Last time, we used *exploratory factor analysis* to explore potential factor structures from data:

- how many factors/dimensions?
- which items load onto the different factors?

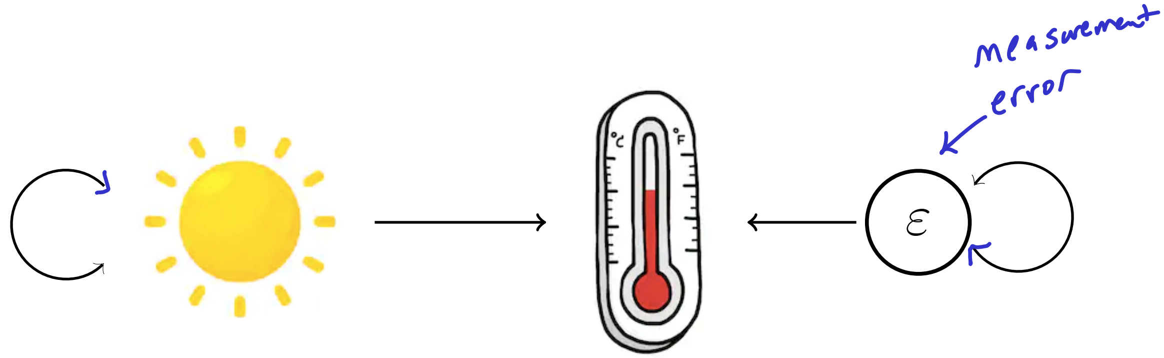
This time, we will use *confirmatory factor analysis* to **test** these factor structures and **estimate** their components.

To do this, we need to talk about "measurement models" and "path diagrams"

How do we measure temperature? By looking at a *thermometer*!

- for this to make sense, we need to assume the following:
  - temperature *causes* the reading on the thermometer
  - the thermometer has relatively little measurement error

So we have a *causal* hypothesis, which we can instantiate as a *measurement model*:

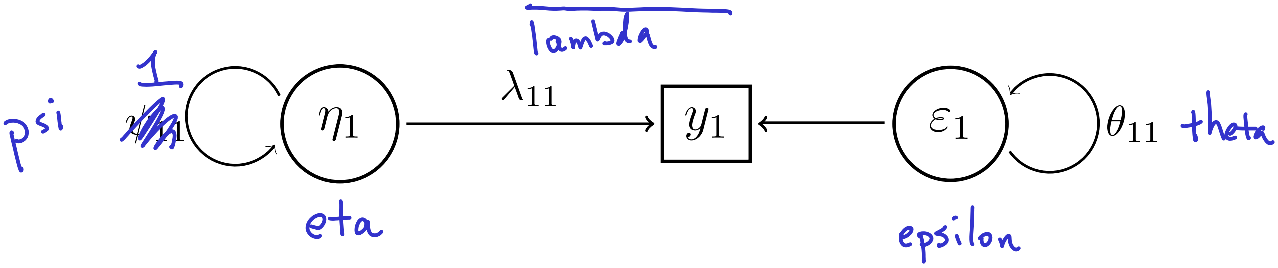


- the sun is a *latent variable* (not observable)
- the thermometer is a *observed variable*
  - also called an "indicator" of a latent variable

— nodes

- unidirectional links = causal effects
- bidirectional links = (co)variances

Let's formalize this idea with a *path diagram*:



- circles = latent (unobserved) factors
- squares = observed variables
- y_1 is *indicated by* factor η_1

This diagram encodes a lot of information about the causal relationship between factor η_1 and observation y_1

- $y_1 = \lambda_{11}\eta_1 + \varepsilon_1$
  - λ_{11} is the **loading** of factor η_1 onto observation y_i, and ε_1 is the **measurement error**
- $\eta_1 \sim \mathcal{N}(0, \sqrt{\underset{1}{\cancel{\psi_{11}}}})$
  - η_1 is assumed to be normally distributed with a mean of 0 and a variance of ψ_{11} (this is called the **factor variance**)
- $\varepsilon_1 \sim \mathcal{N}(0, \sqrt{\theta_{11}})$
  - ε_1 is assumed to be normally distributed with a mean of 0 and a variance of θ_{11} (this is called the **residual variance**)

