;\; \sigma^2_{\varepsilon 3} \;\; \sigma^2_{\varepsilon 4} \;\; \sigma^2_{\varepsilon 5} \;\; \sigma^2_{\varepsilon 6})
$$

$$\underset{(ny \times ny)}{\Sigma_y} = \underset{(ny \times ne)}{\Lambda}\ \underset{(ne \times ne)}{\Psi}\ \underset{(ne \times ny)}{\Lambda^t} + \underset{(ny \times ny)}{\Theta}$$

factor correlation matrix ($\Psi$)

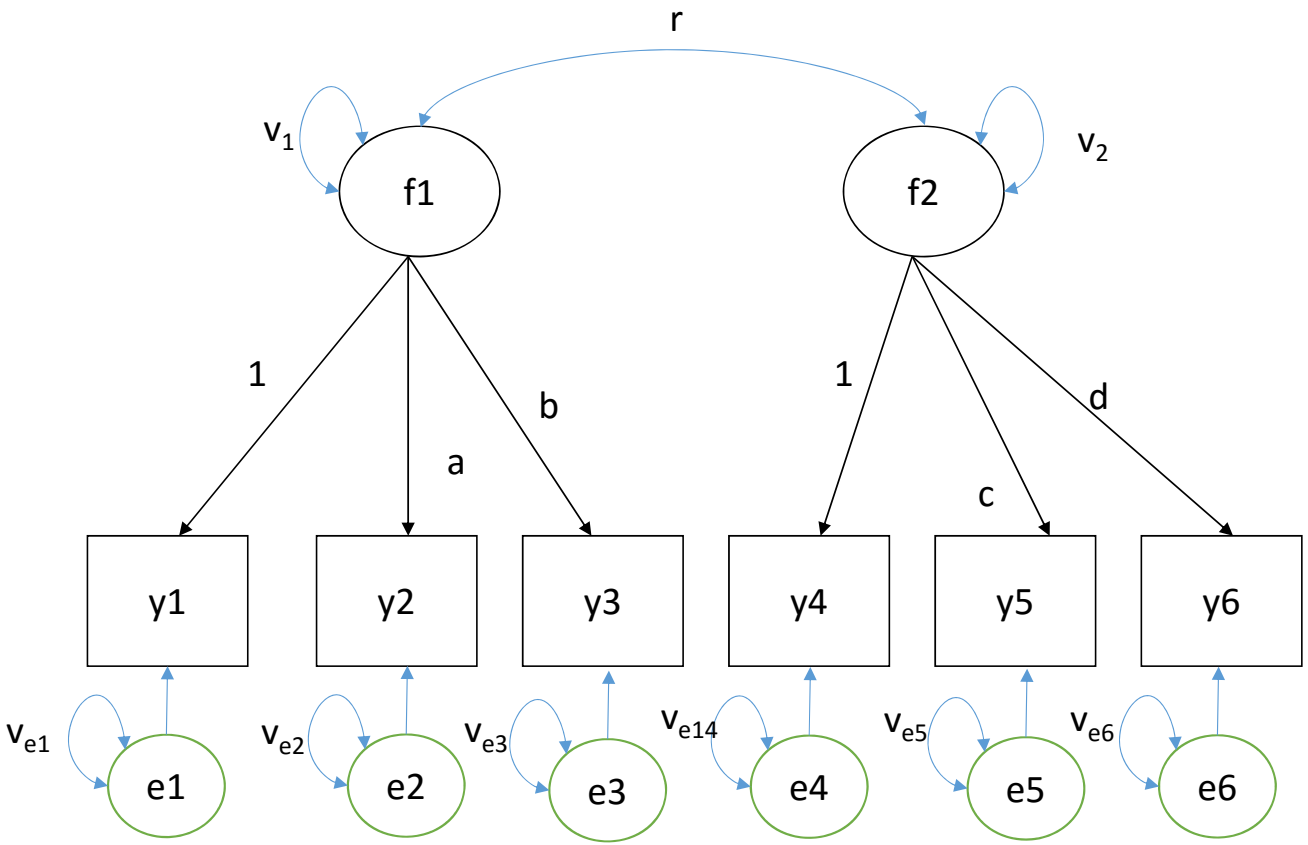
factor loading matrix ($\Lambda$)

residual covariance matrix ($\Theta$)

