SEM 1: Confirmatory Factor Analysis

Lecture 1-2: Measurement and basics of CFA

Sacha Epskamp

SEM1 - 2020

Psychometrics

Psychometrics

The field of science concerned with measurement of psychological constructs.

Psychometrics

The field of science concerned with measurement of psychological constructs.

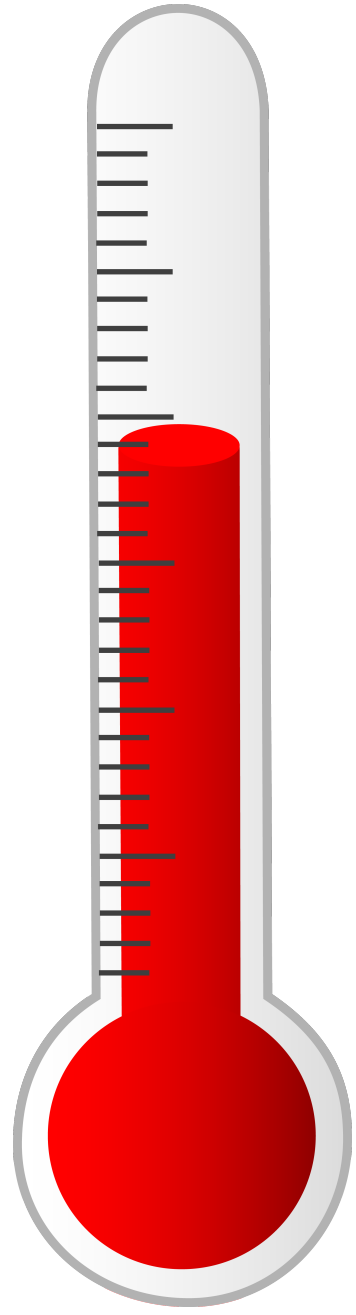
For example:

- Personality traits
- Cognitive skills
- Ability
- Psychopathological disorders
- Types

How do you measure temperature?

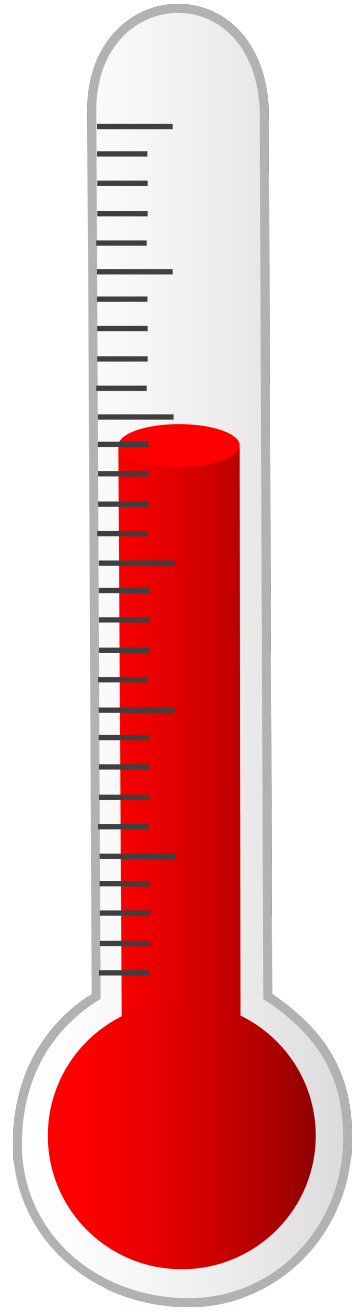
How do you measure temperature?

- By looking at a thermometer



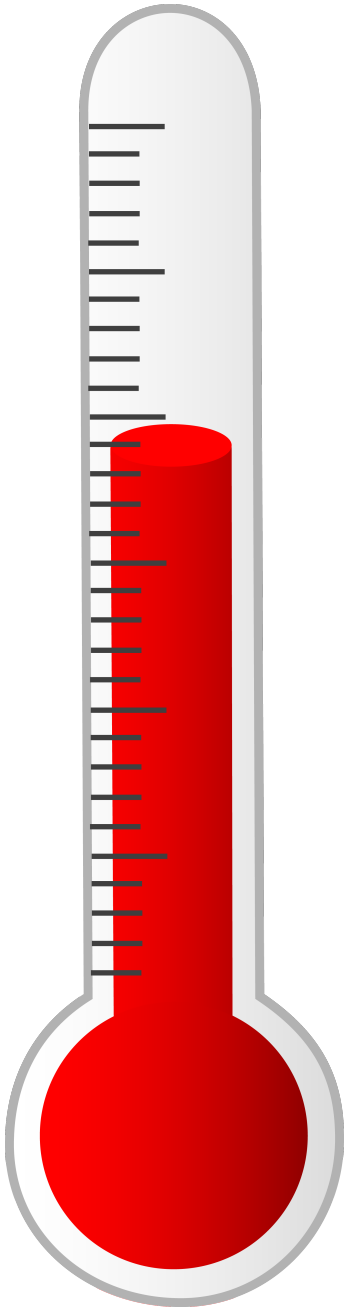
How do you measure temperature?

- By looking at a thermometer
- For this to make sense, we need to assume that:



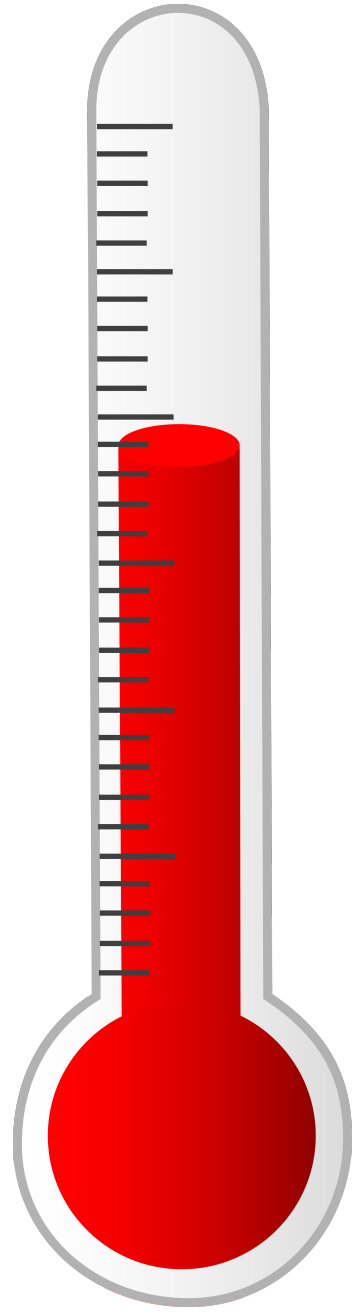
How do you measure temperature?

- By looking at a thermometer
- For this to make sense, we need to assume that:
  - Temperature causes the level shown in a thermometer

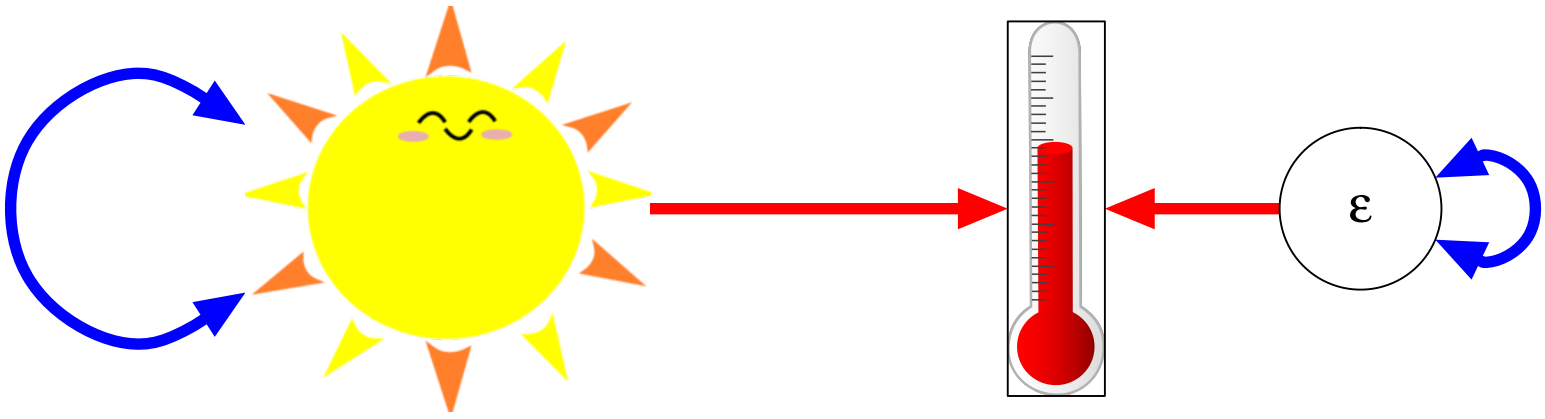


How do you measure temperature?

