Why SEM? How structural equation modelling can help you make the most of your data

Uku Vainik
University of Tartu & McGill University

Health & Personality Development meeting
April 27,2020

uku.vainik@ut.ee

Uku Vainik, PhD

- 2015 – PhD in psychology from Tartu, Estonia
- 2016-19 – post-doc at Montreal Neurological Institute, McGill University, Canada
- 2019 – research fellow in Institute of Psychology, University of Tartu, Estonia
- 2020 – adjunct professor at MNI, McGill, Canada
- I study behavioural signature of obesity and overeating
  - Personality, cognitive abilities, brain structure, genetics

Experience with SEM

- Modelling jangle fallacy with bifactor models and IRT
  - Vainik et al., 2015, Appetite
  - Mason, Vainik et al., 2017. Frontiers in Psychology
- Testing assumptions of a factor model – indifference of indicator
  - Vainik et a., 2015, European Journal of Personality
- Modelling heritability, genetic correlations, and causal inference
  - Vainik et al., 2018, PNAS
  - Arumäe, ... , Vainik, 2020, Nutrixiv

Outline of the talk

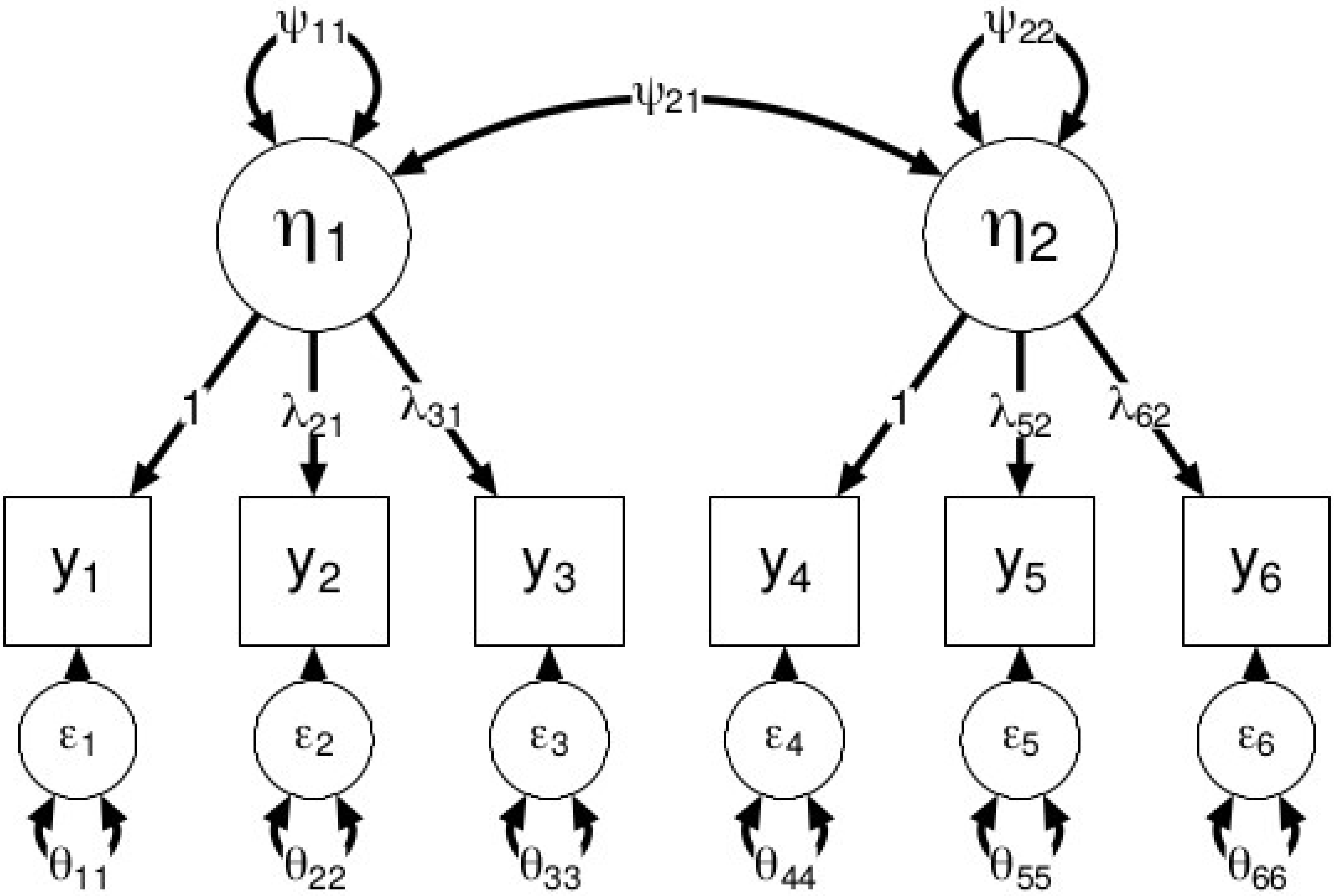
- What is SEM
- Factor analysis
- Path models
- Useful resources
- → A very short version, occasionally borrowing material from Yves Rosseel (R package lavaan) and Sacha Epskamp (R package semPlot)

Factor analysis

- Psychological research uses psychological constructs
  - Intelligence, personality, disorders, etc
- We have to measure these constructs
  - Questionnaires, tests
  - Measurement is indirect..
- Factor analysis tests, how well a test measures underlying trait
  - Exploratory factor analysis (EFA) – no hypothesis underlying causal structure
  - Confirmatory Factor Analysis (CFA) – hypothesizing a particular structure

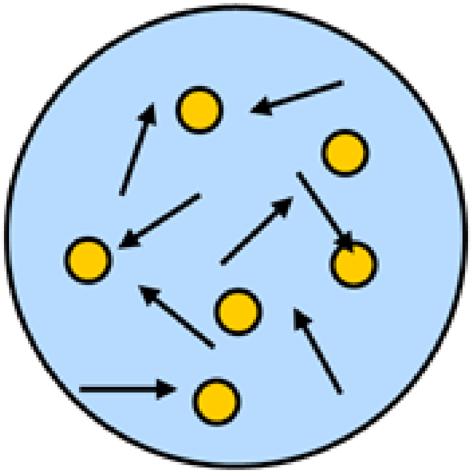
Based on Epskamp slides

A two-factor model



Based on Epskamp slides

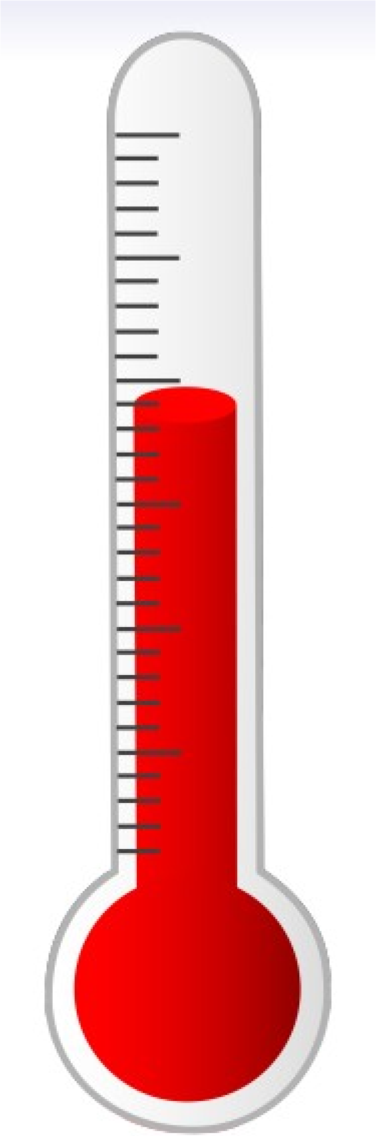
How do we measure temperature?



