Factor & Components Analysis II

Lecture 9

Multivariate statistics

Psychology 613 – Spring 2022

Differences between FA and CA

Components analysis	Factor analysis
Data driven	Hypothesis driven
Model free	*Model-based (SEM!)*
No latent variables	Latent and observed vars
Orthogonal or oblique	Oblique only
Arbitrary number of components	Number of factors specified in advance
No unique solution	Unique solution possible
Exploratory	Confirmatory

Selection primarily depends on the phase of your research

Factor analysis

Used when we want to test whether an underlying latent factor *causes* the indicators

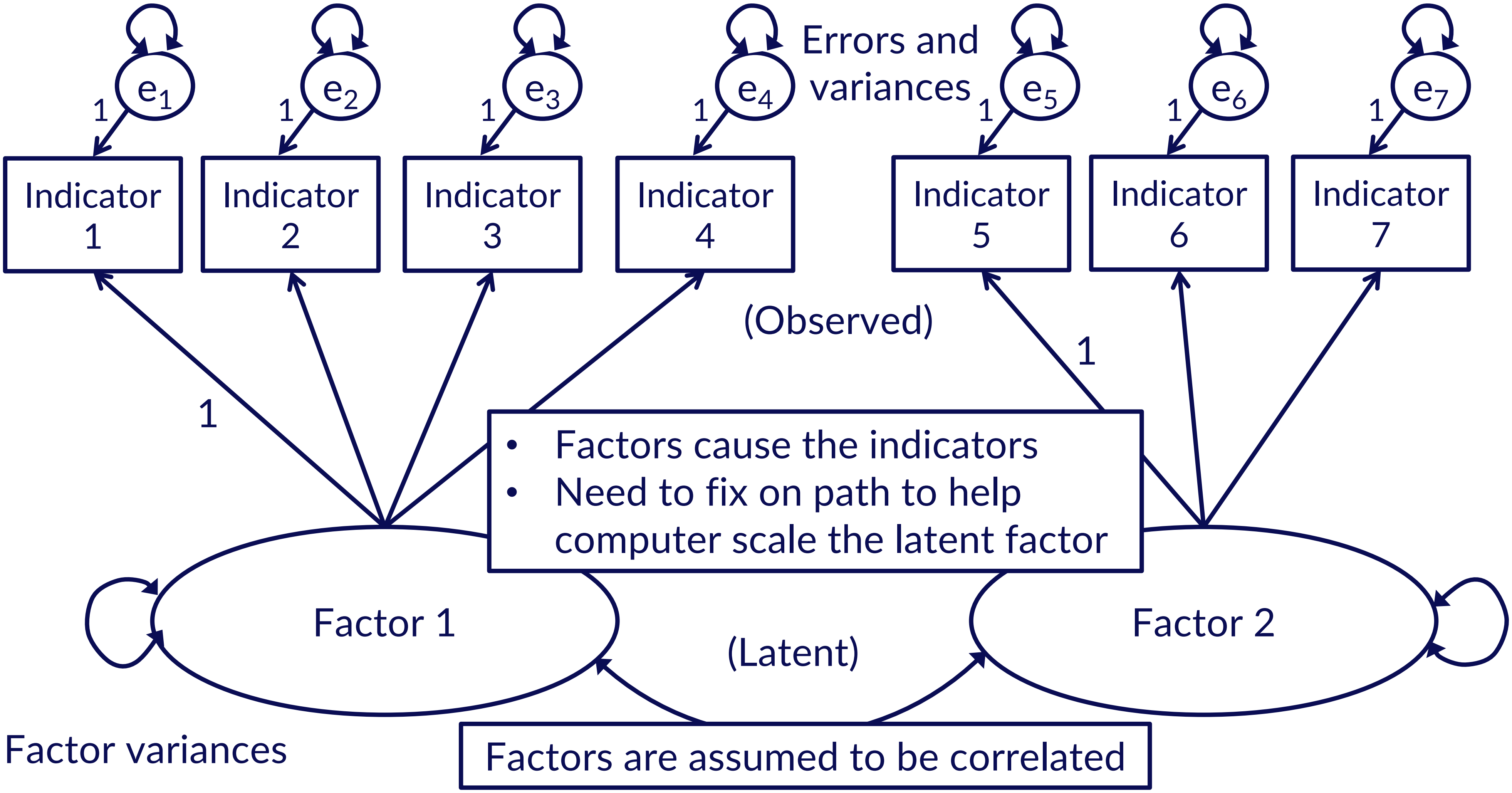
(In contrast to components analysis, where the indicators cause the component)