- By looking at a thermometer
- For this to make sense, we need to assume that:
  - Temperature causes the level shown in a thermometer
  - The thermometer features relatively little measurement error

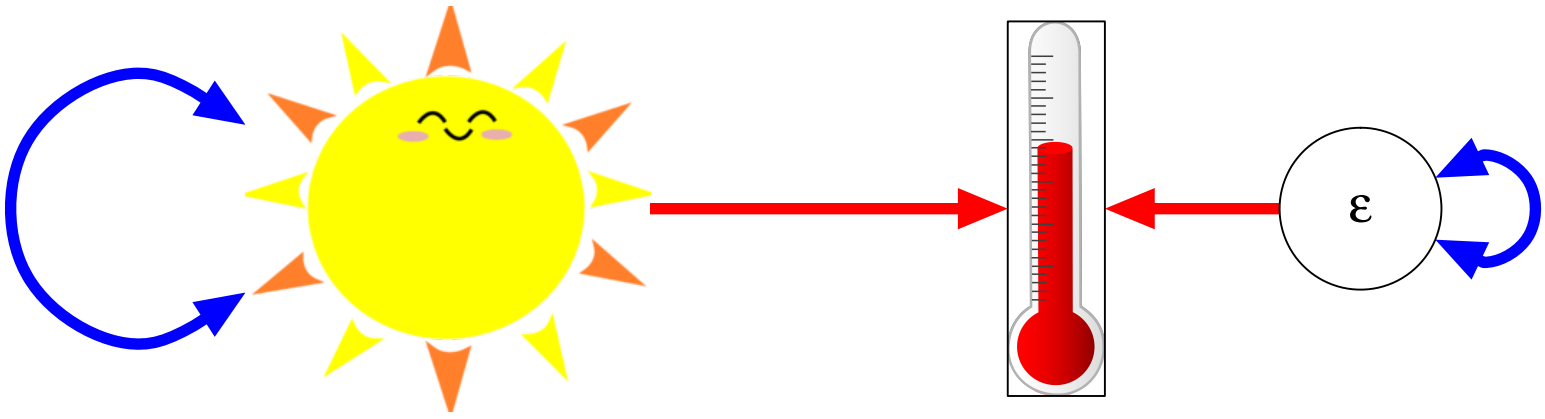


We can summarize a causal hypothesis in a path diagram:



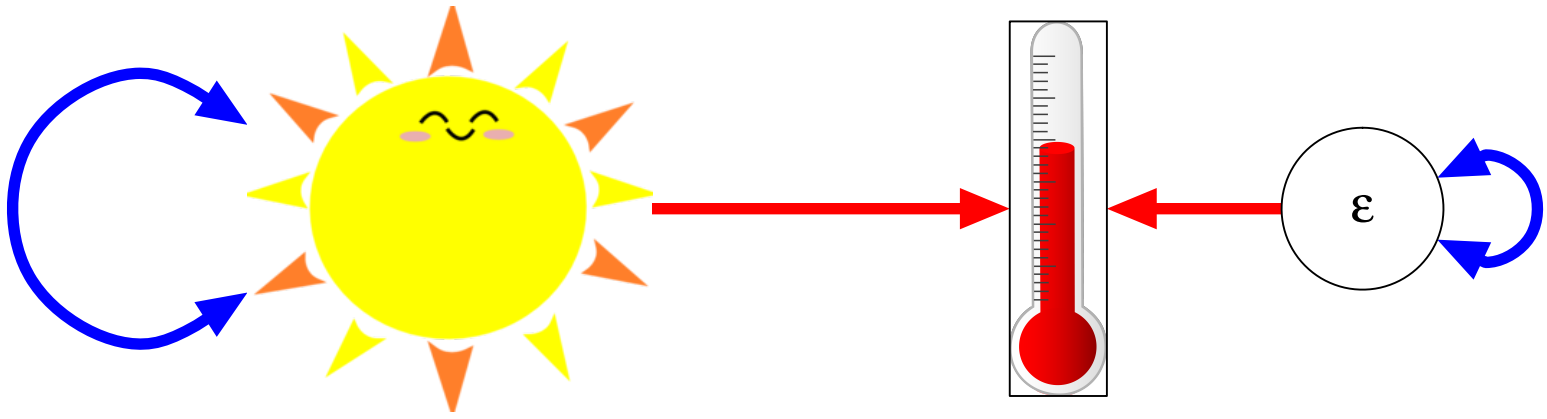
- Circular nodes (or suns): latent (unobserved) variables

We can summarize a causal hypothesis in a path diagram:



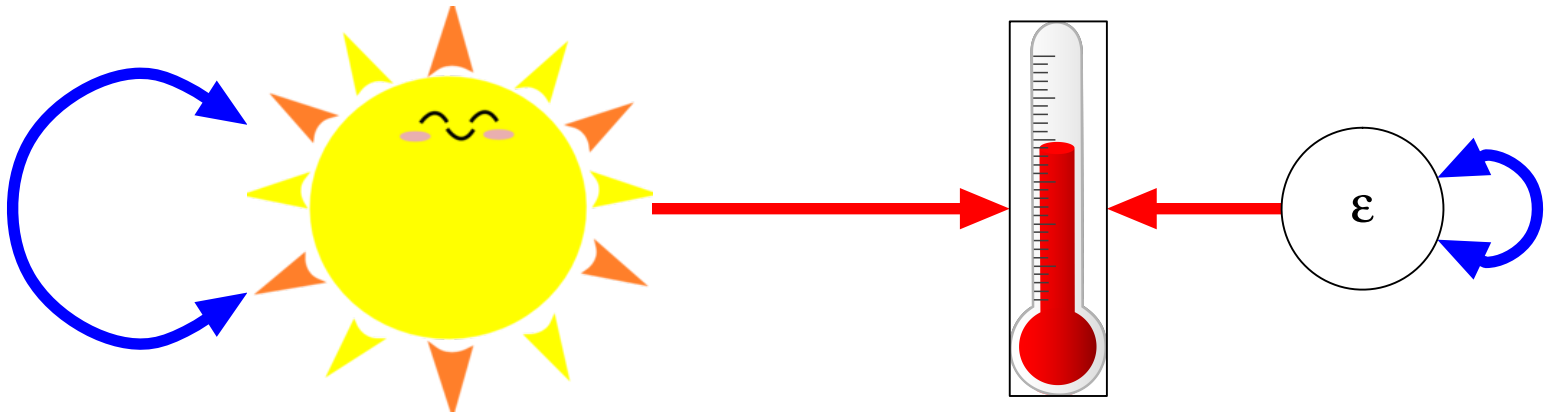
- Circular nodes (or suns): latent (unobserved) variables
- Square nodes: observed variables

We can summarize a causal hypothesis in a path diagram:



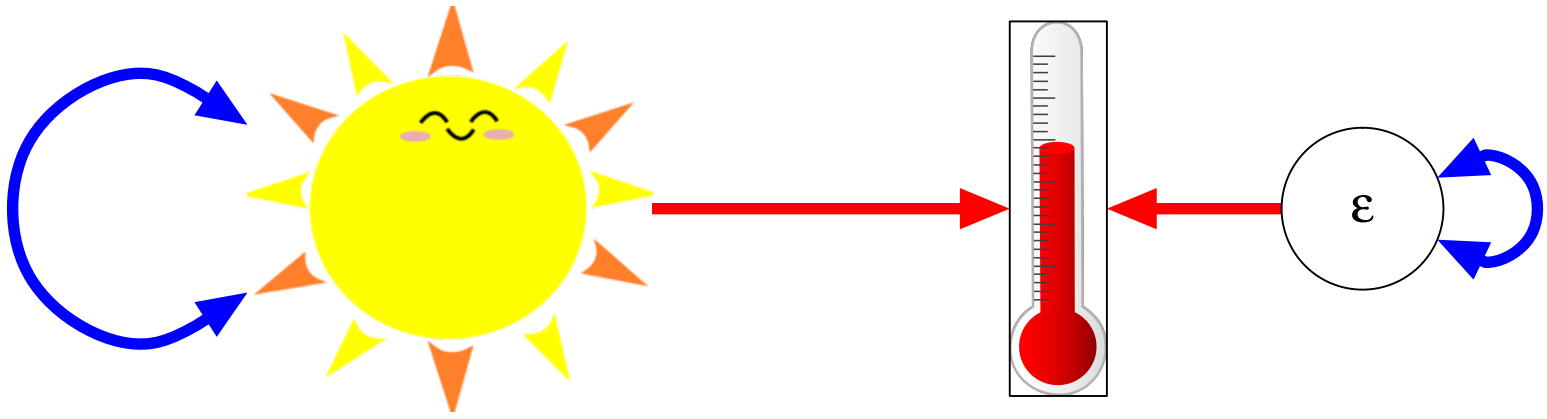
- Circular nodes (or suns): latent (unobserved) variables
- Square nodes: observed variables
  - Also termed *indicators* of a latent variable