Goal: given *observed data* y_1, we want to estimate the unknown *parameters* of the model:

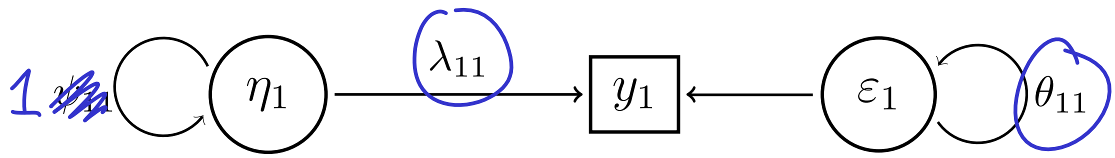
- the factor loading(s) λ_{11}
- the factor variance(s) ψ_{11}
- the residual variance(s) θ_{11}

To fit data to one of these **structural equation models**, we must make sure that two conditions hold:

1. we must *scale* the factors, either by
   - setting one of the loadings from each factor equal to 1, or
   - setting the factor variances equal to 1 (JASP does this one by default)
2. we must make sure that the number of observations (observed variances and covariances) **exceeds** the number of parameters (factor loadings/variances + residual variances)
   - the amount by which observations exceeds parameters is called the **degrees of freedom**

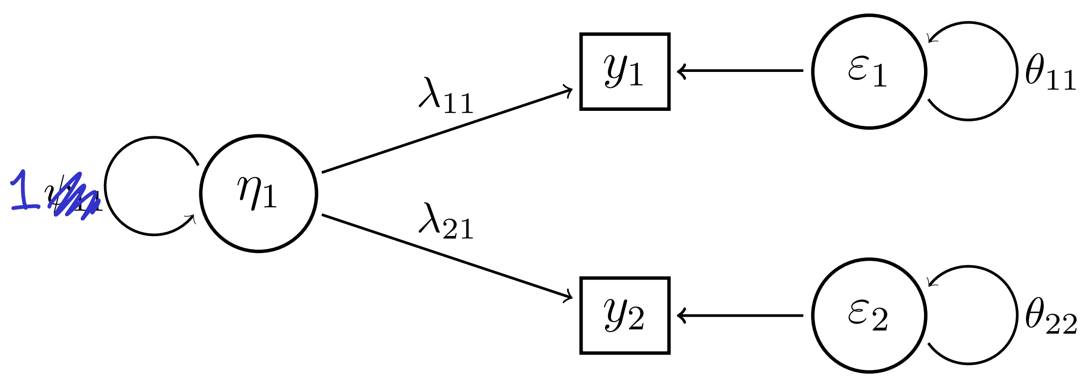
If these two conditions hold, we say that the model is **identified**.

Let's do some examples



- Number of observations: 1
- Number of parameters: 2
- Degrees of freedom: # obs − # par = 1 − 2 = −1