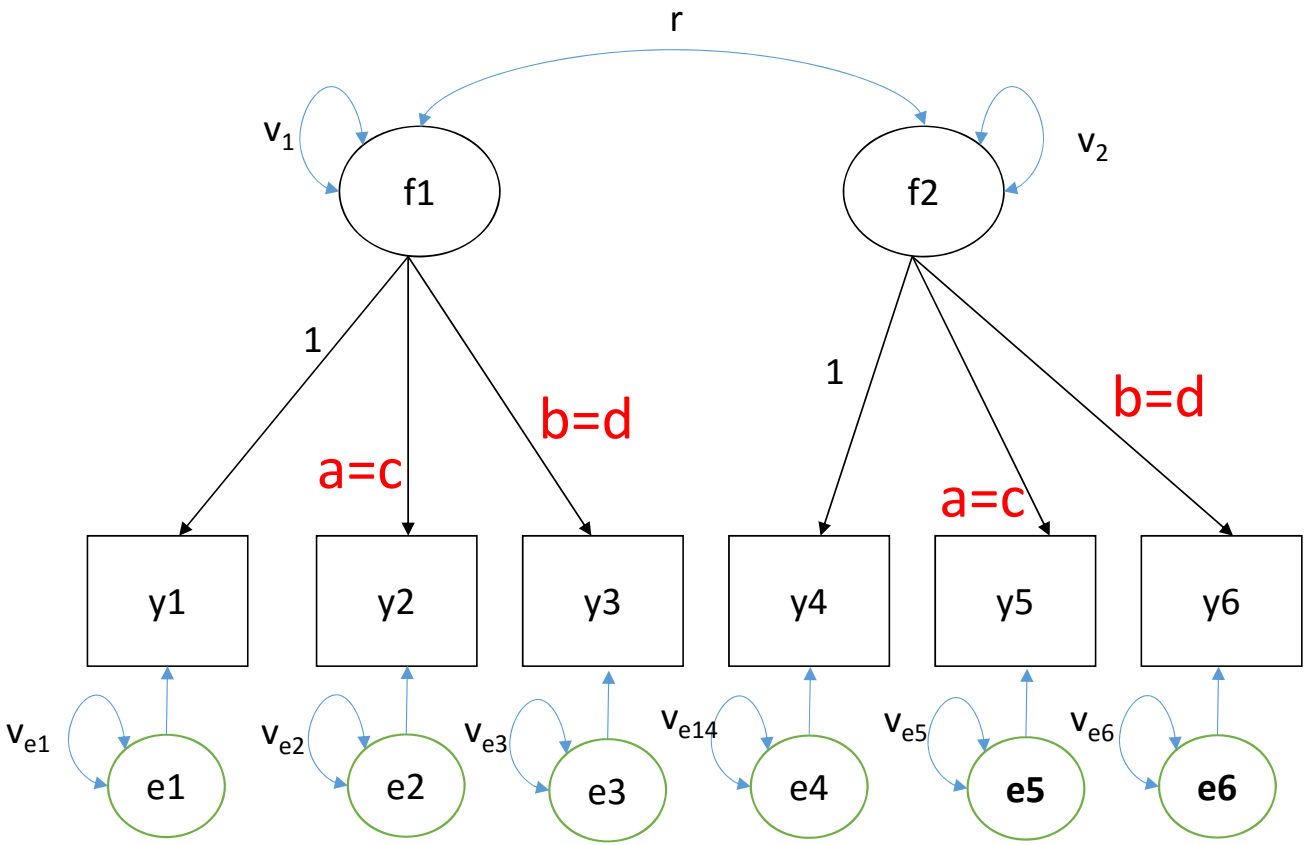
In CFA, in contrast to EFA, you can impose all kinds of constraints on the parameters

In CFA, in constrast to EFA, you can estimate off-diagonal elements in the cov matrix of the residuals $\Theta$
(e.g. to accommodate violations of local independence)