We can summarize a causal hypothesis in a path diagram:



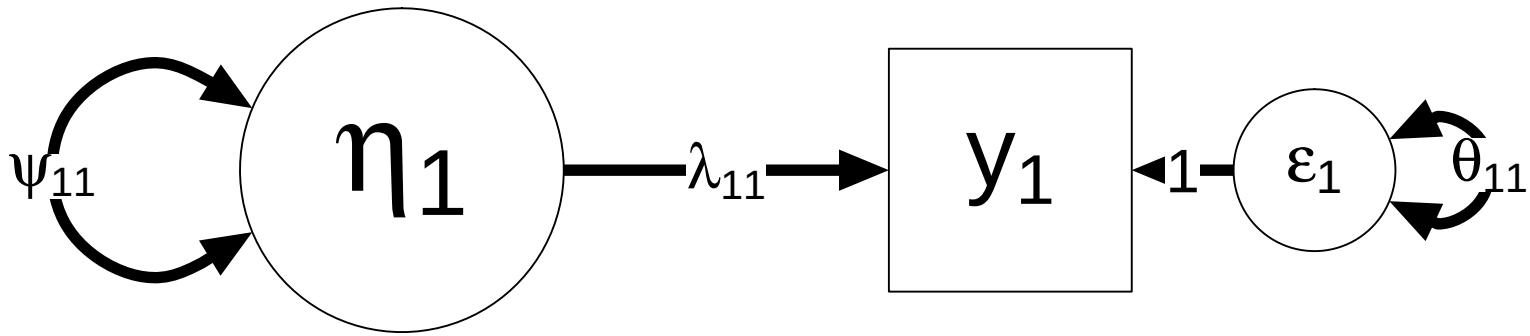
- Circular nodes (or suns): latent (unobserved) variables
- Square nodes: observed variables
  - Also termed *indicators* of a latent variable
- Unidirectional links: causal effects

We can summarize a causal hypothesis in a path diagram:



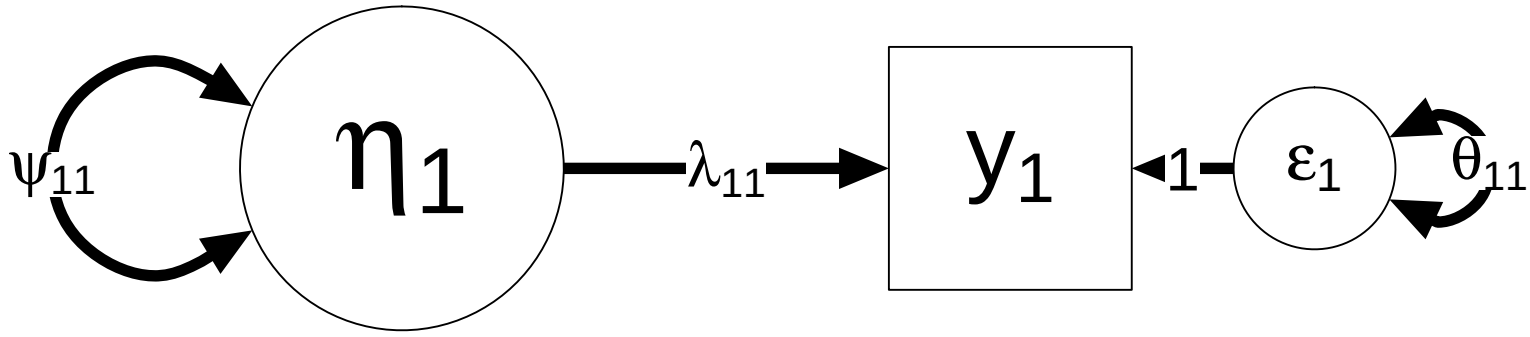
- Circular nodes (or suns): latent (unobserved) variables
- Square nodes: observed variables
  - Also termed *indicators* of a latent variable
- Unidirectional links: causal effects
- Bidirectional links: (co)variances

Assume all observed and latent variables are normally distributed and all causal effects are linear. Without loss of information, we can center data and assume all means are 0 (we don't use means until week 3). Now, the path diagram encodes a causal equation:



$$y_{i1} = \lambda_{11}\eta_{i1} + \varepsilon_{i1}$$
$$\eta_1 \sim N(0, \sqrt{\psi_{11}})$$
$$\varepsilon_1 \sim N(0, \sqrt{\theta_{11}})$$

λ_{11} is called a factor loading, ε_{i1} the residual variance and ψ_{11} the factor variance.



$$y_{i1} = \lambda_{11}\eta_{i1} + \varepsilon_{i1}$$

$$\eta_1 \sim N(0, \sqrt{\psi_{11}})$$

$$\varepsilon_1 \sim N(0, \sqrt{\theta_{11}})$$

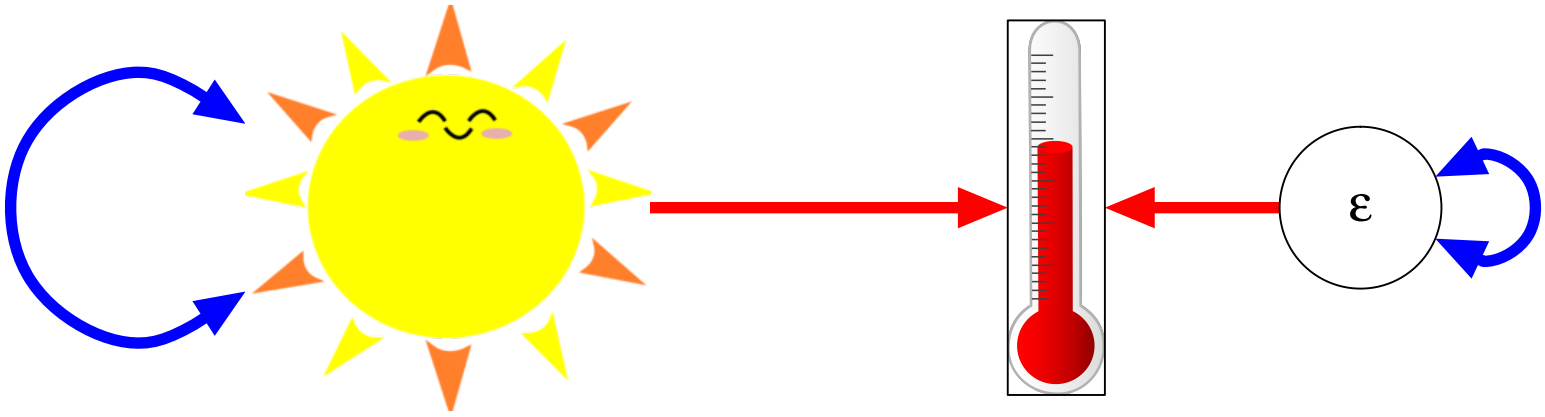
Variance of y_1:

$$\mathrm{Var}(y_1) = \lambda_{11}^2\psi_{11} + \theta_{11}$$

How much variance does the latent variable explain?

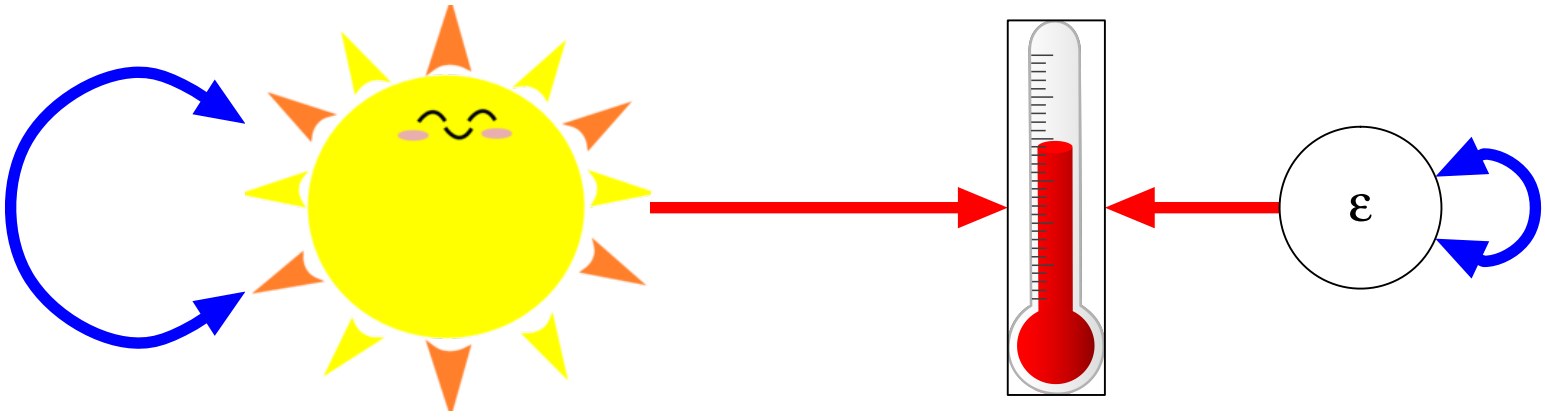
$$\frac{\lambda_{11}^2\psi_{11}}{\lambda_{11}^2\psi_{11} + \theta_{11}}$$