  Not identified

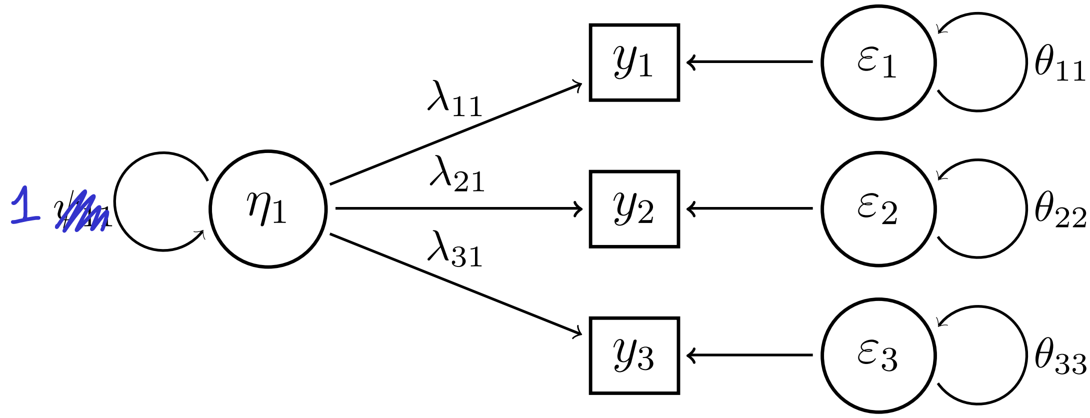


- Number of observations: obs. variances & covariances

  3
- Number of parameters: loadings = 2

  resid. variances = 2 → 4
- Degrees of freedom: # obs − # par = 3 − 4 = −1

  Not identified



Shortcut: p - obs. variables

$$\boxed{\# \text{ obs} = \frac{p(p+1)}{2}}$$

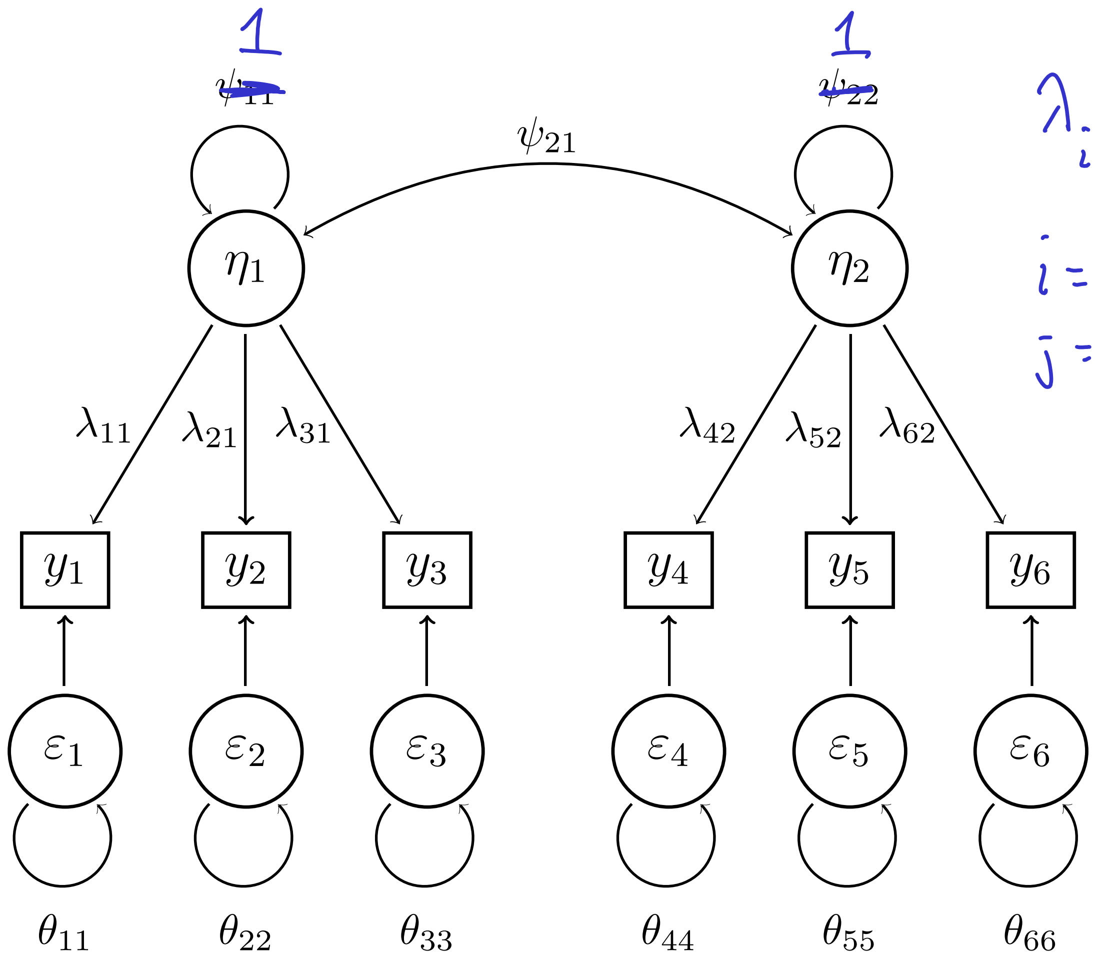
- Number of observations: $3 + 2 + 1 = 6$

  $\frac{3 \cdot 4}{2} = 6$!