Based on Epskamp slides

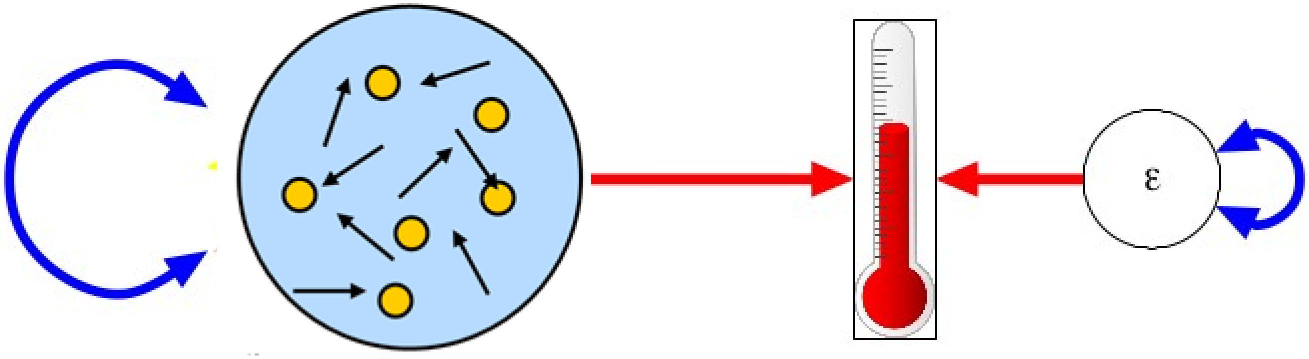
How do we measure temperature?

- We look at a thermometer
- For this to make sense, we need to assume that:
  - Temperature causes the level of the thermometer
  - Thermometer has little measurement error



Based on Epskamp slides

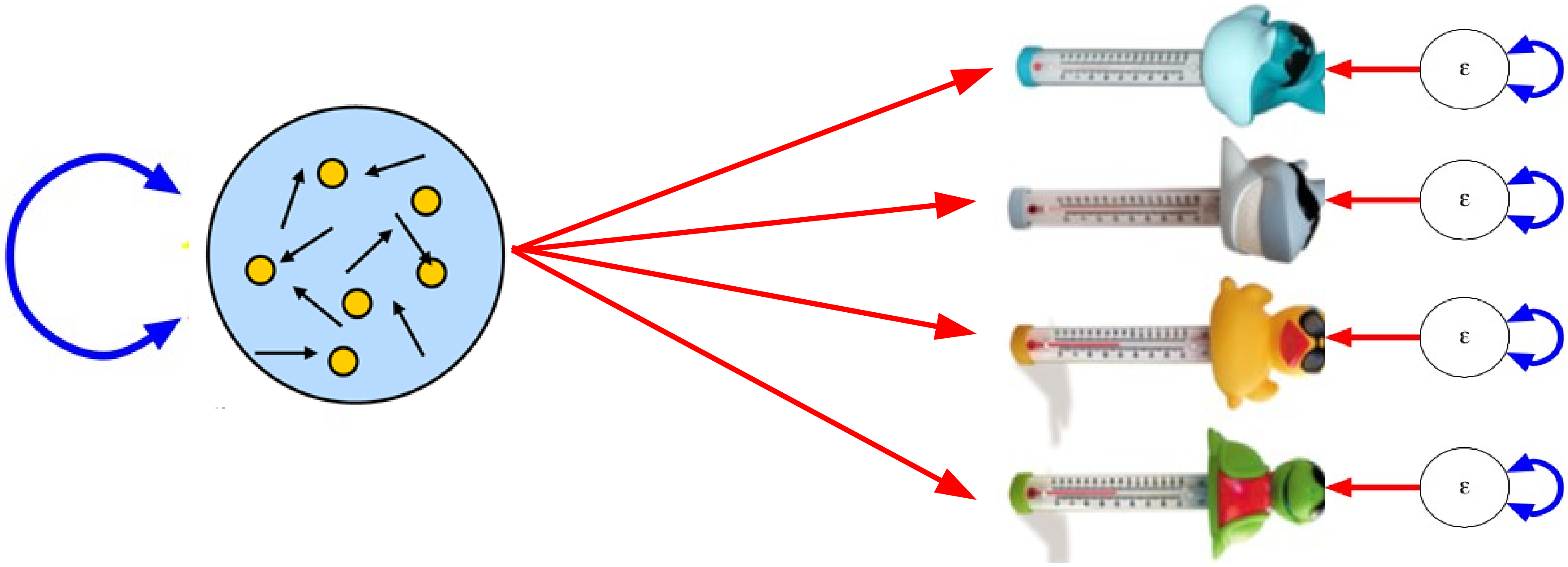
Path diagram of causal hypothesis



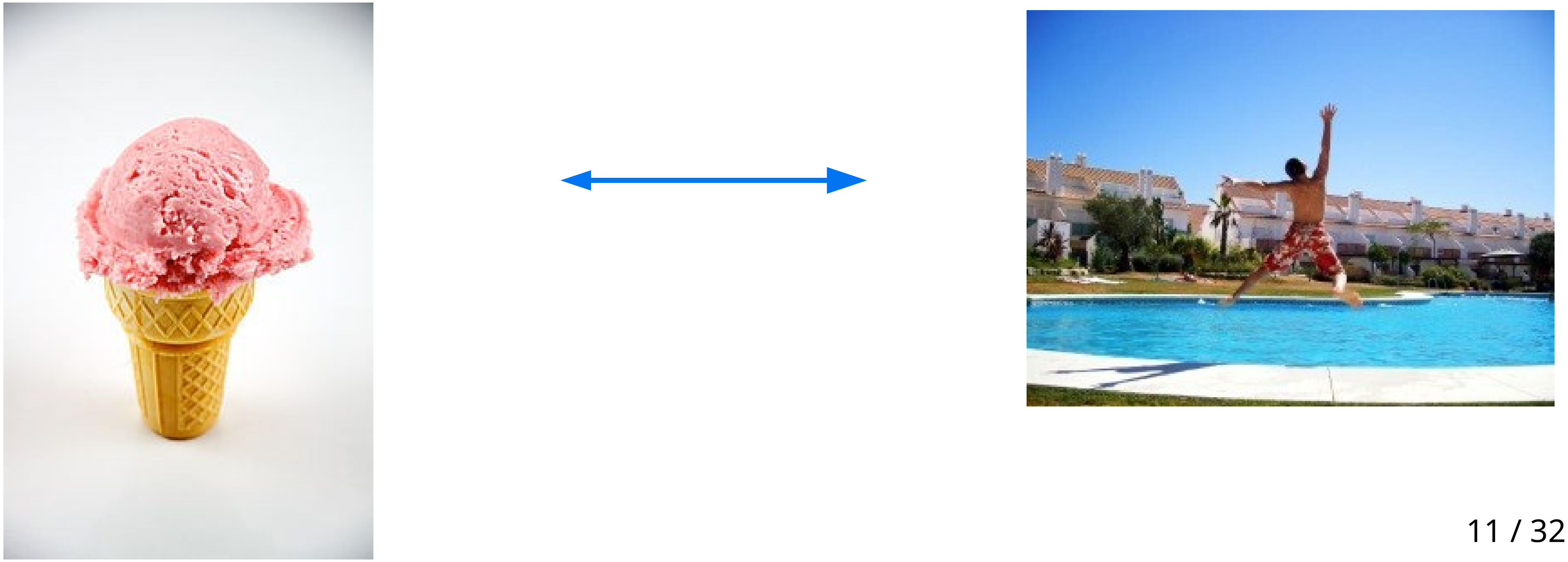
- Circular nodes – latent (unobserved variables)
- Rectangular nodes – observed variables
  - Indicators of latent variable
- Unidirectional links → causal effects
- Bidirectional links ← → (co)variances