Identification 1: scaling



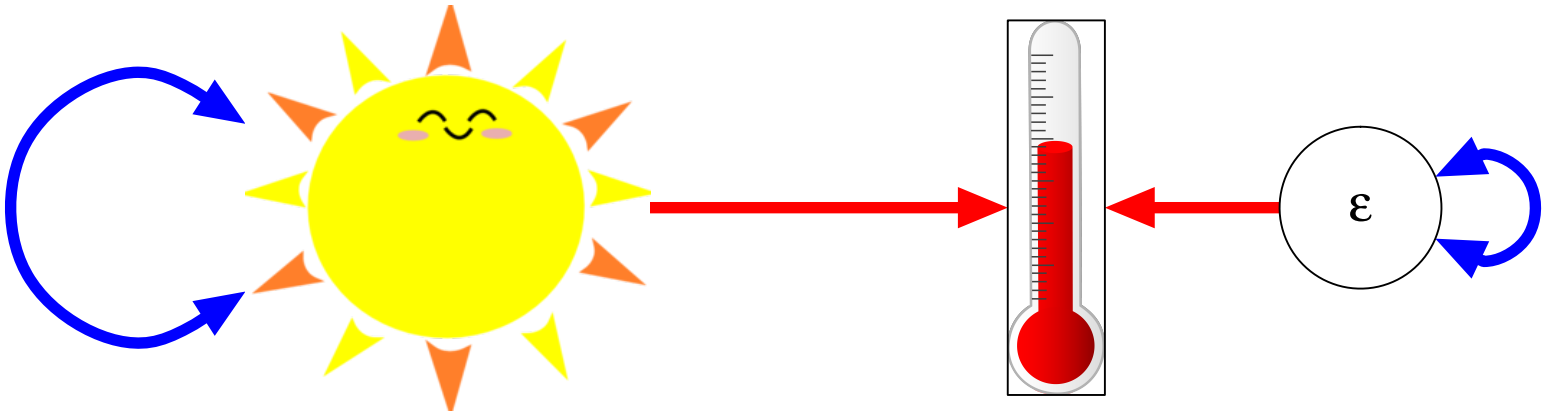
- The *unit* of our thermometer is known (e.g., Celsius).

Identification 1: scaling



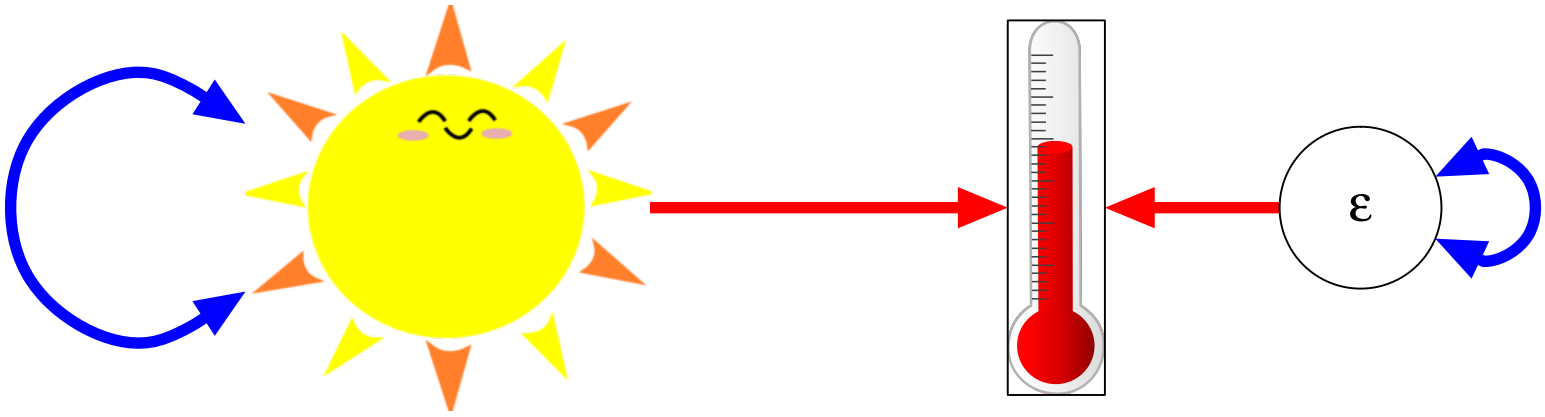
- The *unit* of our thermometer is known (e.g., Celsius).
- But the unit of the true temperature is unknown! It needs to be defined.

Identification 1: scaling



- The *unit* of our thermometer is known (e.g., Celsius).
- But the unit of the true temperature is unknown! It needs to be defined.
  - Our celsius thermometer could also measure Fahrenheit with a simple transformation!

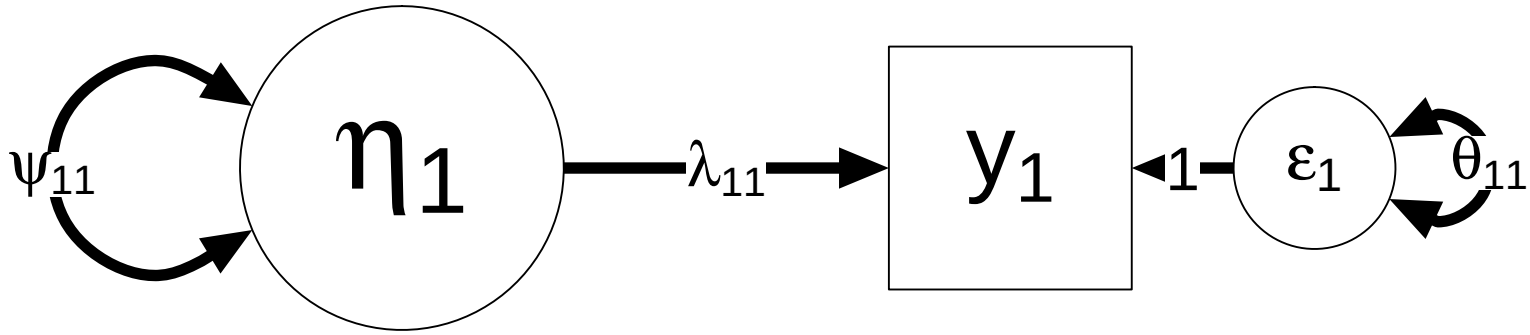
Identification 1: scaling



- The *unit* of our thermometer is known (e.g., Celsius).
- But the unit of the true temperature is unknown! It needs to be defined.
  - Our celsius thermometer could also measure Fahrenheit with a simple transformation!
- We need to *identify* the mean and variance of the latent variable!

Until May 2019 the official definition of a kilogram:

https://en.wikipedia.org/wiki/SI_base_unit



Multiplying λ_{11} by constant c and dividing ψ_{11} by c^2 leads to the *same* variance. Thus, these parameters are not identified. We need to scale the latent variable by fixing one of these parameters. Usually by setting $\lambda_{11} = 1$ or $\psi 11 = 1$ (not both!).

$$y_{i1} = \eta_{i1} + \varepsilon_{i1}$$

$$\eta_1 \sim N(0, \sqrt{\psi_{11}})$$

$$\varepsilon_1 \sim N(0, \sqrt{\theta_{11}})$$

$$\mathrm{Var}(y_1) = \psi_{11} + \theta_{11}$$

Problems in Estimating Latent Variables

- Well known problem in Psychometrics: you can not jointly estimate the *parameters* (parameters stable over people, such as λ_{11}) and *latent variables* (e.g., η_{i1})

Problems in Estimating Latent Variables

- Well known problem in Psychometrics: you can not jointly estimate the *parameters* (parameters stable over people, such as λ_{11}) and *latent variables* (e.g., η_{i1})
  - We need to be able to add data to get better estimates (e.g., smaller standard error), but adding an item adds parameters and adding a person adds latent variables!