- Number of parameters: 6

- Degrees of freedom: $\#\text{ obs} - \#\text{ par} = 6 - 6 = 0$

  identified - "saturated model"
  - exact fit
  - one, unique solution to this model



λ_{ij}

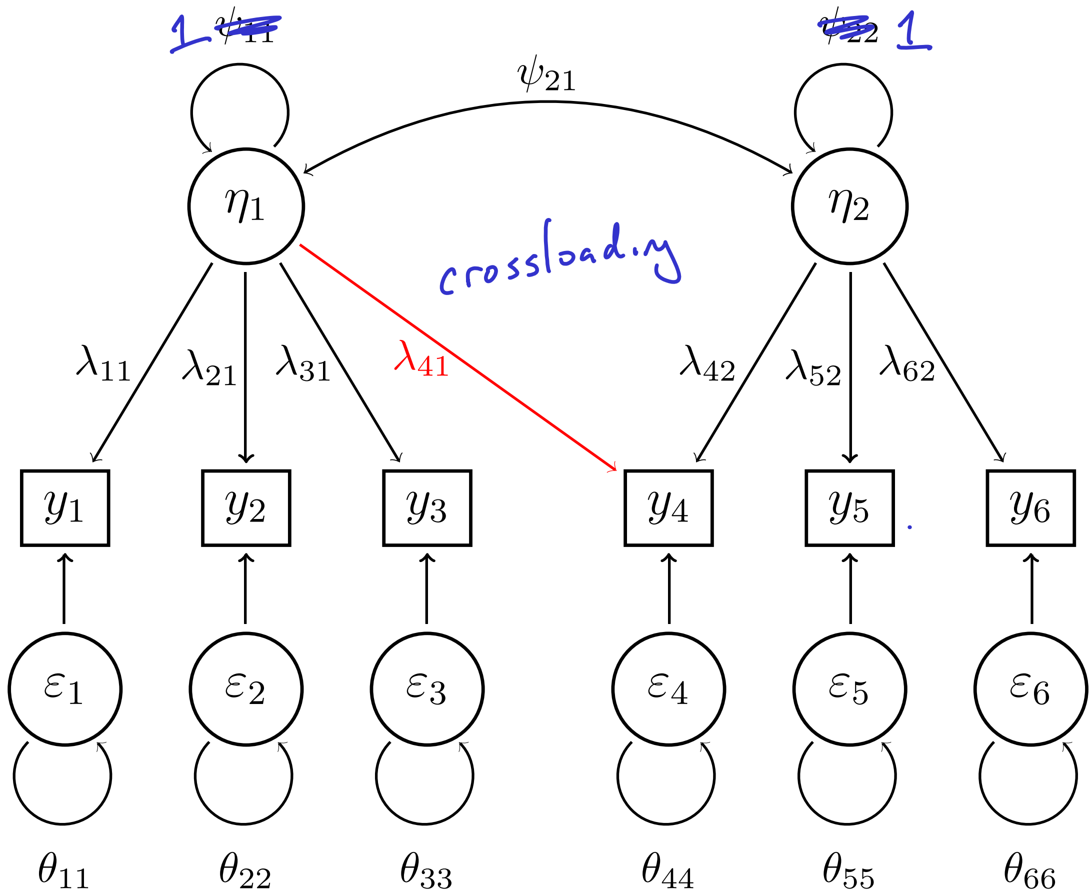
i = item

j = factor

- Number of observations: $\frac{6\cdot(6+1)}{2} = \frac{6\cdot 7}{2} = 21$
- Number of parameters: 6 loadings, 6 res. var, 1 factor cov