Error in measurement is explicitly included in model; only overlapping variance is of interest

(In contrast to CA, where all variance is analyzed simultaneously)

The factor analysis model

Based on structural equation modeling



- Factors cause the indicators
- Need to fix on path to help computer scale the latent factor

Factors are assumed to be correlated

Factor analysis: Output

Factor loadings + SEs for each path

(Standardized and unstandardized)

Error variances + SEs for each observed indicator

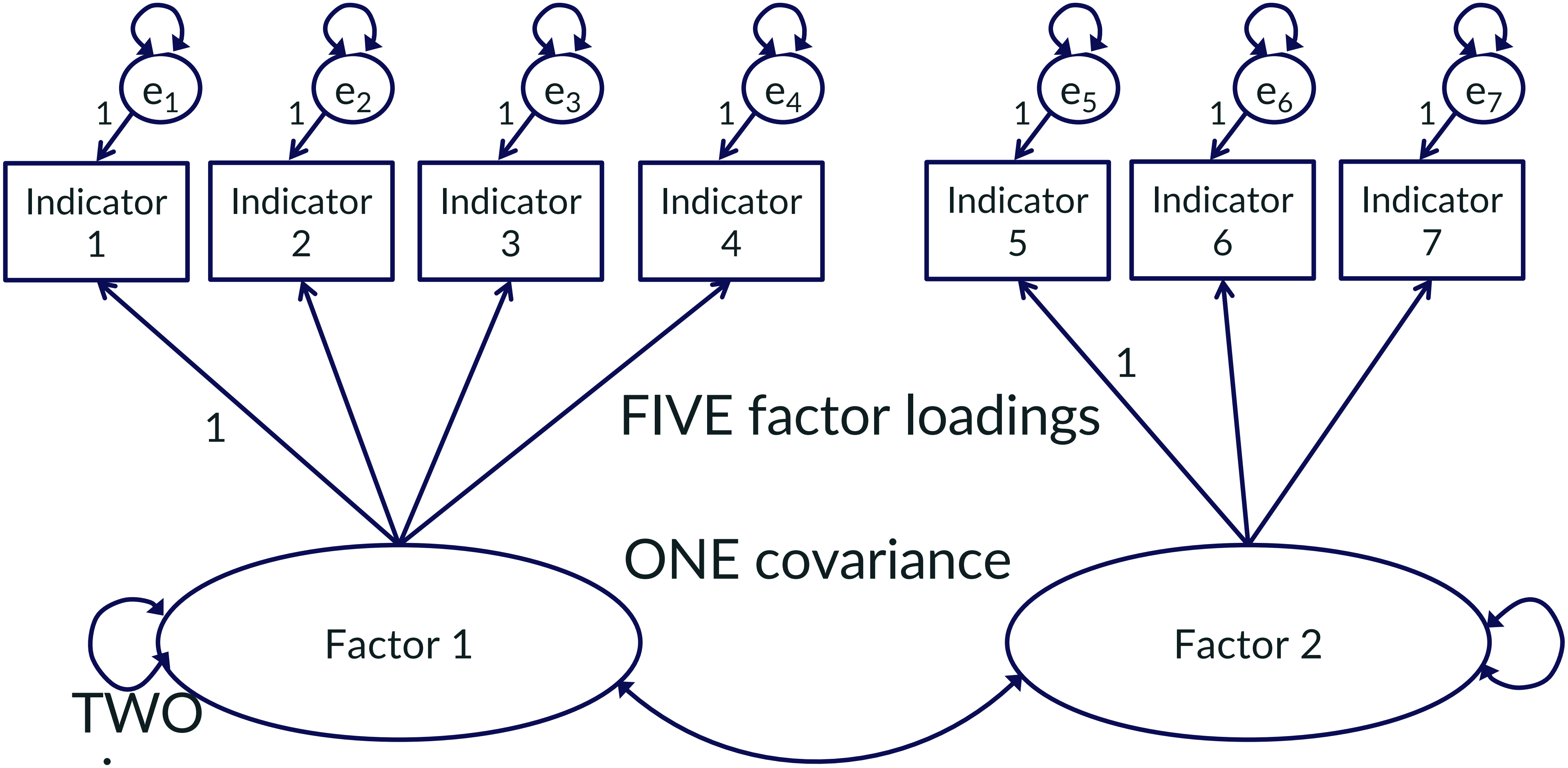
Factor variances and covariance + SEs

Model fit indices

How many parameters?