Problems in Estimating Latent Variables

- Well known problem in Psychometrics: you can not jointly estimate the *parameters* (parameters stable over people, such as λ_{11}) and *latent variables* (e.g., η_{i1})
  - We need to be able to add data to get better estimates (e.g., smaller standard error), but adding an item adds parameters and adding a person adds latent variables!
- Paradoxically, the first step in solving this problem is getting rid of the latent variable η_{i1}

Problems in Estimating Latent Variables

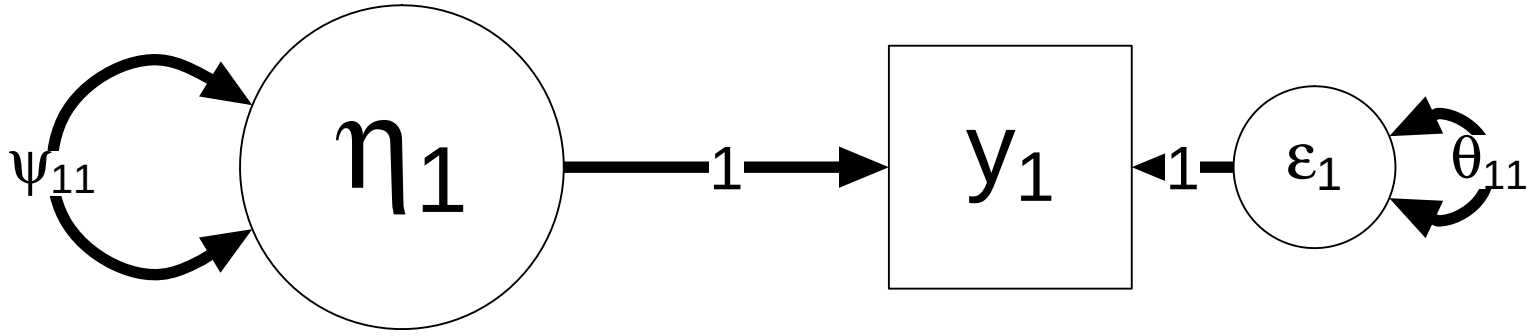
- Well known problem in Psychometrics: you can not jointly estimate the *parameters* (parameters stable over people, such as λ_{11}) and *latent variables* (e.g., η_{i1})
  - We need to be able to add data to get better estimates (e.g., smaller standard error), but adding an item adds parameters and adding a person adds latent variables!
- Paradoxically, the first step in solving this problem is getting rid of the latent variable η_{i1}
  - In item-response theory (binary items), the latent variable is integrated out

Problems in Estimating Latent Variables

- Well known problem in Psychometrics: you can not jointly estimate the *parameters* (parameters stable over people, such as λ_{11}) and *latent variables* (e.g., η_{i1})
  - We need to be able to add data to get better estimates (e.g., smaller standard error), but adding an item adds parameters and adding a person adds latent variables!
- Paradoxically, the first step in solving this problem is getting rid of the latent variable η_{i1}
  - In item-response theory (binary items), the latent variable is integrated out
  - In factor analysis, we make use of *covariance modeling*

Problems in Estimating Latent Variables

- Well known problem in Psychometrics: you can not jointly estimate the *parameters* (parameters stable over people, such as λ_{11}) and *latent variables* (e.g., η_{i1})
  - We need to be able to add data to get better estimates (e.g., smaller standard error), but adding an item adds parameters and adding a person adds latent variables!
- Paradoxically, the first step in solving this problem is getting rid of the latent variable η_{i1}
  - In item-response theory (binary items), the latent variable is integrated out
  - In factor analysis, we make use of *covariance modeling*
- After estimating parameters, the latent variable can be extrapolated. Although as we will see in SEM 2, often this is not needed and we can use covariance modeling to test all our hypotheses!



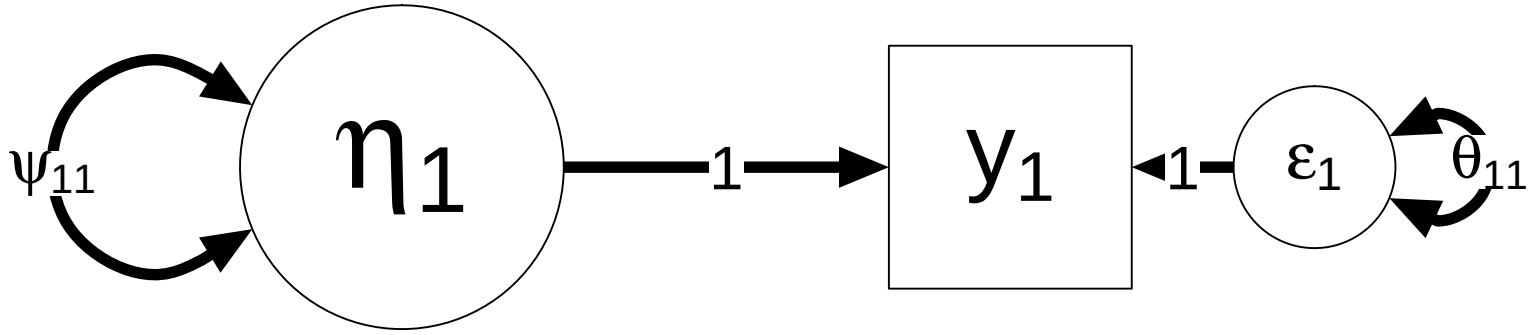
We match observed variances and covariances to parameters so that we do not need to estimate latent variables themselves. If the parameters reproduce the observed variances and covariances, we explain the data!

$$y_{i1} = \eta_{i1} + \varepsilon_{i1}$$

$$\eta_1 \sim N(0, \sqrt{\psi_{11}})$$

$$\varepsilon_1 \sim N(0, \sqrt{\theta_{11}})$$

$$\mathrm{Var}(y_1) = \psi_{11} + \theta_{11}$$



We match observed variances and covariances to parameters so that we do not need to estimate latent variables themselves. If the parameters reproduce the observed variances and covariances, we explain the data!

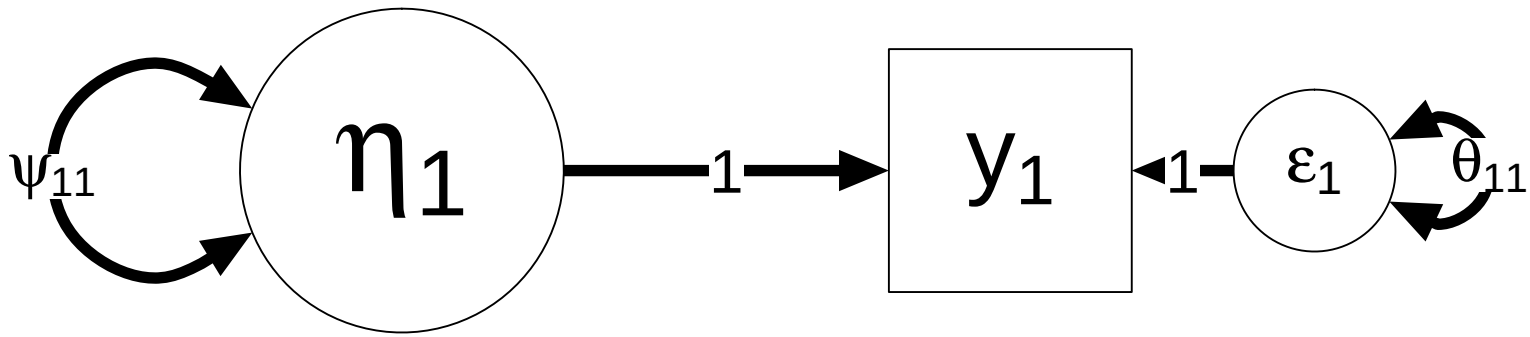
$$y_{i1} = \eta_{i1} + \varepsilon_{i1}$$
$$\eta_1 \sim N(0, \sqrt{\psi_{11}})$$
$$\varepsilon_1 \sim N(0, \sqrt{\theta_{11}})$$
$$\mathrm{Var}(y_1) = \psi_{11} + \theta_{11}$$

Problem: two parameters and only one observed variance, model is not identified!

Identification 2: Degrees of freedom

In addition to scaling, we need non-negative degrees of freedom (DF) for a model to be identified!