Based on Epskamp slides

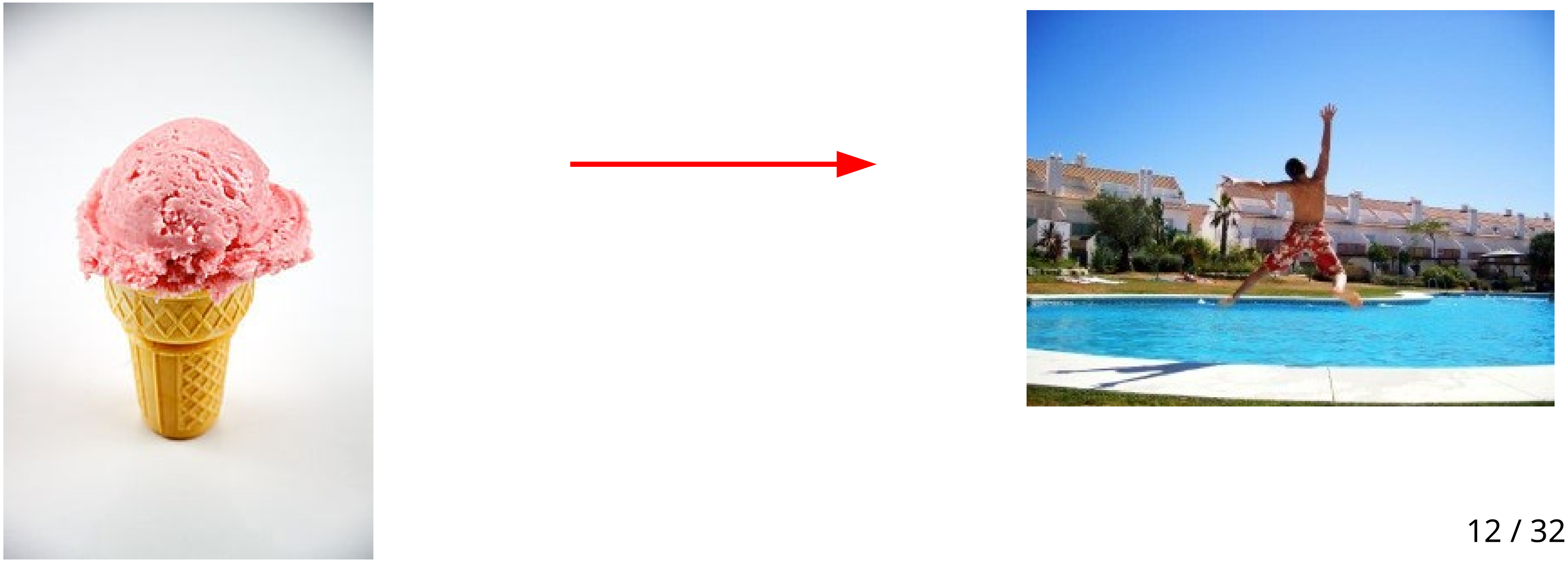
To reduce measurement error, we have multiple indicators for latent variable