Estimating 15 numbers (each with SE) for this simple model!

SEVEN error variances



FIVE factor loadings

ONE covariance

TWO variances

Pattern vs. Structure coefficients

Pattern coefficients are the DIRECT paths between factors/indicators

- I.e., what you see on the diagram
- Controlling for any indirect paths

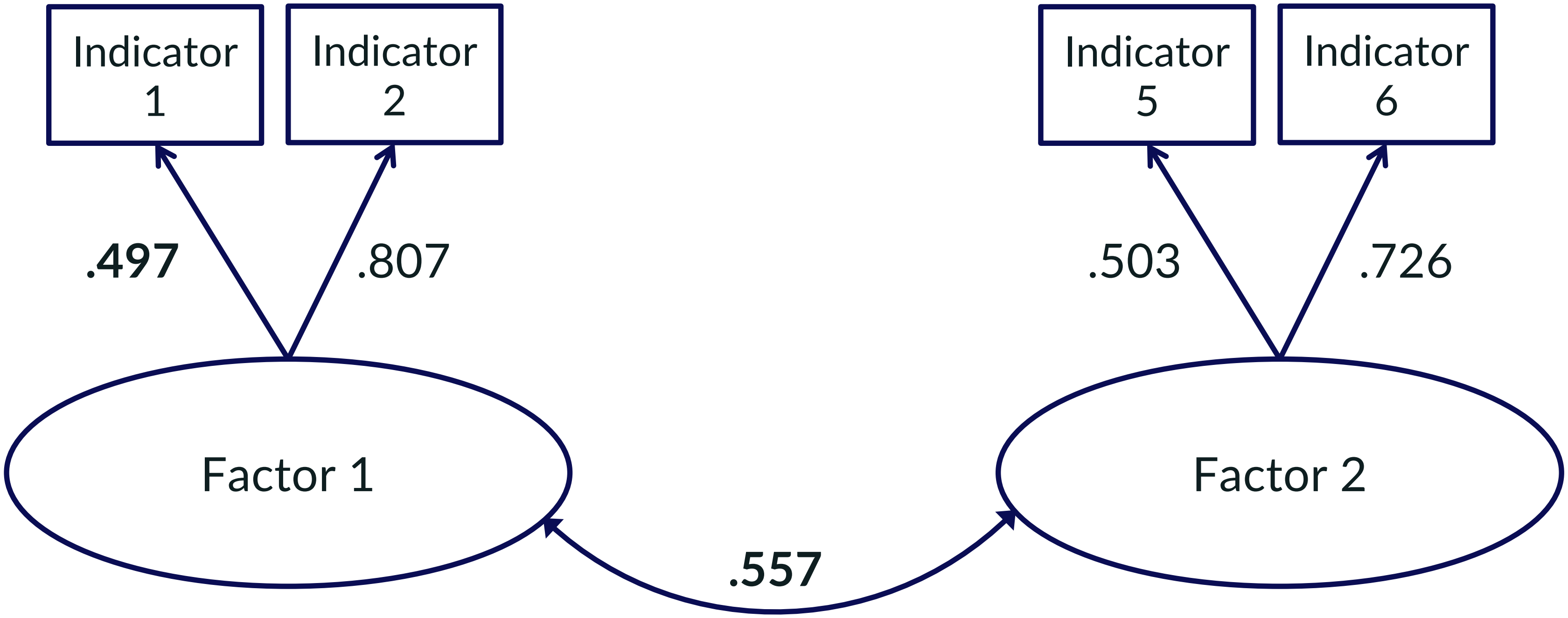
Structure coefficients are the TOTAL paths between factors/indicators

- I.e., correlations predicted by the model

Pattern vs. structure

Pattern correlation between Indicator 1 and Factor 2: Zero
Structure correlation between Indicator 1 and Factor 2: Not zero

Indicator 1 → Factor 1 → Factor 2: .497*.557 = 0.277
Structure correlation between Indicator 1 and Factor 2: 0.277



Problems reproducing?

The *residual correlation matrix* is calculated as:

Reproduced – original

(By convention, the reproduced correlation of each item with itself is the communality, the sum of squared loadings)

If the residual value is high for a given pair, it indicates possible correlation of errors between those items

What is in the error term?