  = 13 par.
- Degrees of freedom: # obs − # par = 21 − 13 = 8

  identified

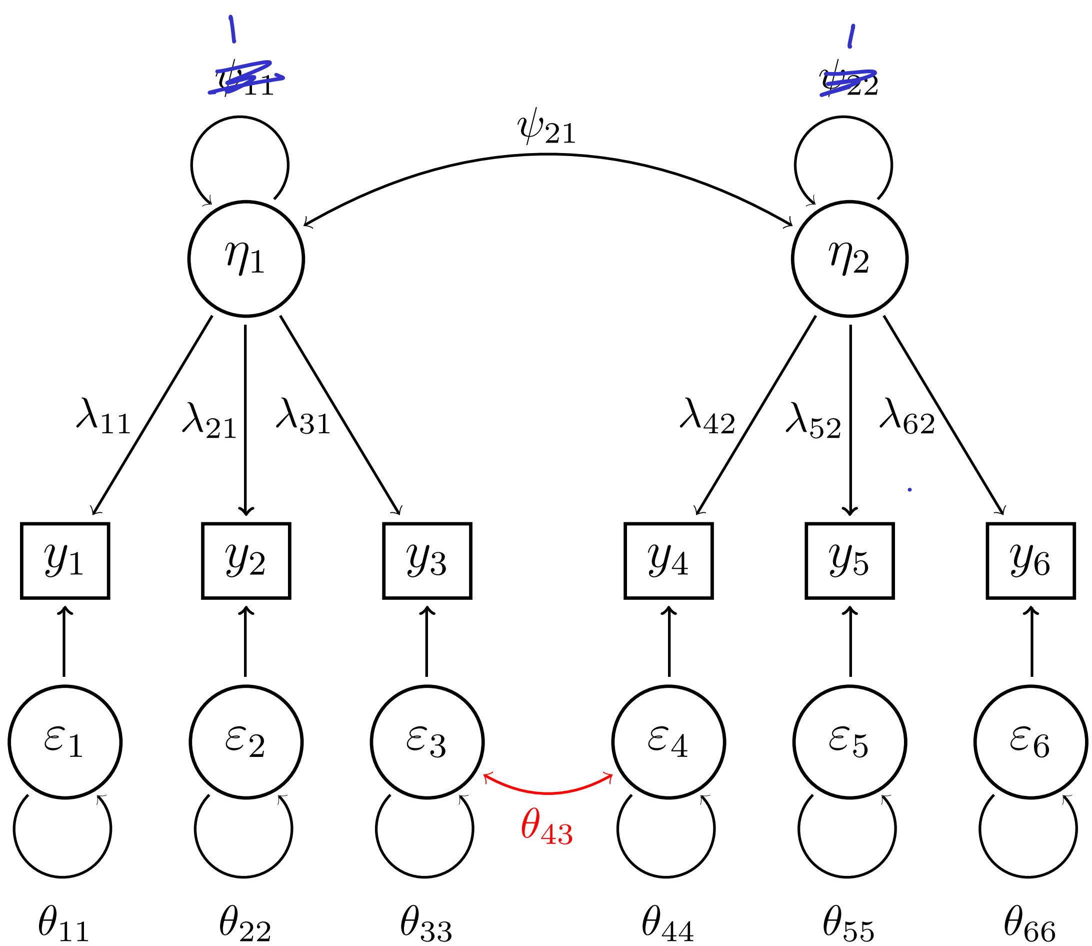


- Number of observations: 21

- Number of parameters: $1 + 7 + 6 = 14$

- Degrees of freedom: # obs - # par = $21 - 14 = 7$

  identified



- Number of observations: 21

- Number of parameters:
  1 factor cov.
  6 factor loadings $\longrightarrow$ 14
  7 res. (co)variances

- Degrees of freedom: # obs − # par = 21 − 14 = 7
  identified

Let's try fitting a model in JASP.