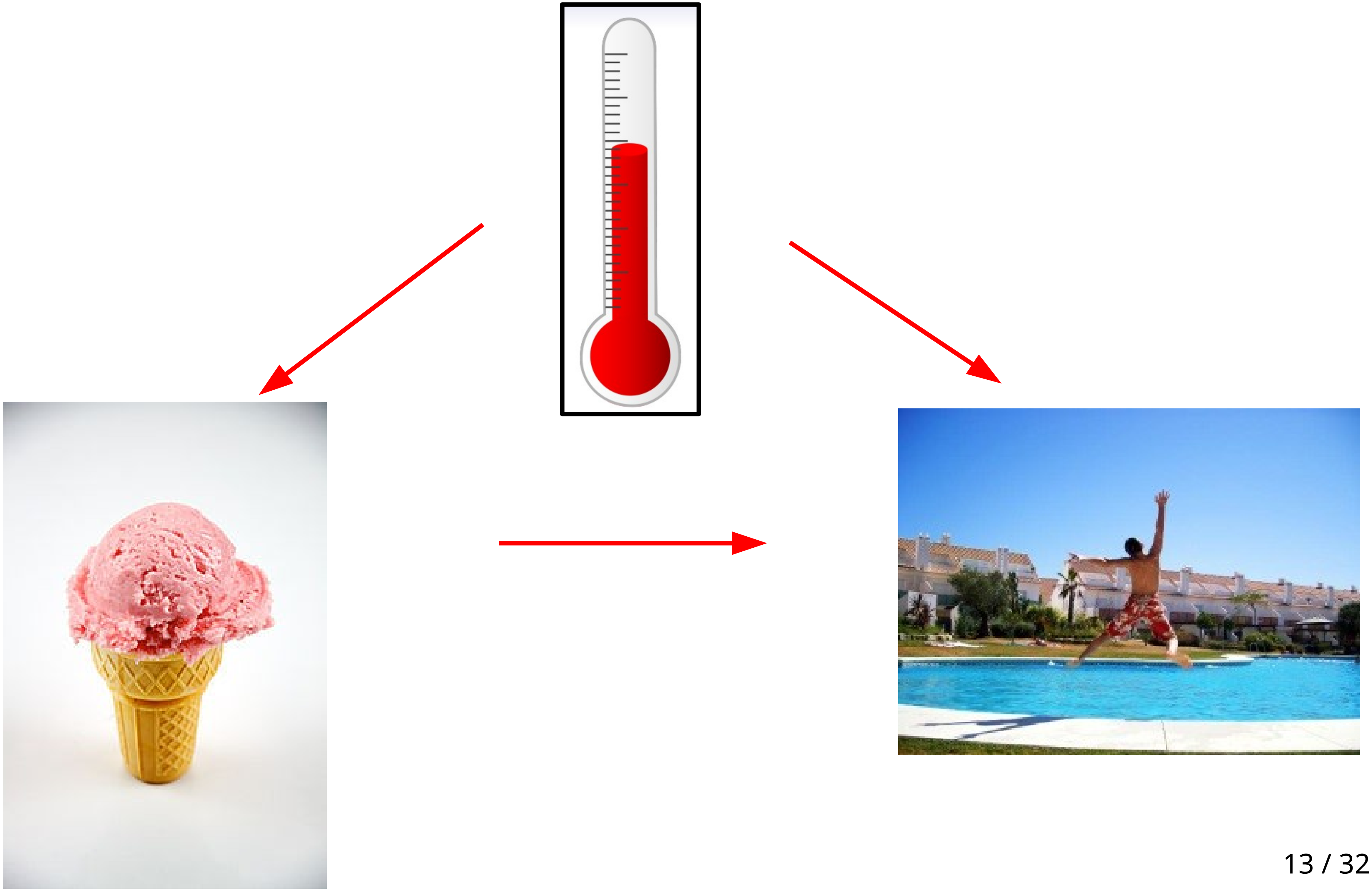
Path analysis



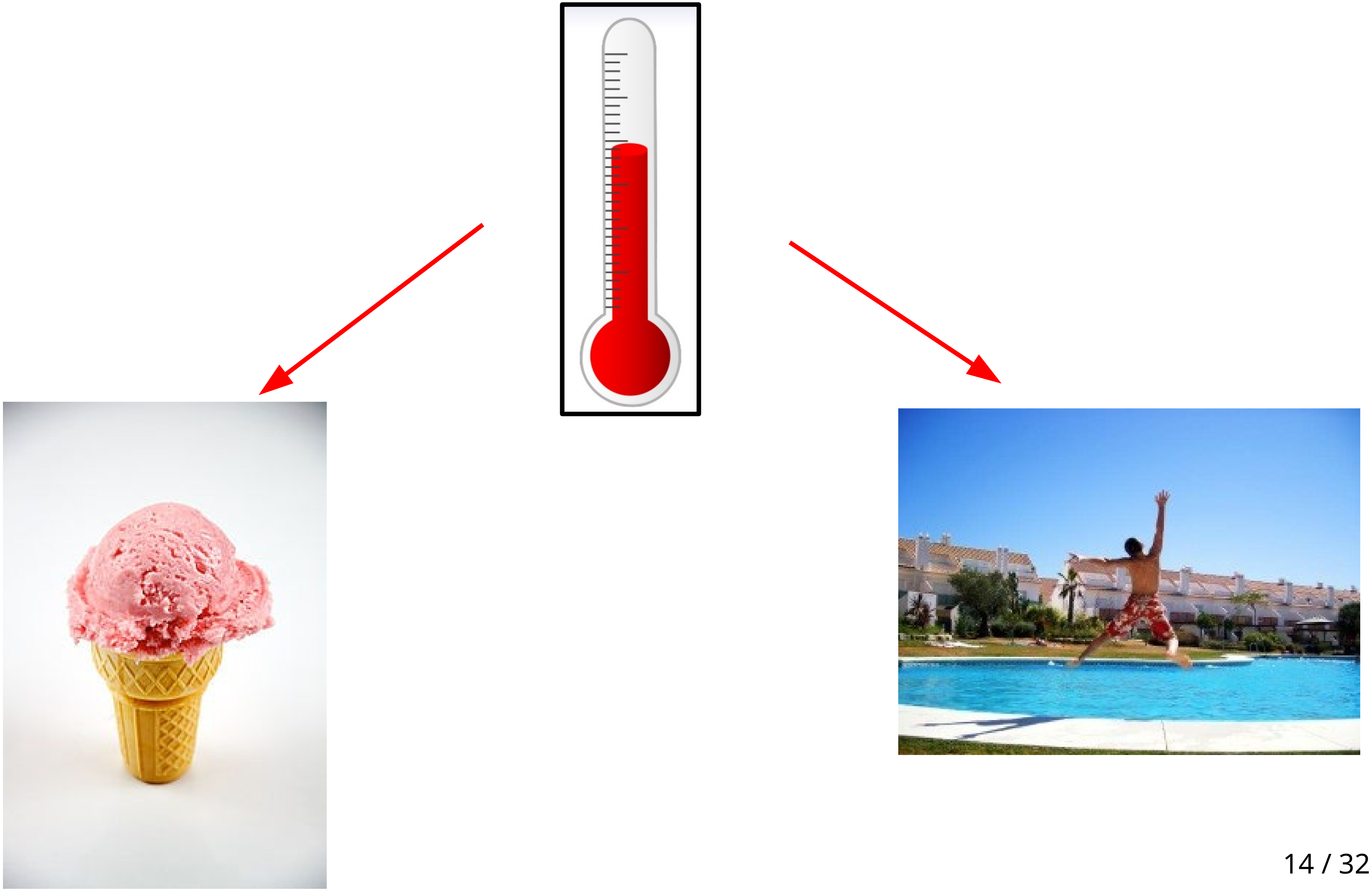
Path analysis



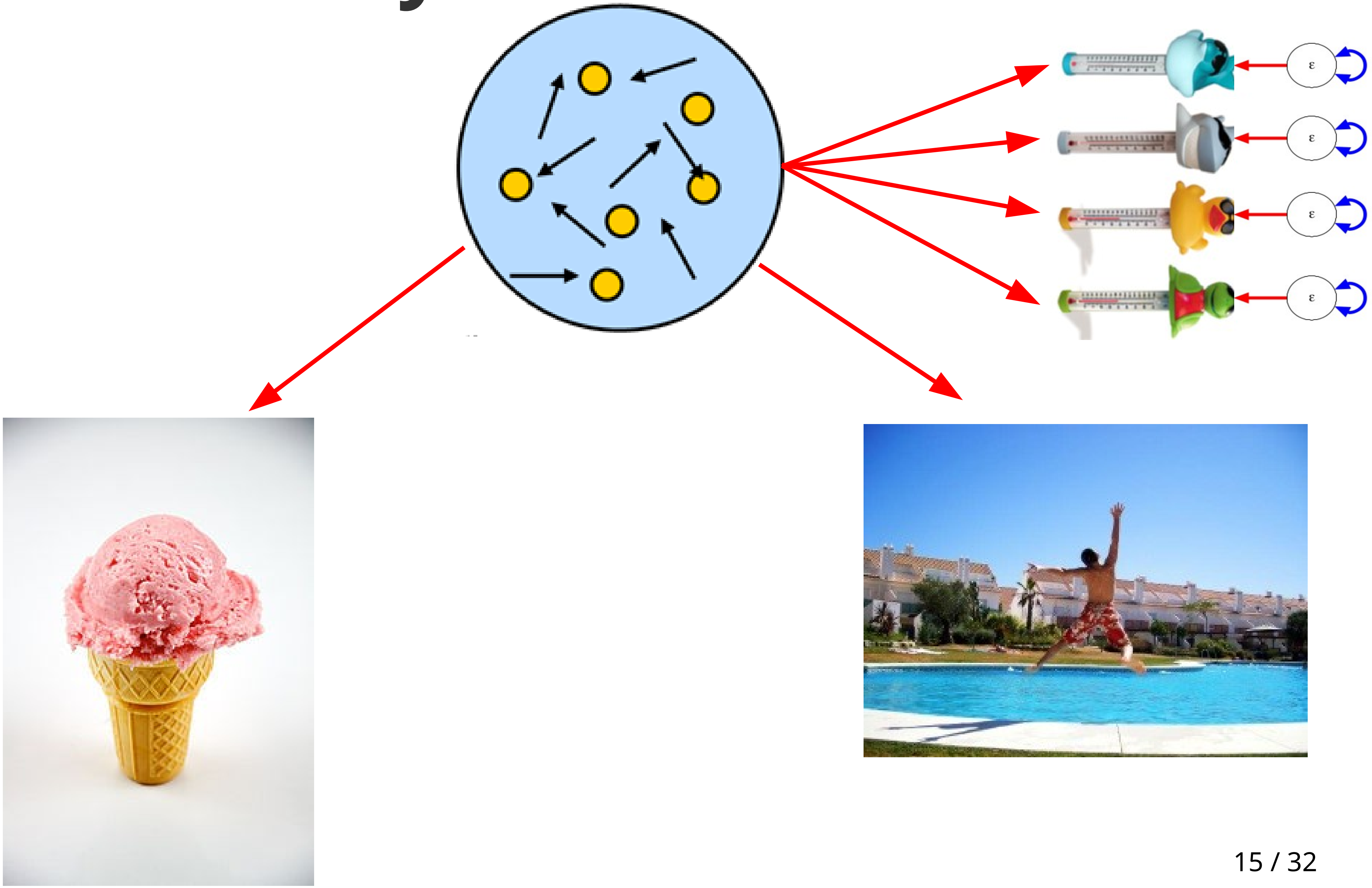
Path analysis



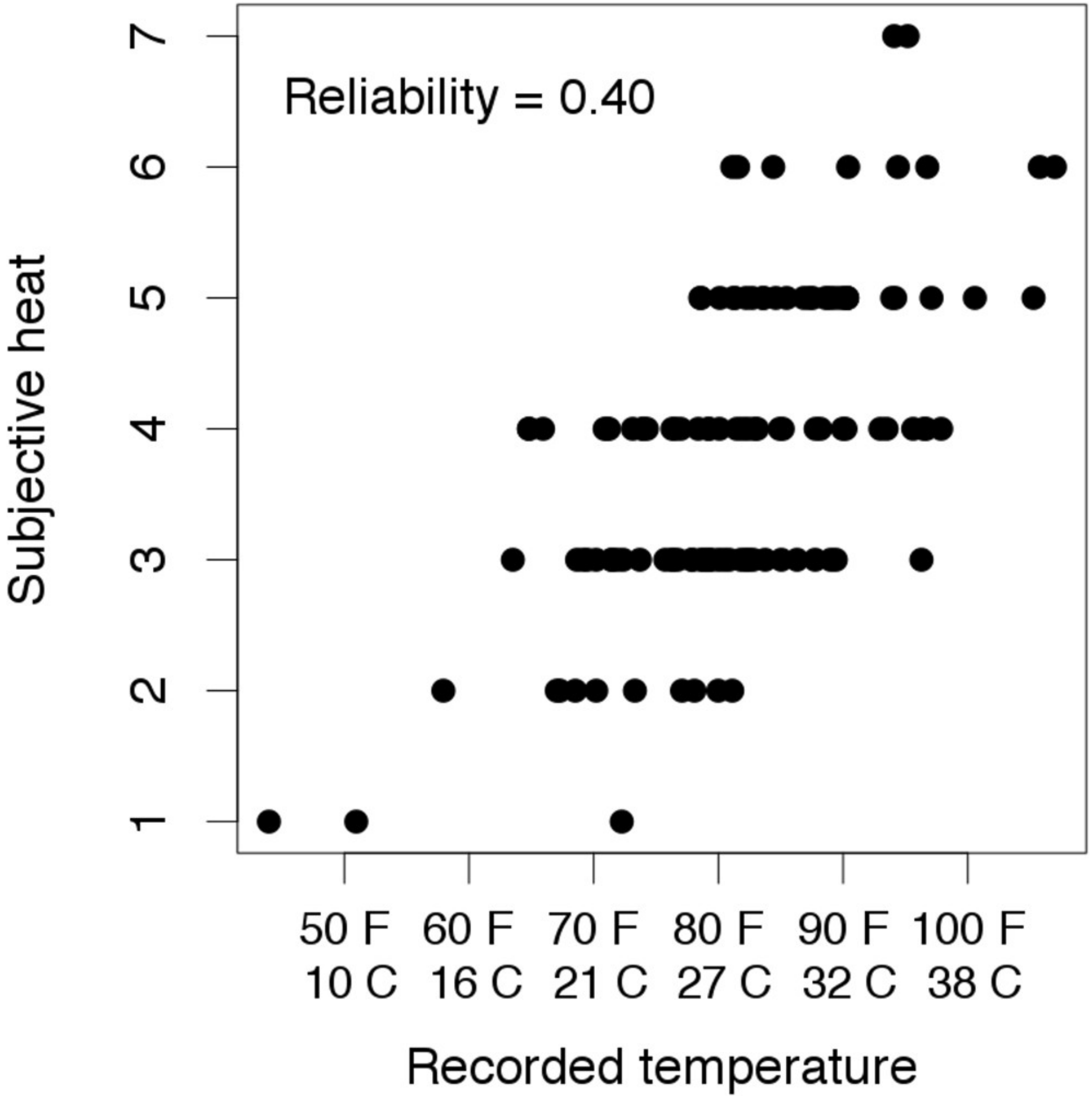
Path analysis



Path analysis + SEM

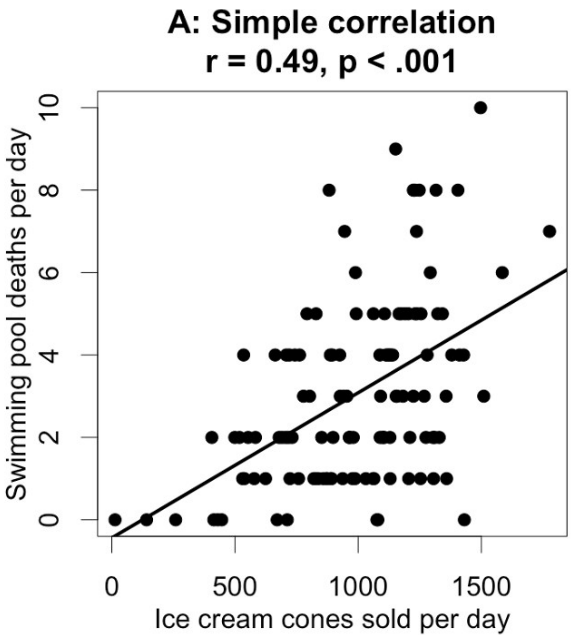


SEM can correct for reliability



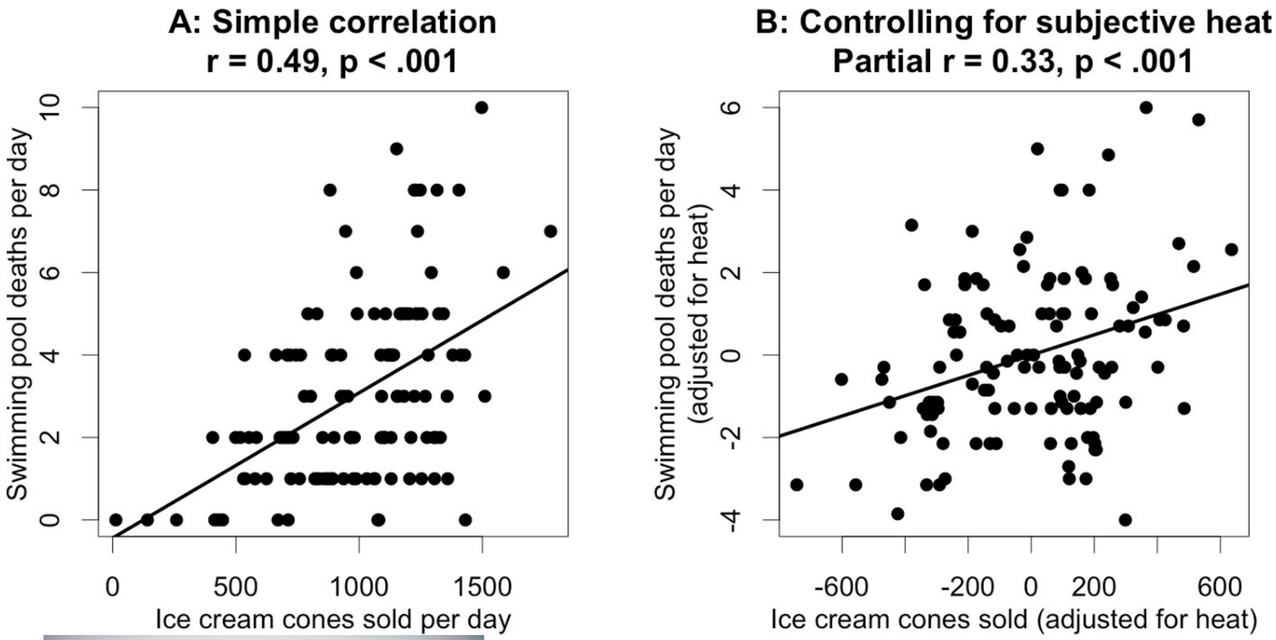
Westfall & Yarkoni, 2016, PlosONE

Simple correlation



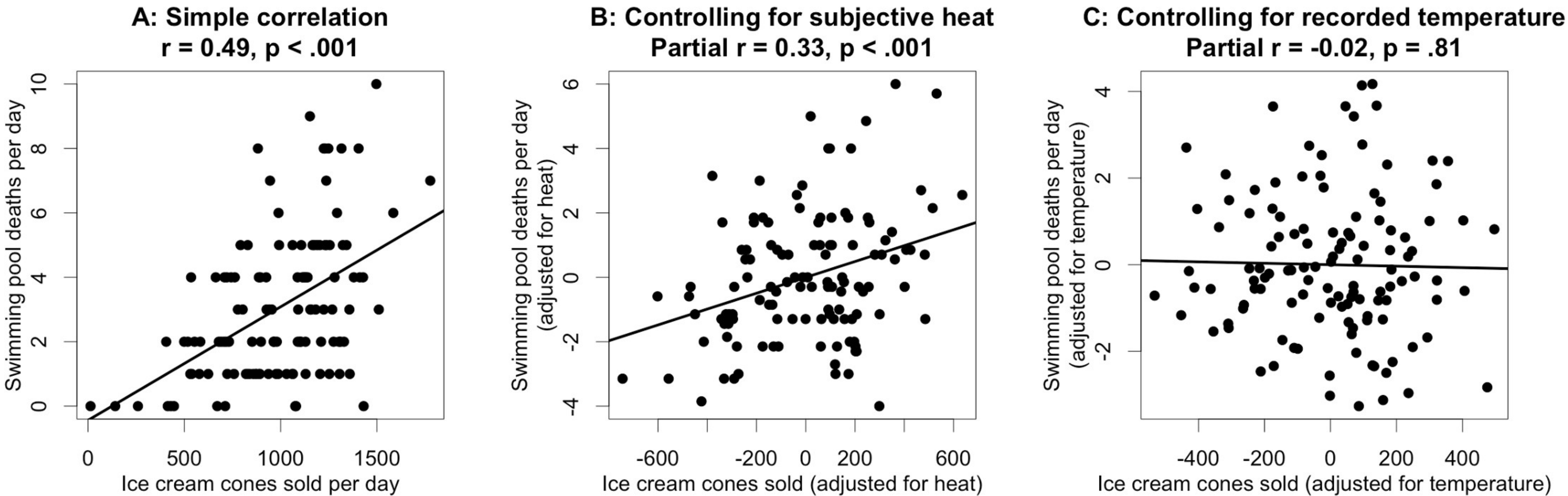
Westfall & Yarkoni, 2016, PlosONE

Controlling for subjective heat



Westfall & Yarkoni, 2016, PlosONE

Controlling for recorded temp



A: Simple correlation
r = 0.49, p < .001

B: Controlling for subjective heat
Partial r = 0.33, p < .001

C: Controlling for recorded temperature
Partial r = -0.02, p = .81

Westfall & Yarkoni, 2016, PlosONE

Factor analysis + Rasch = IRT

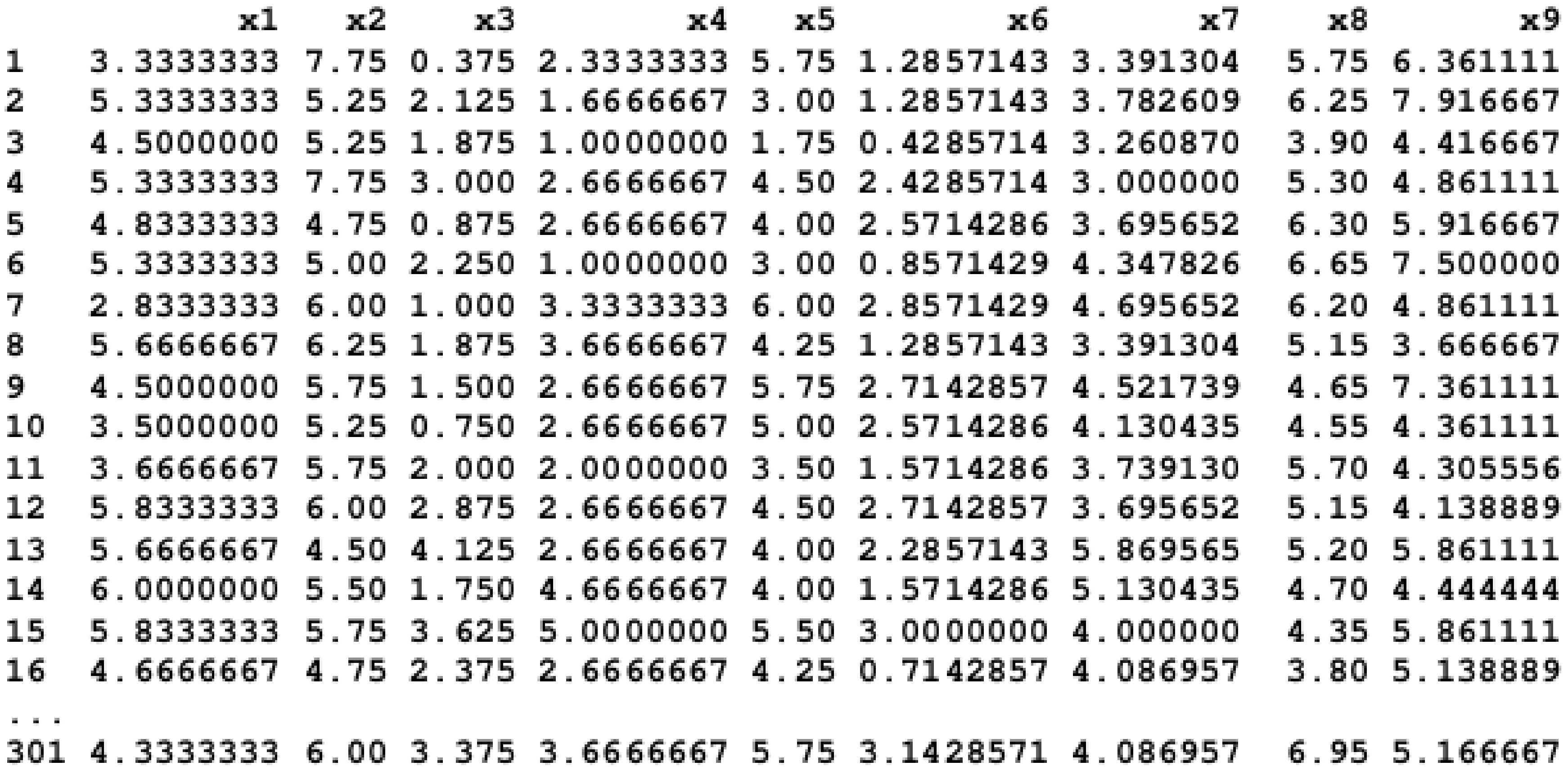
- FA – construct indicators vary in contribution
  - Self report temperature worse than home thermometer