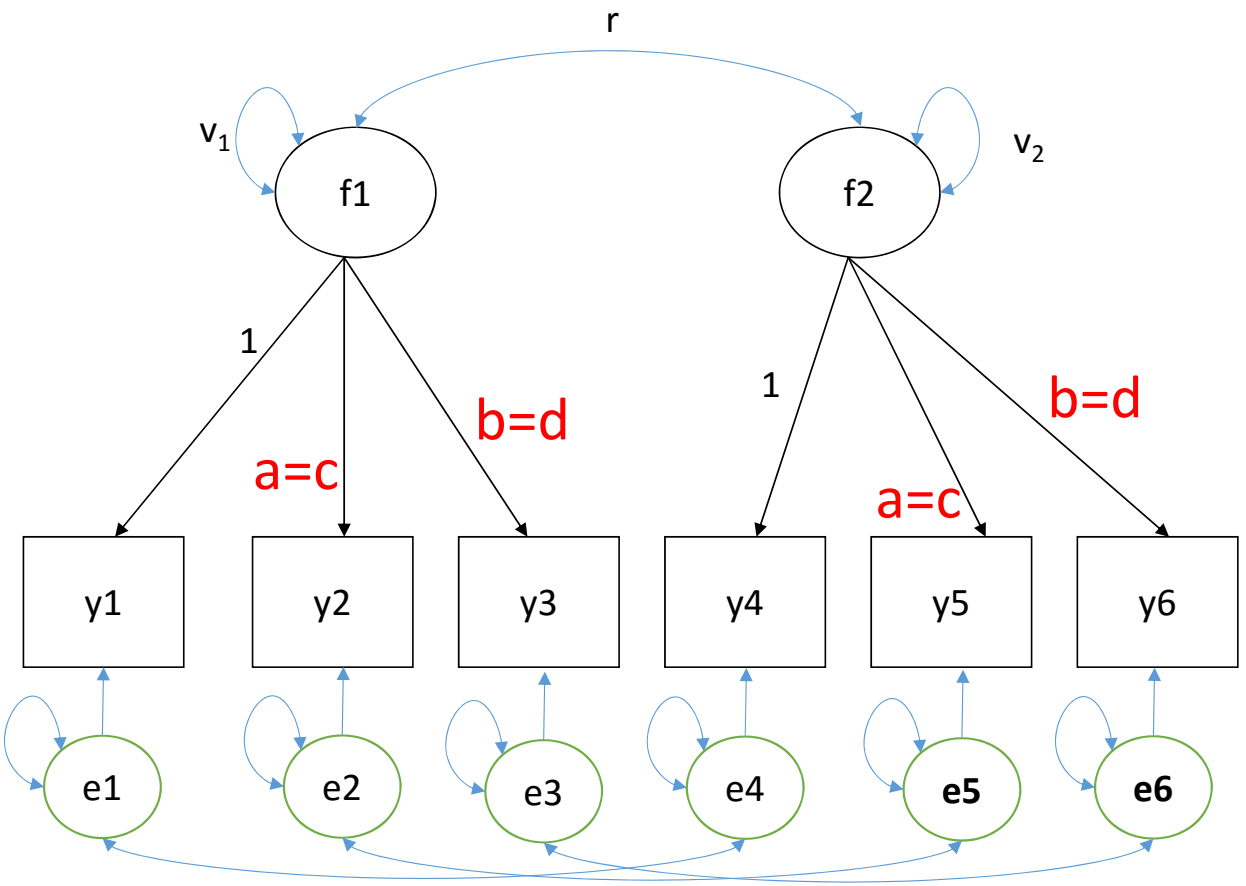
# Suppose 3 indicators at 2 time points



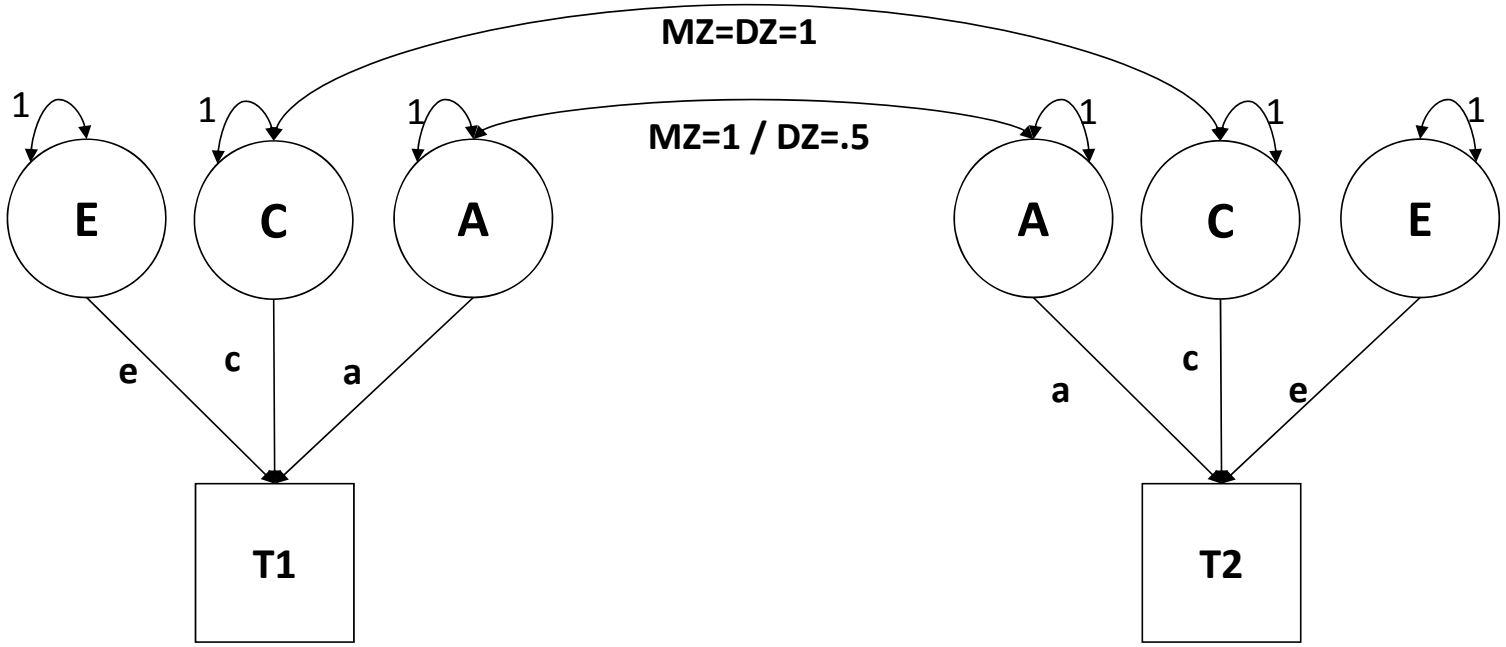
# Suppose 3 indicators at 2 time points



# Suppose 3 indicators at 2 time points




$v_{e1}=v_{e4}$ $v_{e2}=v_{e5}$ $v_{e3}=v_{e6}$ $v_{e1}=v_{e4}$ $v_{e2}=v_{e5}$ $v_{e3}=v_{e6}$




Equality constraints are nothing new!

What if I want to carry out a phenotypic factor analysis given twin data?
N pairs, but N*2 individual...

1) Ignore family relatedness treat N twin pairs as 2*N individuals ? OK does not effect estimate of the covariance matrix, but renders statistical tests invalid

(eigenvalues and scree plots are ok)

2) Ignore family relatedness treat N twin pairs as 2*N individuals use a correction for family clustering? OK convenient, the correction is built in in the Mplus program, can be programmed in OpenMx

3) Do the factor analysis in N twins and replicate the model in the other N twins? Ok, but not true replication (call it pseudo replication)

4) Do the factor analysis in twins separately and simultaneously, but include the twin 1 – twin 2 phenotypic covariances. Ok, but possibly unwieldy (especially is you have extended pedigrees).

**Summary I**: Exploratory factor analysis

a) how many common factors “underlie” the test scores....?

b) can we interpret the common factors by inspecting the rotated factor loadings...to understand the meaning of the common factors?

c) does it help interpretation to allow for correlated factors (oblique rotation)?

d) evaluate the reliability of the indicators.

**Summary II**: Confirmatory factor analysis.

a) test preconceived model in which the number of common factors and the pattern of factor loadings are given

b) evaluate the overall goodness of fit

c) evaluate the significance of the factor loadings, the values of the common factor correlations

d) evaluate measurement invariance w.r.t. a given external variable X: is the relationship between X and the indicators mediated by the common factor(s)?

**Summary III**: relevance of factor analysis to twin studies

a) It may be of interest to determine the dimensionality of items if the items are use to create a sum score (sum of items or symptom endorsement)

b) The decomposition of $\mathbf{S}_{ph} = \mathbf{S}_A + \mathbf{S}_C + \mathbf{S}_E$ is based on a cholesky decomposition. Each covariance matrix SA, SC and SE may be subjected to factor analysis.

**Summary IV**: Relevance of twin studies to phenotypic factor analysis:

1) Common pathway model vs Independent pathway model.

Phenotypic Factor Analysis practical

1) EFA using the R routine **factanal (a routine in R, not part of OpenMx)**:

    13 WAIS subtests, unrotated and rotated 3 factor model

2) CFA using OpenMx.