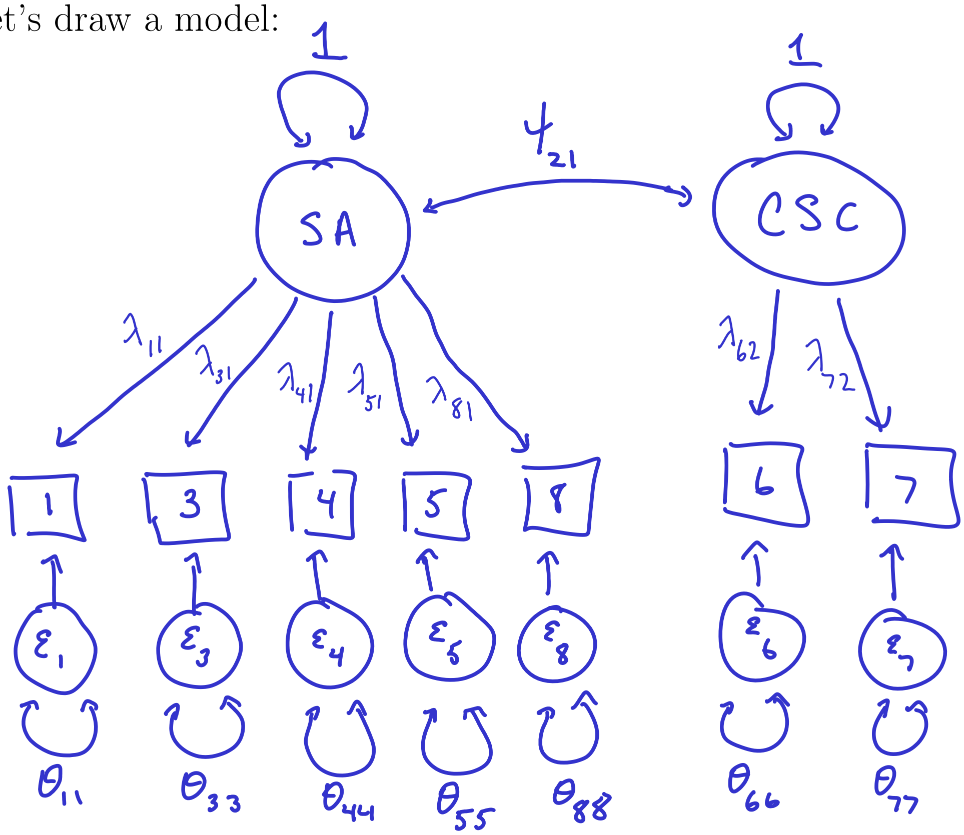
Suppose we are measuring statistics anxiety with the *SAQ-8* – an 8-item "statistics anxiety questionnaire". Each item is Likert scaled with 1 = strongly disagree and 5 = strongly agree.
Items:

1. Statistics makes me cry
2. My friends will think I'm stupid for not being able to use statistical software
3. Standard deviations excite me
4. I dream that Pearson is attacking me with correlation coefficients
5. I don't understand statistics
6. I have little experience with computers
7. All computers hate me
8. I have never been good at mathematics

From last we found the following (potential) factor structure:

- Factor 1: "statistics anxiety"
  - 1. Statistics makes me cry
  - 3. Standard deviations excite me
  - 4. I dream that Pearson is attacking me with correlation coefficients
  - 5. I don't understand statistics
  - 8. I have never been good at mathematics
- Factor 2: "computer self concept"
  - 6. I have little experience with computers
  - 7. All computers hate me

Let's draw a model:



$\#\,obs = \frac{7(8)}{2} = 28$

$\#\,par = 1 + 7 + 7 = 15$

$df = \#\,obs - \#\,par$

$= 28 - 15$

$= 13$

So how does the model fit?

- JASP computes a fit statistic T
- If the model fits **exactly**, then T is distributed as a χ^2 distribution
- so, JASP reports a χ^2 test
  - if $p < 0.05$, we reject $\mathcal{H}_0$, which implies the model does NOT fit
  - if $p > 0.05$, we accept $\mathcal{H}_0$, which implies the model DOES fit

Some notes about χ^2 test:

- χ^2 is a measure of "exact fit" – smaller is better
- for large N, the χ^2 test tend to reject models even when the fit is close (this is a problem!)

Alternative method of assessing fit - *RMSEA*

- "root mean squared error of approximation"
- it is a measure of "absolute fit" (i.e., there is no comparison model)
- smaller is better
- Guidelines:
  - < 0.05 = very good fit
  - $0.05 - 0.08$ = good fit
  - > 0.08 = unacceptable fit
- RMSEA is one of the only fit indices for which the sampling distribution is known. Thus, confidence intervals can be computed (and are reported in JASP)