- Rasch – construct indicators vary in difficulty
  - Home grade thermometer useless below -30 C or above 50 C.
  - We also need a very focused thermometer for body temperature
  - Rasch= 1 parameter item response theory, IRT
- FA + Rasch = 2 parameter IRT

http://www.personality-project.org/r/book/#chapter8

Goal of SEM

- To find a model that explains the data with the least complex model
- Estimation of certain model parameters
- Model-implied variance-covariance matrix should resemble observed variance-covariance matrix

data

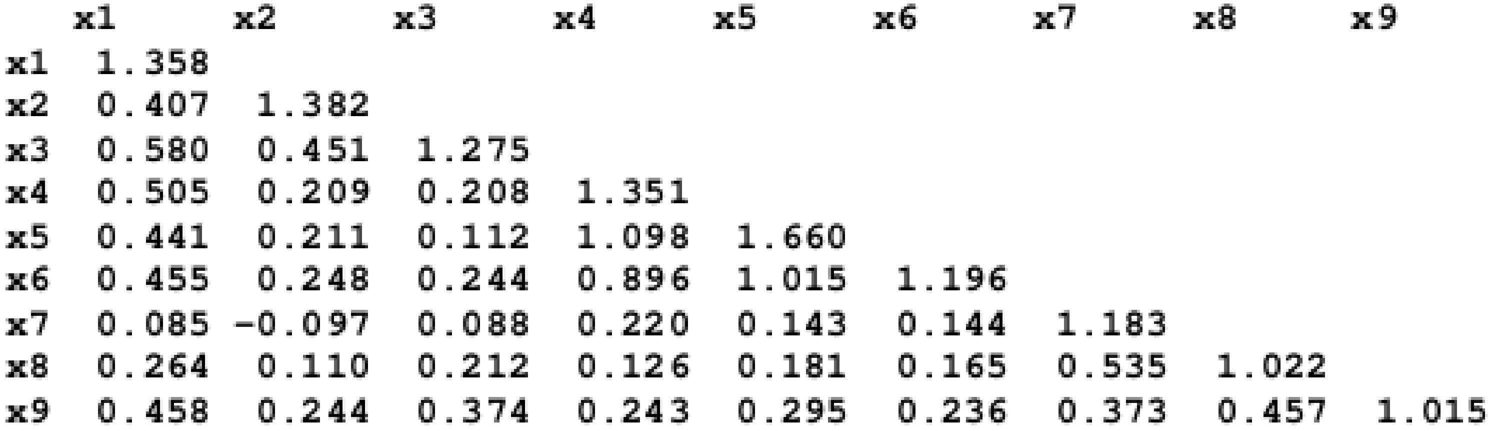


```
           x1   x2    x3        x4   x5        x6       x7   x8       x9
1   3.3333333 7.75 0.375 2.3333333 5.75 1.2857143 3.391304  5.75 6.361111
2   5.3333333 5.25 2.125 1.6666667 3.00 1.2857143 3.782609  6.25 7.916667
3   4.5000000 5.25 1.875 1.0000000 1.75 0.4285714 3.260870  3.90 4.416667
4   5.3333333 7.75 3.000 2.6666667 4.50 2.4285714 3.000000  5.30 4.861111