    Big 5 Neuroticism and Extroversion, 12 subtest: correlated 2 factor model

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:

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

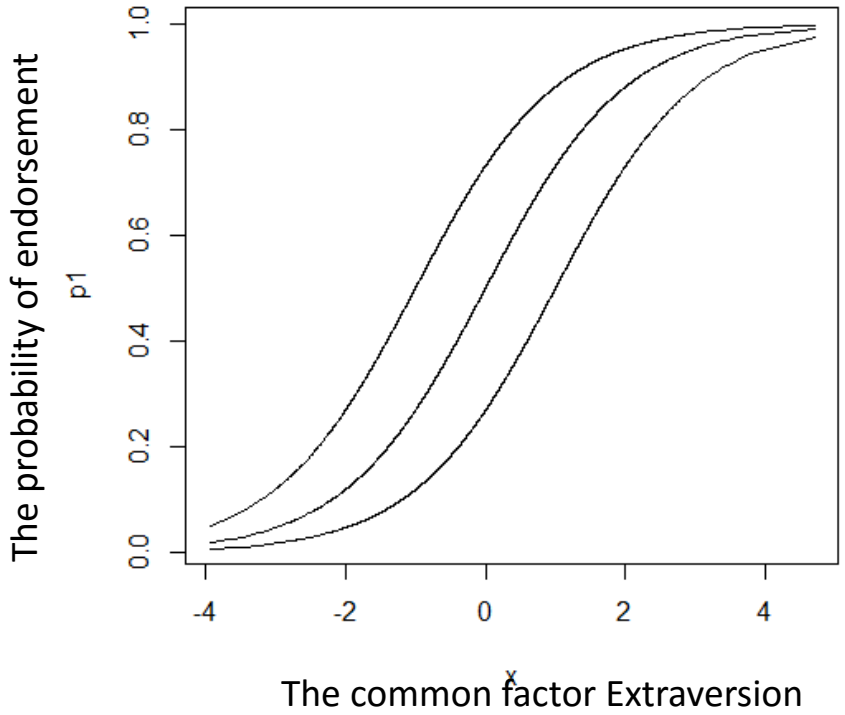
Probit:

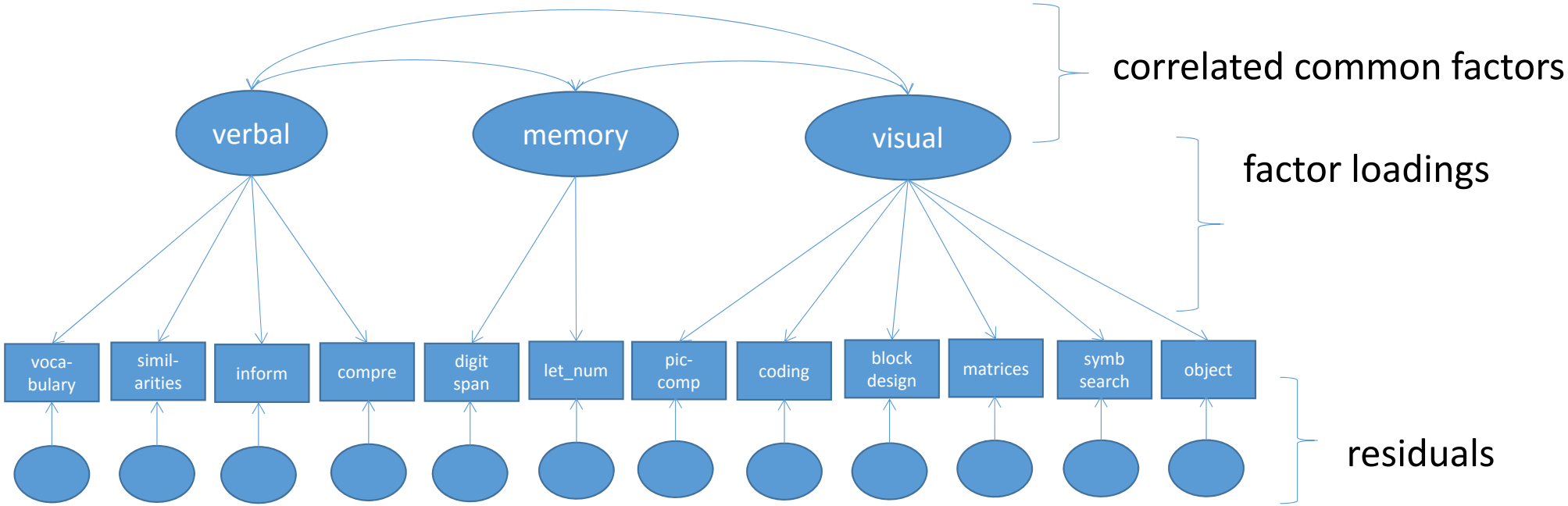
$E[Z|X=x^\circ] = Prob(Z=0|X=x^\circ) = \Phi(b0 + b1*x^\circ)$, $\Phi(.)$ cumulative st. normal distribution