Random error (assumed IID~N(0,1))

Measurement error (*un*reliability)

“Specific error” that is accounted for by other (unmodeled) factors

“Left-out-variable error”, or LOVE

What is in the error term?

If the model is *misspecified*, it may need additional variables/paths to improve fit

Alternatively, there may be a *measurement artifact* that accounts for the correlation between the variables (e.g., both items use some unusual word that causes the correlation)

Sample size

SEM requires large sample sizes, factor analysis less so

Follow the *N:q* rule

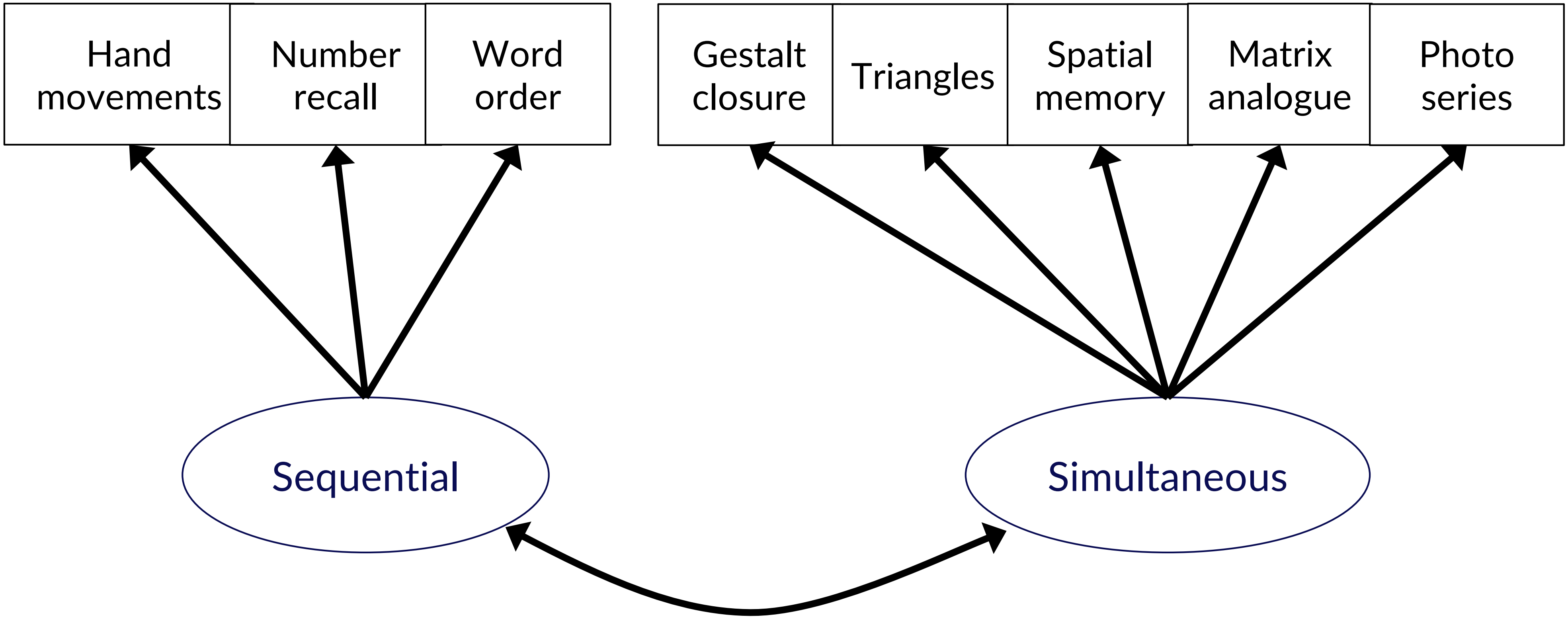
- Want a ratio of samples to parameters of at least 20:1

More needed for non-normality

Follow the KISS rule

An actual example

Eight observed indicators for two latent factors of intelligence



The Data

	HandMov	NumRec	WordOrd	Gestalt	Triangles	Spatial Mem	MatrixAn	PhotoSer
HandMov	11.56	3.1824	3.451	1.9278	2.9376	5.712	3.7128	3.978
NumRec	3.1824	5.76	4.6632	0.7128	1.7496	2.9232	2.1504	2.088
WordOrd	3.451	4.6632	8.41	1.2528	2.2707	3.4104	2.436	3.219
Gestalt	1.9278	0.7128	1.2528	7.29	2.7702	3.402	2.3436	3.402
Triangles	2.9376	1.7496	2.2707	2.7702	7.29	5.3298	3.1752	4.698
SpatialMem	5.712	2.9232	3.4104	3.402	5.3298	17.64	4.8216	6.426
MatrixAn	3.7128	2.1504	2.436	2.3436	3.1752	4.8216	7.84	3.528
PhotoSer	3.978	2.088	3.219	3.402	4.698	6.426	3.528	9