5   4.8333333 4.75 0.875 2.6666667 4.00 2.5714286 3.695652  6.30 5.916667
6   5.3333333 5.00 2.250 1.0000000 3.00 0.8571429 4.347826  6.65 7.500000
7   2.8333333 6.00 1.000 3.3333333 6.00 2.8571429 4.695652  6.20 4.861111
8   5.6666667 6.25 1.875 3.6666667 4.25 1.2857143 3.391304  5.15 3.666667
9   4.5000000 5.75 1.500 2.6666667 5.75 2.7142857 4.521739  4.65 7.361111
10  3.5000000 5.25 0.750 2.6666667 5.00 2.5714286 4.130435  4.55 4.361111
11  3.6666667 5.75 2.000 2.0000000 3.50 1.5714286 3.739130  5.70 4.305556
12  5.8333333 6.00 2.875 2.6666667 4.50 2.7142857 3.695652  5.15 4.138889
13  5.6666667 4.50 4.125 2.6666667 4.00 2.2857143 5.869565  5.20 5.861111
14  6.0000000 5.50 1.750 4.6666667 4.00 1.5714286 5.130435  4.70 4.444444
15  5.8333333 5.75 3.625 5.0000000 5.50 3.0000000 4.000000  4.35 5.861111
16  4.6666667 4.75 2.375 2.6666667 4.25 0.7142857 4.086957  3.80 5.138889
...
301 4.3333333 6.00 3.375 3.6666667 5.75 3.1428571 4.086957  6.95 5.166667
```