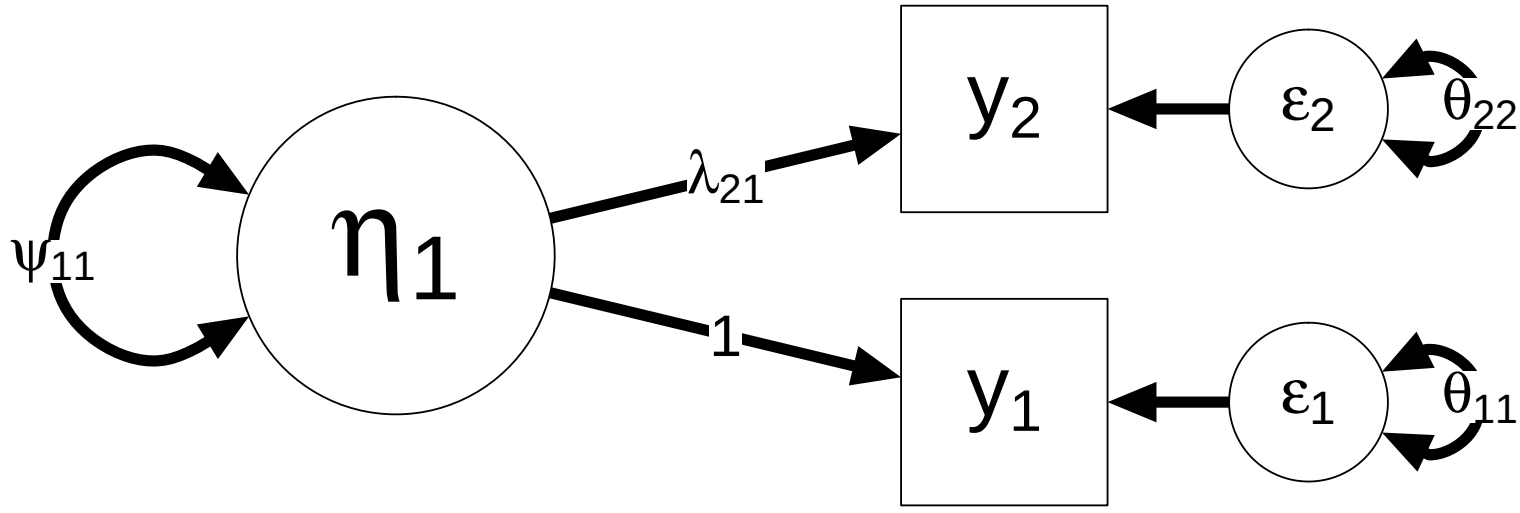
- $\mathrm{DF} = a - b$
- a: number of observations: $a = p(p+1)/2$ variances and covariances.
- b: number of parameters we need to estimate (do not count parameters we fixed for scaling)
- In general, we need 3 indicators for a single latent variable model, or 2 per factor for models with multiple (correlated) latent variables.



We need at least as many observed variances and covariances as parameters!

$$y_{i1} = \eta_{i1} + \varepsilon_{i1}$$
$$\eta_1 \sim N(0, \sqrt{\psi_{11}})$$
$$\varepsilon_1 \sim N(0, \sqrt{\theta_{11}})$$
$$\mathrm{Var}(y_1) = \psi_{11} + \theta_{11}$$

Two parameters, one observation. $\mathrm{DF} = -1$. Not identified!



$$y_{i1} = \eta_{i1} + \varepsilon_{i1}$$
$$y_{i2} = \lambda_{21}\eta_{i1} + \varepsilon_{i2}$$
$$\eta_1 \sim N(0, \sqrt{\psi_{11}})$$
$$\varepsilon_1 \sim N(0, \sqrt{\theta_{11}})$$
$$\varepsilon_2 \sim N(0, \sqrt{\theta_{22}})$$

Number of parameters: 4 (λ_{21}, ψ_{11}, θ_{11} and θ_{22})

η_{i1} is now a common cause. y_{i1} and y_{i2} are assumed independent after controlling for η_{i1}: local independence. Number of observations: 2 variances ($\mathrm{Var}(y_1)$ and $\mathrm{Var}(y_2)$) + 1 covariance ($\mathrm{Cov}(y_1, y_2)$) = 3. Model is not identified!

General factor analysis framework:

$$\begin{aligned} \boldsymbol{y}_i &= \boldsymbol{\Lambda}\boldsymbol{\eta}_i + \boldsymbol{\varepsilon}_i \\ \boldsymbol{y} &\sim N(\mathbf{0}, \boldsymbol{\Sigma}) \\ \boldsymbol{\eta} &\sim N(\mathbf{0}, \boldsymbol{\Psi}) \\ \boldsymbol{\varepsilon} &\sim N(\mathbf{0}, \boldsymbol{\Theta}), \end{aligned}$$

in which:

- $\boldsymbol{y}_i$ is a p-length vector of item responses
- $\boldsymbol{\eta}_i$ an m-length vector of latent variables
- $\boldsymbol{\varepsilon}_i$ an p-length vector of residuals
- $\boldsymbol{\Lambda}$ a $p \times m$ matrix of factor loadings
- $\boldsymbol{\Psi}$ an $m \times m$ symmetric variance–covariance matrix (assume always all latent variables are correlated)
- $\boldsymbol{\Theta}$ is a $p \times p$ symmetric variance–covariance matrix, mostly diagonal (unless you explicitly expect violations of local independence)

The general framework:

$$
\begin{aligned}
\boldsymbol{y}_i &= \boldsymbol{\Lambda}\boldsymbol{\eta}_i + \boldsymbol{\varepsilon}_i \\
\boldsymbol{y} &\sim N(\mathbf{0}, \boldsymbol{\Sigma}) \\
\boldsymbol{\eta} &\sim N(\mathbf{0}, \boldsymbol{\Psi}) \\
\boldsymbol{\varepsilon} &\sim N(\mathbf{0}, \boldsymbol{\Theta}),
\end{aligned}
$$

Allows you to derive the model-implied variance–covariance matrix:

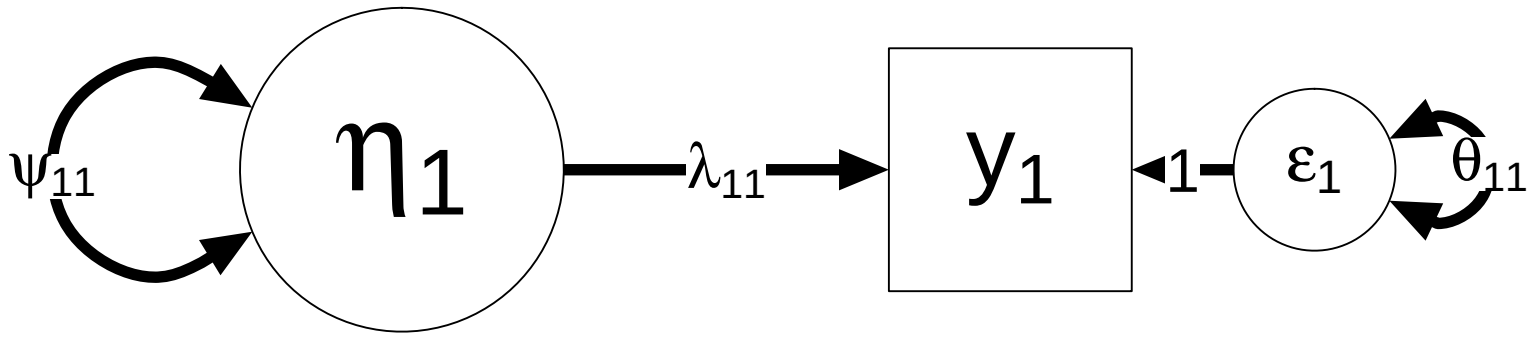
$$
\boldsymbol{\Sigma} = \boldsymbol{\Lambda}\boldsymbol{\Psi}\boldsymbol{\Lambda}^{\top} + \boldsymbol{\Theta}
$$

Covariance modeling

In general, we aim to estimate parameters that lead to a model implied variance–covariance matrix $\boldsymbol{\Sigma}$ by minimizing (note that the Brown book makes a mistake here):

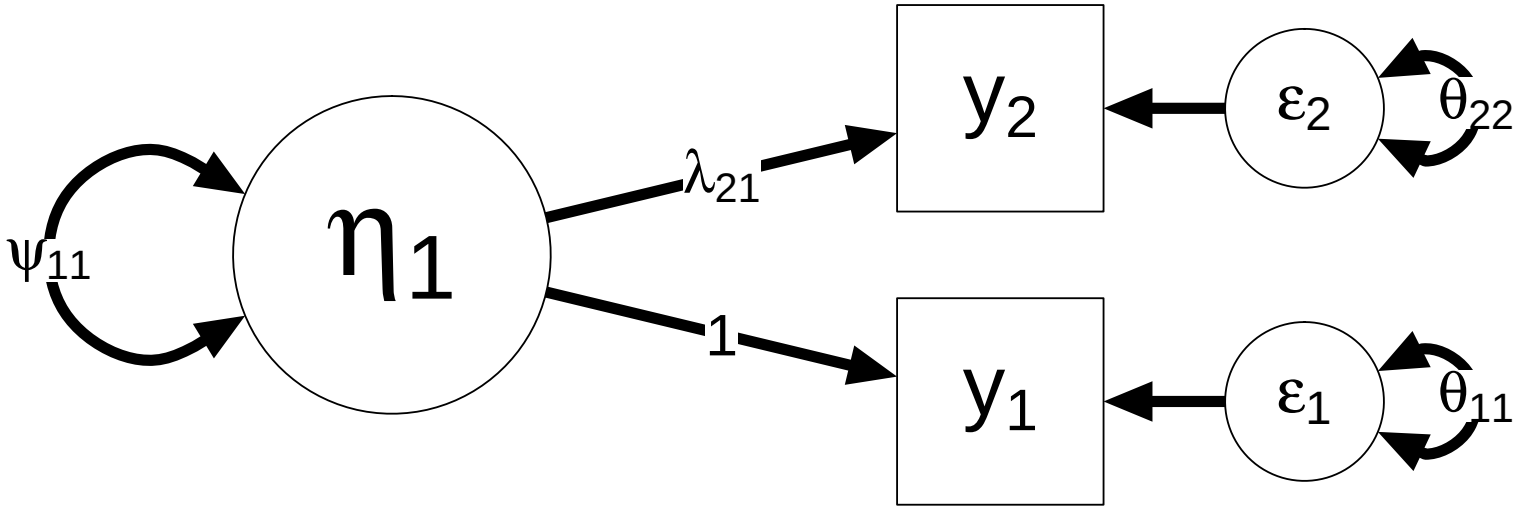
$$F_{\mathrm{ML}} = \mathrm{trace}\left(\boldsymbol{S}\boldsymbol{\Sigma}^{-1}\right) - \ln|\boldsymbol{S}\boldsymbol{\Sigma}^{-1}| - p,$$