- This is a variance/covariance matrix: Variances along the diagonal, covariances elsewhere
- These are the raw data to pass to the software

Install the required packages

Here, I’m using the awesome SEM package “LAVAAN”

```
> require(‘lavaan’)
```

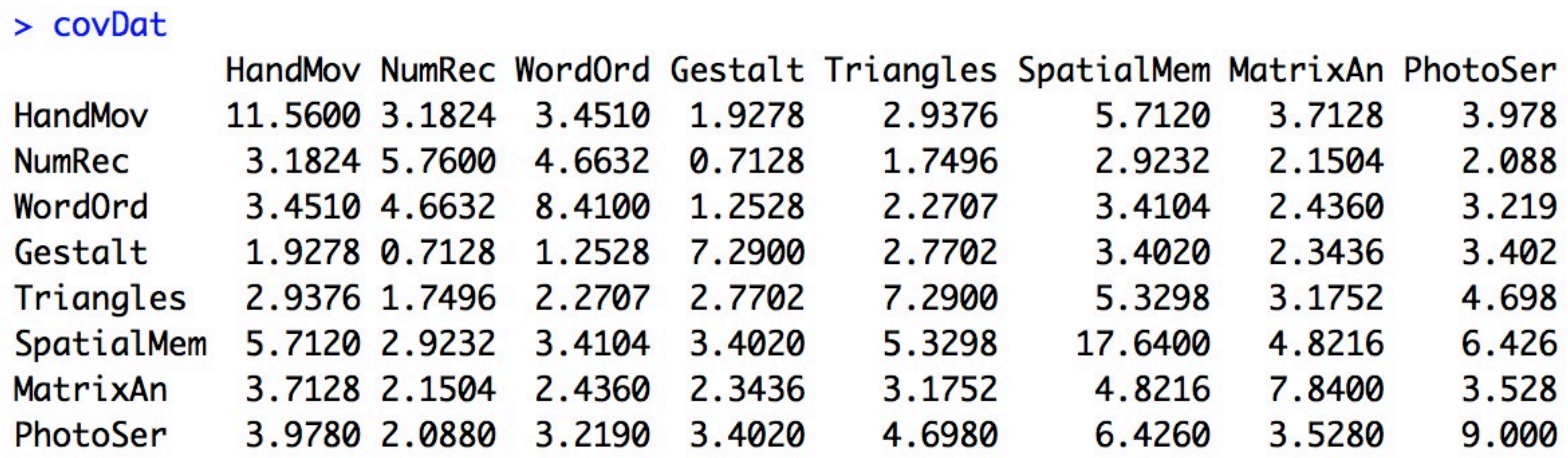
Also useful is “SEMPlot”

```
> require(‘semPlot’)
```

Read in the data

> covDat = read.table("Lecture9KaufmanCov.dat",
header=TRUE,row.names=1)

> covDat



```
> covDat
           HandMov NumRec WordOrd Gestalt Triangles SpatialMem MatrixAn PhotoSer
HandMov    11.5600 3.1824  3.4510  1.9278    2.9376     5.7120   3.7128    3.978
NumRec      3.1824 5.7600  4.6632  0.7128    1.7496     2.9232   2.1504    2.088
WordOrd     3.4510 4.6632  8.4100  1.2528    2.2707     3.4104   2.4360    3.219
Gestalt     1.9278 0.7128  1.2528  7.2900    2.7702     3.4020   2.3436    3.402
Triangles   2.9376 1.7496  2.2707  2.7702    7.2900     5.3298   3.1752    4.698
SpatialMem  5.7120 2.9232  3.4104  3.4020    5.3298    17.6400   4.8216    6.426
MatrixAn    3.7128 2.1504  2.4360  2.3436    3.1752     4.8216   7.8400    3.528
PhotoSer    3.9780 2.0880  3.2190  3.4020    4.6980     6.4260   3.5280    9.000
```

Specify the model (Lavaan)