- data is complete
- the ‘covariance matrix’ contains all information about the interrelations among the observed variables

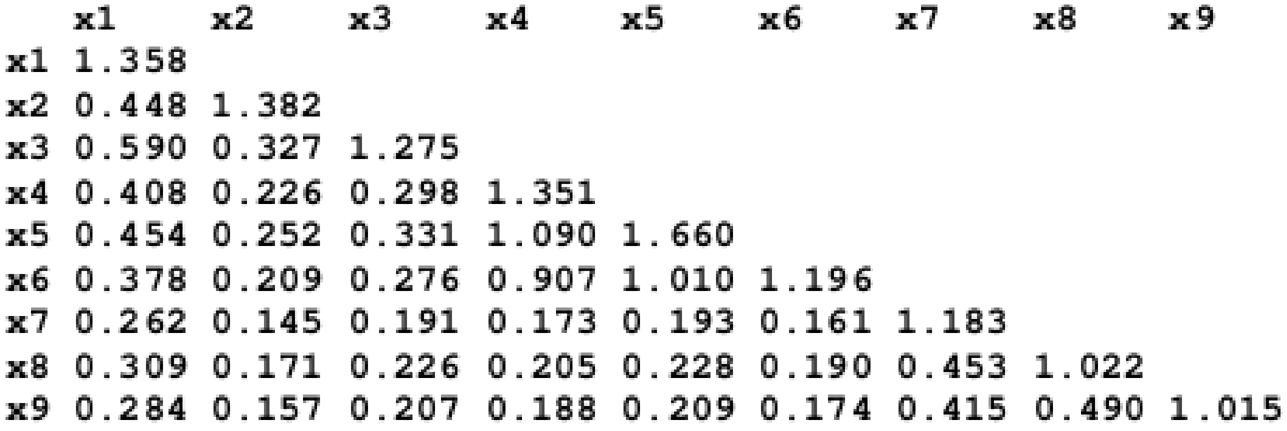
Slide by Yves Rosseel

observed covariance matrix



```
      x1     x2     x3     x4     x5     x6     x7     x8     x9
x1  1.358
x2  0.407  1.382
x3  0.580  0.451  1.275
x4  0.505  0.209  0.208  1.351
x5  0.441  0.211  0.112  1.098  1.660
x6  0.455  0.248  0.244  0.896  1.015  1.196
x7  0.085 -0.097  0.088  0.220  0.143  0.144  1.183
x8  0.264  0.110  0.212  0.126  0.181  0.165  0.535  1.022
x9  0.458  0.244  0.374  0.243  0.295  0.236  0.373  0.457  1.015
```

model-implied covariance matrix



```
      x1    x2    x3    x4    x5    x6    x7    x8    x9
x1 1.358
x2 0.448 1.382
x3 0.590 0.327 1.275
x4 0.408 0.226 0.298 1.351
x5 0.454 0.252 0.331 1.090 1.660
x6 0.378 0.209 0.276 0.907 1.010 1.196
x7 0.262 0.145 0.191 0.173 0.193 0.161 1.183
x8 0.309 0.171 0.226 0.205 0.228 0.190 0.453 1.022
x9 0.284 0.157 0.207 0.188 0.209 0.174 0.415 0.490 1.015
```

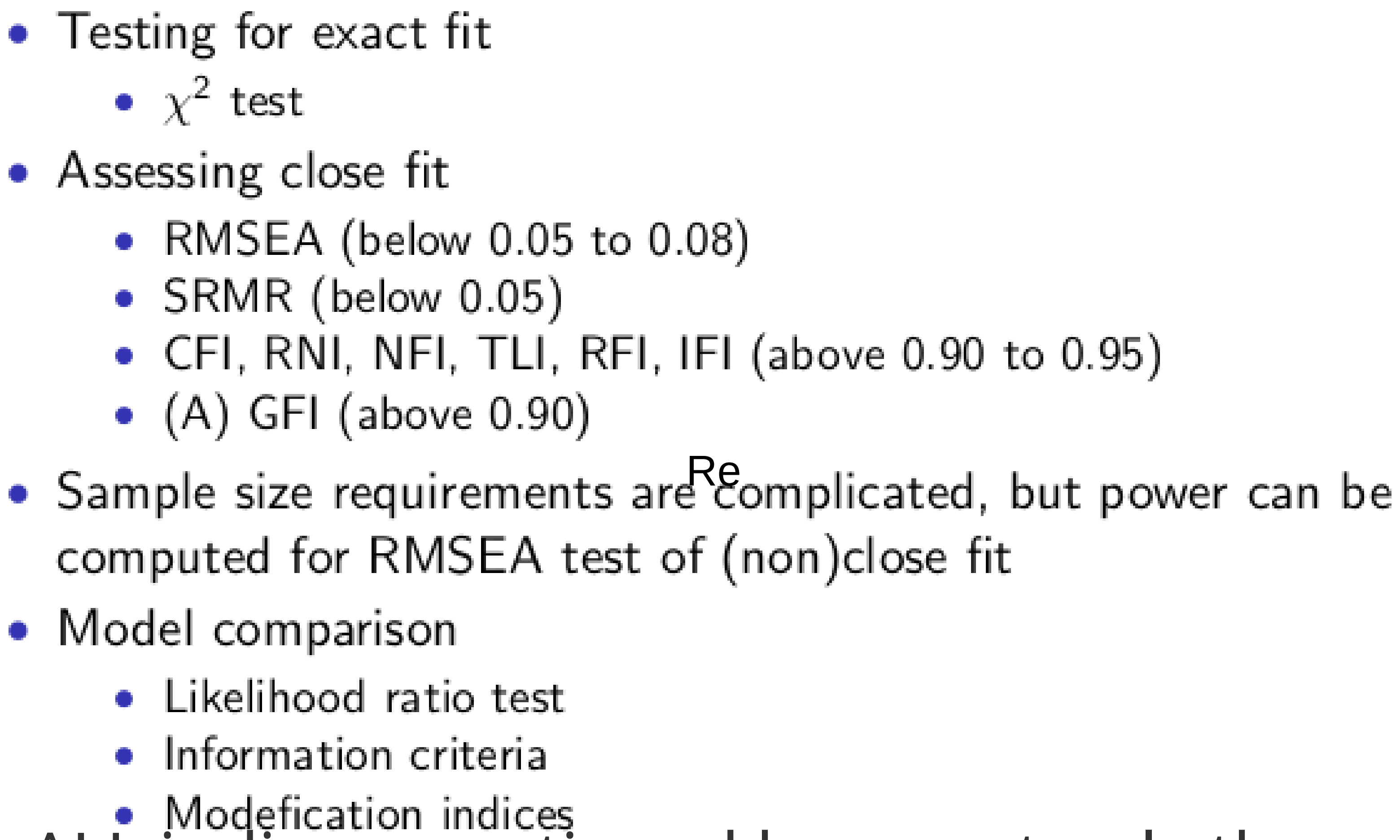
Slide by Yves Rosseel

What the practical covers

- Simple CFA example
- Simple regression example
- Evaluating model fit
- Modification indices
- Model comparison
- Plotting the model
- Question time for your own problems / plans / interests
- “Black box” approach due to lack of time