in which $\boldsymbol{S}$ is the observed variance–covariance matrix, the trace operator takes the sum of diagonal values, the $|\ldots|$ notation indicates the determinant and p is the number of observed variables. This expression is proportional to the log *likelihood ratio* compared to a saturated model, and optimizing this expression is called maximum likelihood estimation. In principle, This expression is optimized if $\boldsymbol{\Sigma}$ resembles $\boldsymbol{S}$ as much as possible! When $\boldsymbol{S} = \boldsymbol{\Sigma}$, $F_{\mathrm{ML}} = 0$.



$$\mathbf{\Lambda} = \left[1\right], \mathbf{\Psi} = \left[\psi_{11}\right], \mathbf{\Theta} = \left[\theta_{11}\right]$$

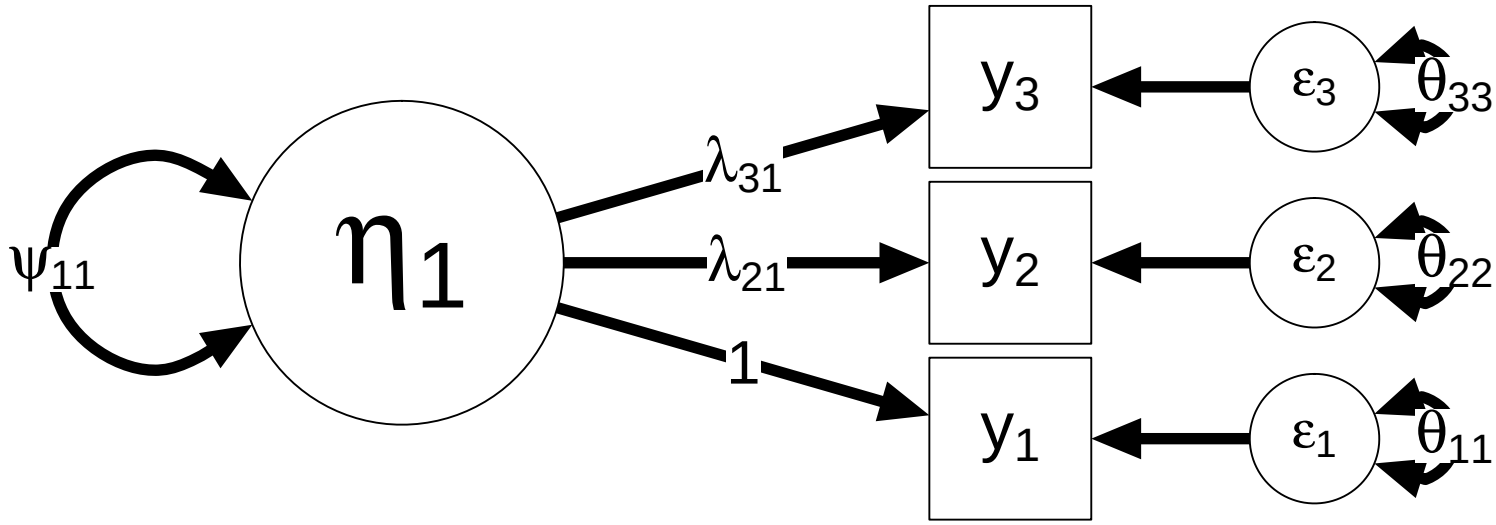
$$\mathbf{\Sigma} = \left[\psi_{11} + \theta_{11}\right]$$



$$\mathbf{\Lambda} = \begin{bmatrix} 1 \\ \lambda_{21} \end{bmatrix}, \mathbf{\Psi} = \begin{bmatrix} \psi_{11} \end{bmatrix}, \mathbf{\Theta} = \begin{bmatrix} \theta_{11} & \\ 0 & \theta_{22} \end{bmatrix}$$

$$\mathbf{\Sigma} = \begin{bmatrix} \psi_{11} + \theta_{11} & \\ \psi_{11}\lambda_{21} & \lambda_{21}^2\psi_{11} + \theta_{22} \end{bmatrix}$$

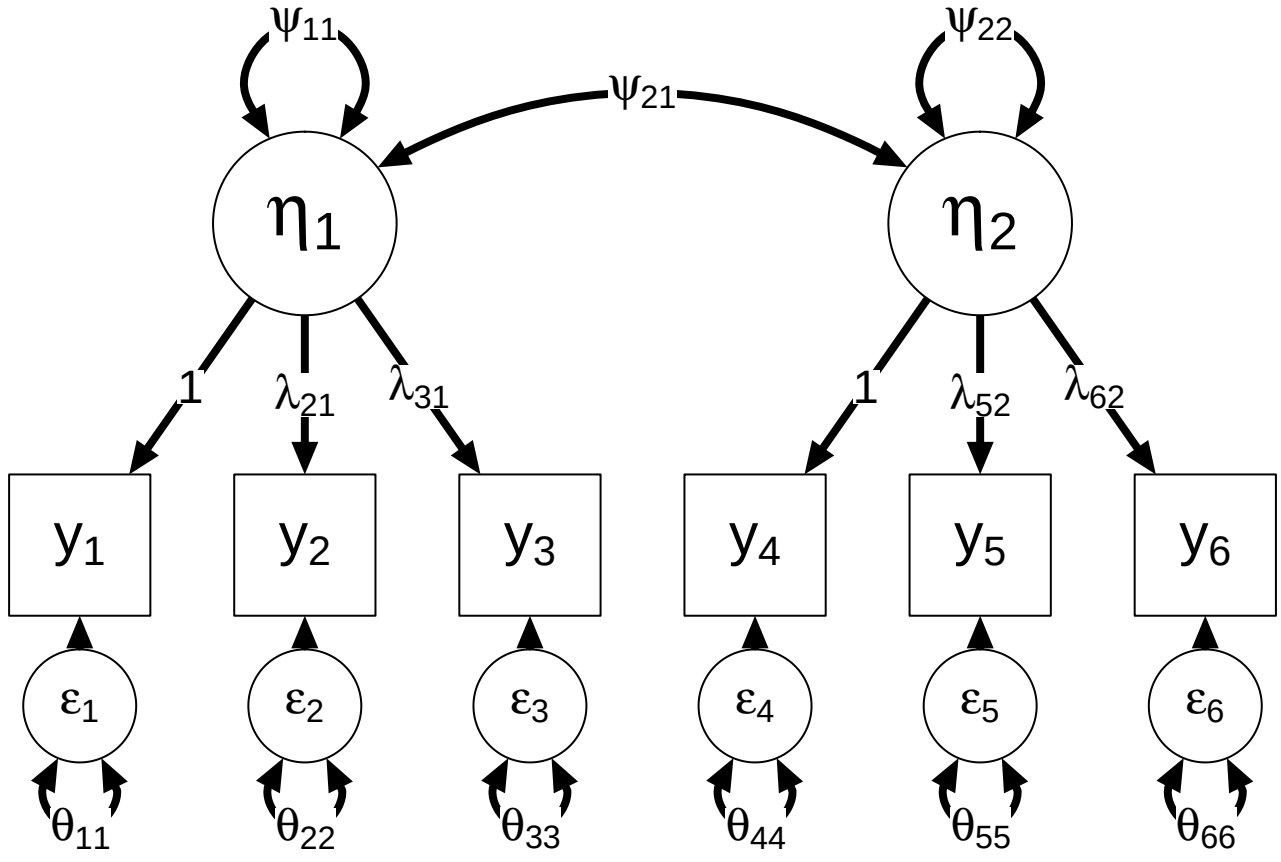
(upper triangular elements in symmetric matrices not shown)