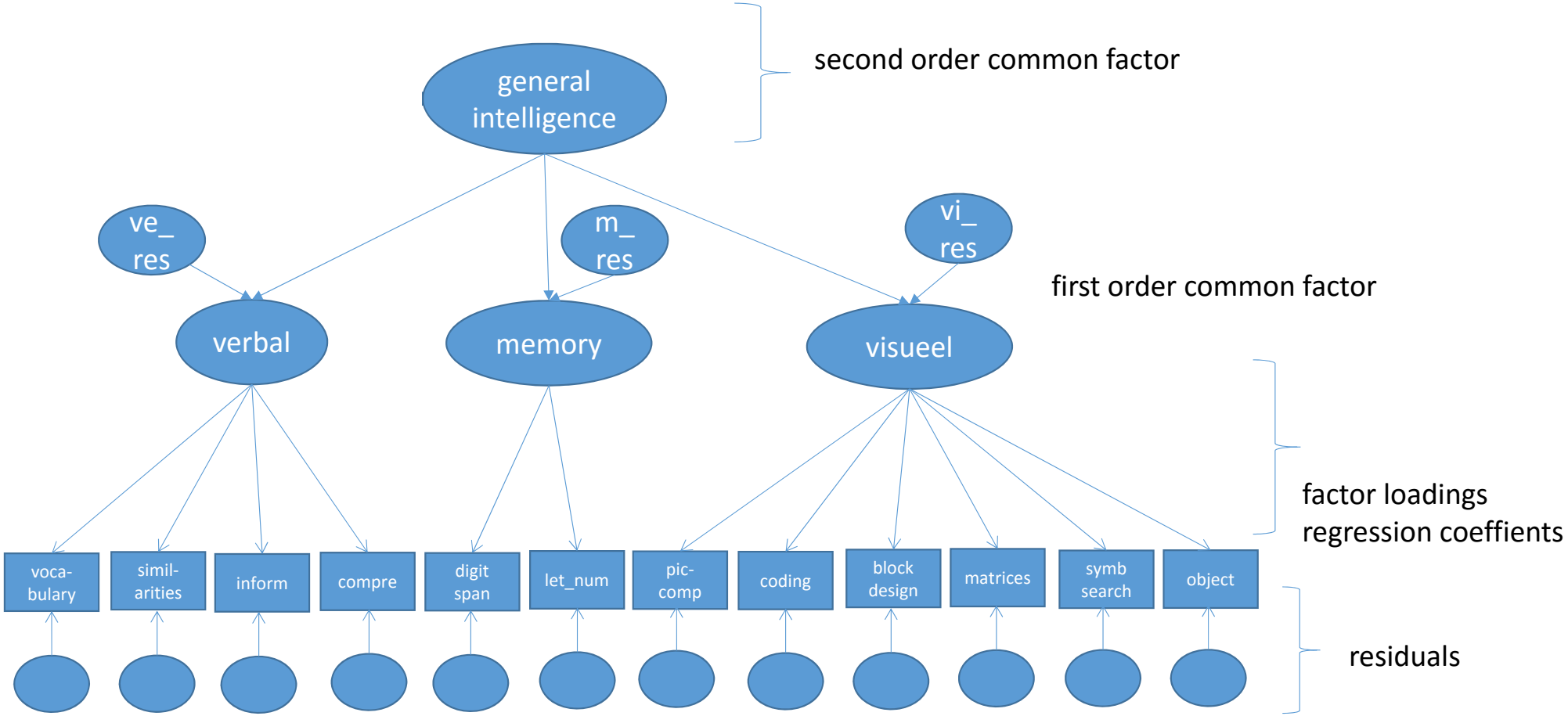
Replace X, observed predictor, but $\eta$, the common factor.

Prob(yes|X=x˚)

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




verbal
memory
visual
voca-bulary
simil-arities
inform
compre
digit span
let_num
pic-comp
coding
block design
matrices
symb search
object
correlated common factors
factor loadings
residuals


general intelligence
second order common factor
ve_ res
m_ res
vi_ res
first order common factor
verbal
memory
visueel
factor loadings
regression coeffients
voca- bulary
simil- arities
inform
compre
digit span
let_num
pic- comp
coding
block design
matrices
symb search
object
residuals