Model fit

Slide by Sacha Epskamp



- Testing for exact fit
  - χ^2 test
- Assessing close fit
  - RMSEA (below 0.05 to 0.08)
  - SRMR (below 0.05)
  - CFI, RNI, NFI, TLI, RFI, IFI (above 0.90 to 0.95)
  - (A) GFI (above 0.90)
- Sample size requirements are complicated, but power can be computed for RMSEA test of (non)close fit
- Model comparison
  - Likelihood ratio test
  - Information criteria
  - Modefication indices

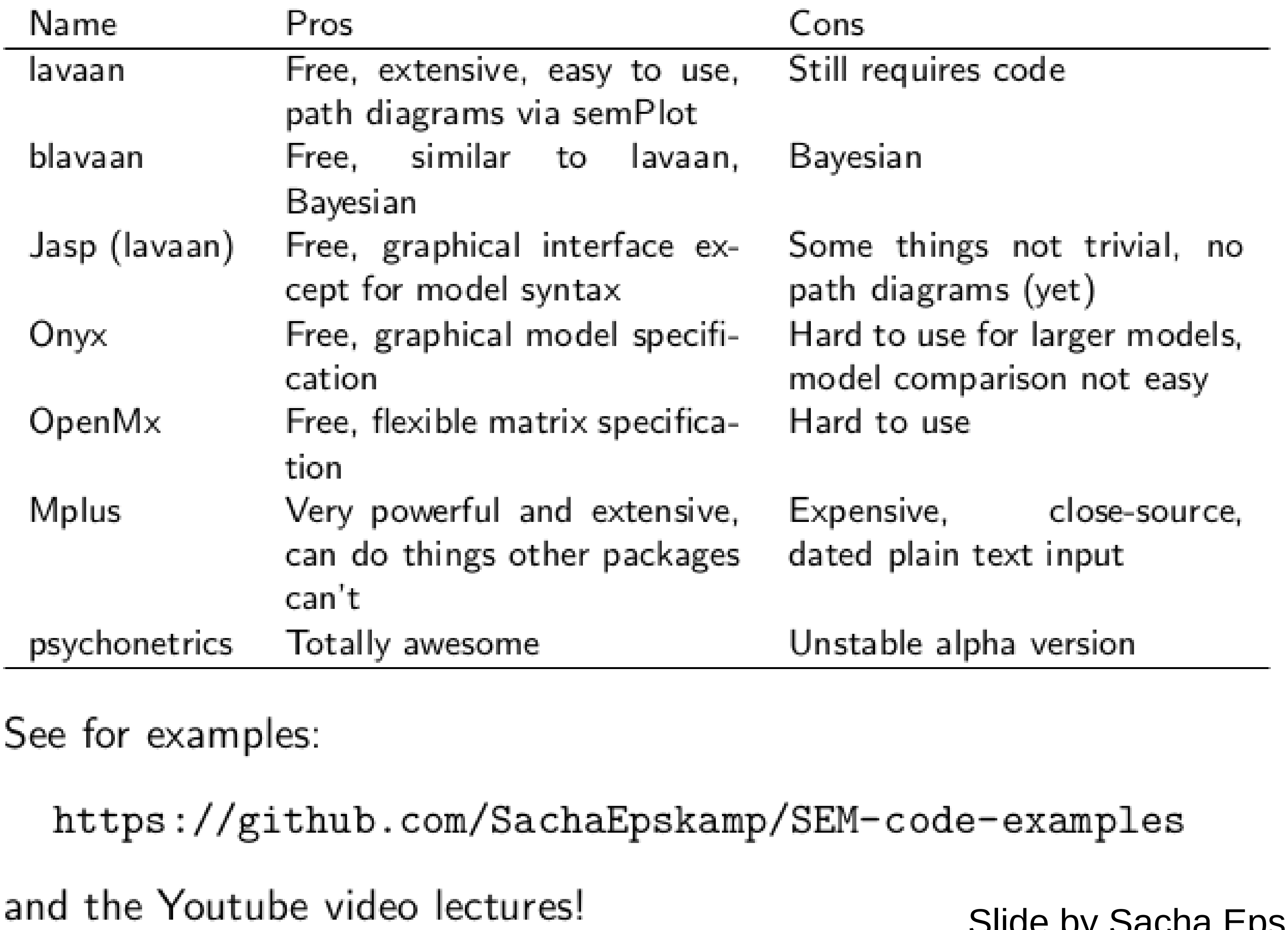
Re

- Report ALL indices mentioned here, not only the ones that “work”
- Good practice includes sharing data / covariance matrix / correlation matrix + mean +SD

What course/practical will lack

- Understanding the black box
- Intro to algebra
- What parameters are estimated and how they work
- Problem of alternative models
- Solution → Do the Epskamp course!

Software



Name	Pros	Cons
lavaan	Free, extensive, easy to use, path diagrams via semPlot	Still requires code
blavaan	Free, similar to lavaan, Bayesian	Bayesian
Jasp (lavaan)	Free, graphical interface except for model syntax	Some things not trivial, no path diagrams (yet)
Onyx	Free, graphical model specification	Hard to use for larger models, model comparison not easy
OpenMx	Free, flexible matrix specification	Hard to use
Mplus	Very powerful and extensive, can do things other packages can't	Expensive, close-source, dated plain text input
psychonetrics	Totally awesome	Unstable alpha version

See for examples:

```
https://github.com/SachaEpskamp/SEM-code-examples
```

and the Youtube video lectures!

Slide by Sacha Epskamp

A glimpse at the universe of SEM

- Hierarchical CFA models
  - e.g., bifactor models
- Measurement invariance
  - are models similar across genders, countries?
- Ordinal data
  - Questionnaire answers not continuous:
  - Agree | somewhat agree | maybe | somewhat disagree | disagree
  - Can be converted to 2-parameter IRT
- Nested data

SEM universe continues

- Complex path models
- Non-normal data
- Missing data
- Longitudinal data
- Recreating other statistical models (ANOVA, multi-level models) in SEM
  - But requires panel-like data
- Twin models (genetic models, causal inference)
- + any of the approaches can be combined with others

Summary

- Factor analysis – measurement model of a construct
- Path model – multiple regression equations in one model
- SEM = factor analysis + path models
- “All models are wrong, but some are useful” Box, 1976

Further resources

- Online courses
  - http://sachaepskamp.com/SEM2019
  - http://davidakenny.net/cm/fit.htm (free during covid)
  - http://www.personality-project.org/r/book/
- Lavaan web page, lavaan.ugent.be
  - Tutorials, books, code, help forum
- Power analysis
  - yilinandrewang.shinyapps.io/pwrSEM/
- Uku.vainik@ut.ee | Twitter: @ukuv

Prepping for the practical

- Install **R 3.6.3** NOT R 4.0.0!
  - Look for “Previous releases” or just 3.6.3 https://cran.r-project.org/
- Install Rstudio https://rstudio.com
- install.packages(c("lavaan","psych","tidyverse","semPlot","GPArotation","summarytools"),dependencies=T)
- Test loading them with
  - library(lavaan); library(psych)