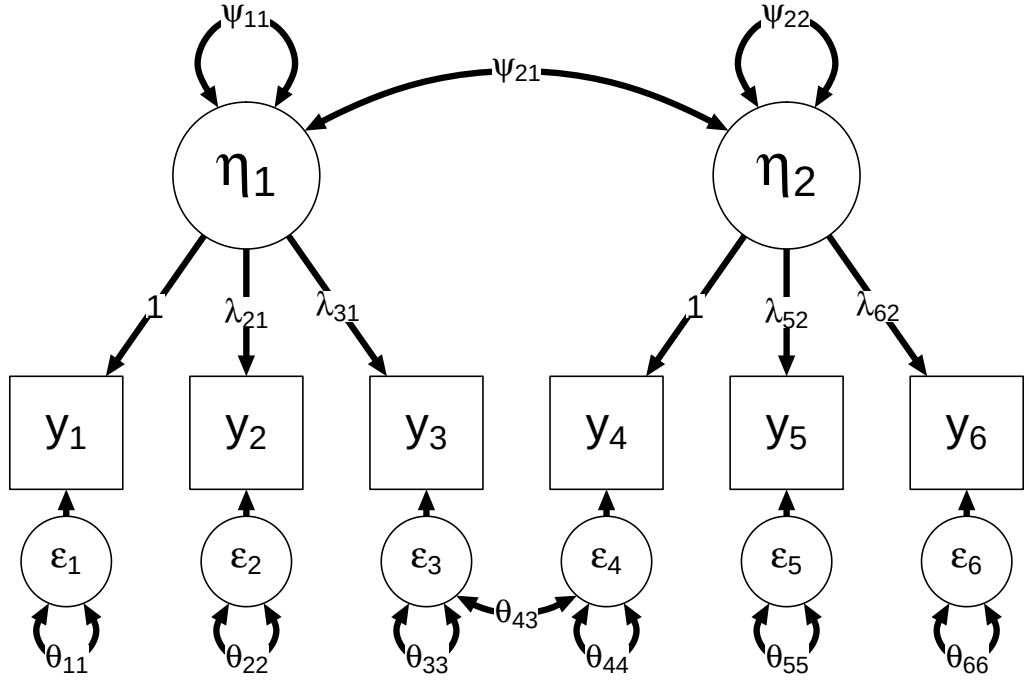
$$\mathbf{\Lambda} = \begin{bmatrix} 1 \\ \lambda_{21} \\ \lambda_{31} \end{bmatrix}, \mathbf{\Psi} = \begin{bmatrix} \psi_{11} \end{bmatrix}, \mathbf{\Theta} = \begin{bmatrix} \theta_{11} & & \\ 0 & \theta_{22} & \\ 0 & 0 & \theta_{33} \end{bmatrix}$$

$$\mathbf{\Sigma} = \begin{bmatrix} \psi_{11} + \theta_{11} & & \\ \psi_{11}\lambda_{21} & \lambda_{21}^2\psi_{11} + \theta_{22} & \\ \psi_{11}\lambda_{31} & \psi_{11}\lambda_{21}\lambda_{31} & \lambda_{31}^2\psi_{11} + \theta_{33} \end{bmatrix}$$

$\mathrm{DF} = 0$, just identified (but saturated, will explain the data perfectly)

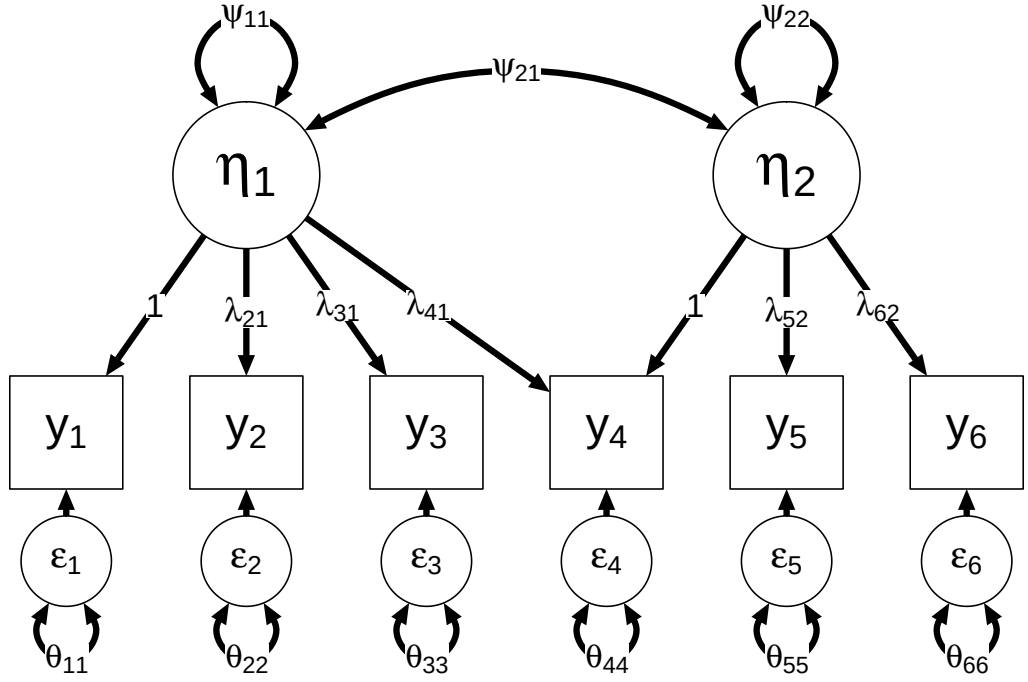


$$\mathbf{\Lambda} = \begin{bmatrix} 1 & 0 \\ \lambda_{21} & 0 \\ \lambda_{31} & 0 \\ 0 & 1 \\ 0 & \lambda_{52} \\ 0 & \lambda_{62} \end{bmatrix}, \mathbf{\Psi} = \begin{bmatrix} \psi_{11} & \\ \psi_{21} & \psi_{22} \end{bmatrix}, \mathbf{\Theta} = \begin{bmatrix} \theta_{11} & & & & & \\ 0 & \theta_{22} & & & & \\ 0 & 0 & \theta_{33} & & & \\ 0 & 0 & 0 & \theta_{44} & & \\ 0 & 0 & 0 & 0 & \theta_{55} & \\ 0 & 0 & 0 & 0 & 0 & \theta_{66} \end{bmatrix}$$



$$\mathbf{\Lambda} = \begin{bmatrix} 1 & 0 \\ \lambda_{21} & 0 \\ \lambda_{31} & 0 \\ 0 & 1 \\ 0 & \lambda_{52} \\ 0 & \lambda_{62} \end{bmatrix}, \mathbf{\Psi} = \begin{bmatrix} \psi_{11} & \\ \psi_{21} & \psi_{22} \end{bmatrix}, \mathbf{\Theta} = \begin{bmatrix} \theta_{11} & & & & & \\ 0 & \theta_{22} & & & & \\ 0 & 0 & \theta_{33} & & & \\ 0 & 0 & \theta_{43} & \theta_{44} & & \\ 0 & 0 & 0 & 0 & \theta_{55} & \\ 0 & 0 & 0 & 0 & 0 & \theta_{66} \end{bmatrix}$$

Residual covariance / correlation



$$\mathbf{\Lambda} = \begin{bmatrix} 1 & 0 \\ \lambda_{21} & 0 \\ \lambda_{31} & 0 \\ \lambda_{41} & 1 \\ 0 & \lambda_{52} \\ 0 & \lambda_{62} \end{bmatrix}, \mathbf{\Psi} = \begin{bmatrix} \psi_{11} & \\ \psi_{21} & \psi_{22} \end{bmatrix}, \mathbf{\Theta} = \begin{bmatrix} \theta_{11} & & & & & \\ 0 & \theta_{22} & & & & \\ 0 & 0 & \theta_{33} & & & \\ 0 & 0 & 0 & \theta_{44} & & \\ 0 & 0 & 0 & 0 & \theta_{55} & \\ 0 & 0 & 0 & 0 & 0 & \theta_{66} \end{bmatrix}$$

Cross-loading

Scale formation

Once we have a factor model and it fits the data (next week), we can use estimated model matrices to estimate the factor scores using a weighted centered score:

$$\hat{\boldsymbol{\eta}}_i = \boldsymbol{W}(\boldsymbol{y}_i - \bar{\boldsymbol{y}}).$$

The *regression method* forms the weights matrix as follows:

$$\boldsymbol{W} = \hat{\boldsymbol{\Psi}}\hat{\boldsymbol{\Lambda}}^\top\boldsymbol{\Sigma}^{-1},$$

And the *Bartlett method* forms the weights matrix as follows:

$$\boldsymbol{W} = \left(\boldsymbol{\Lambda}^\top\boldsymbol{\Theta}^{-1}\boldsymbol{\Lambda}\right)^{-1}\boldsymbol{\Lambda}^\top\boldsymbol{\Theta}^{-1}.$$

Note: $\hat{\boldsymbol{\eta}}_i$ is an estimate based on estimated parameter matrices, and therefore quite biased. For hypothesis testing, we *very rarely* need these estimates! For example, in SEM2 we will test causal hypotheses between latents without these estimates.

Structural Equation Modeling

- Factor analysis is part of the more general framework of Structural Equation Modeling (SEM)
- SEM allows for testing of causal relationships between both observed and latent variables
- We will discuss SEM in SEM 2!