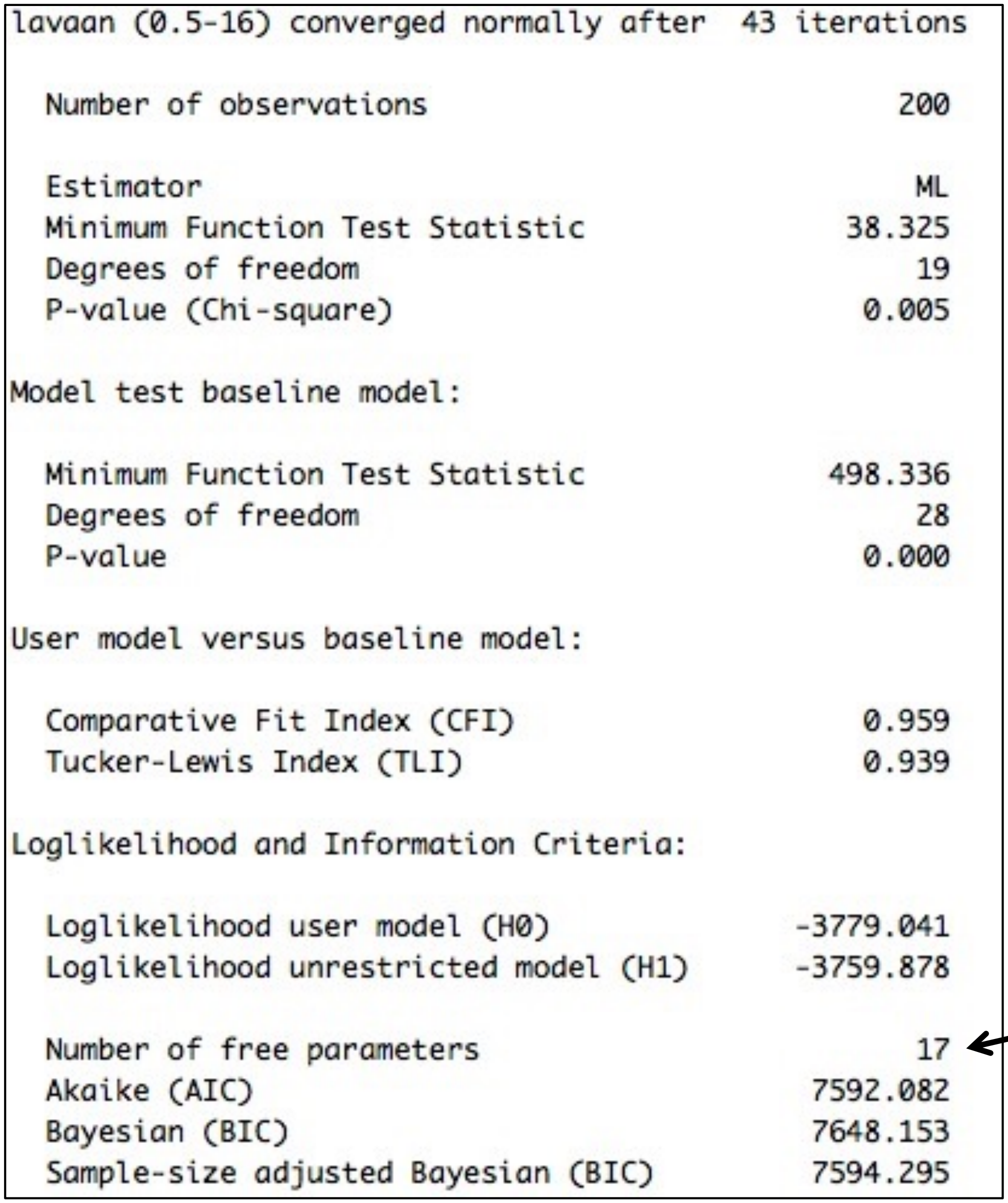
```
> model <- '
>  # Latent variables
>  [latent1] =~ [man1] + [man2] + ...
>  [latent2] =~ [man5] + [man6] + ...
>
>  # Covariances among latents
>  [latent1] ~~ [latent2] + [latent3] + ...
>  [latent2] ~~ [latent3] + ...
>  '
>
>  fit <- cfa(model, sample.cov=[cov object], sample.nobs = [N]
>  summary(fit, fit.measures = TRUE)
```

Specify the model.
“man” = manifest
= observed vars

Estimate it

Display it

Output from Lavaan

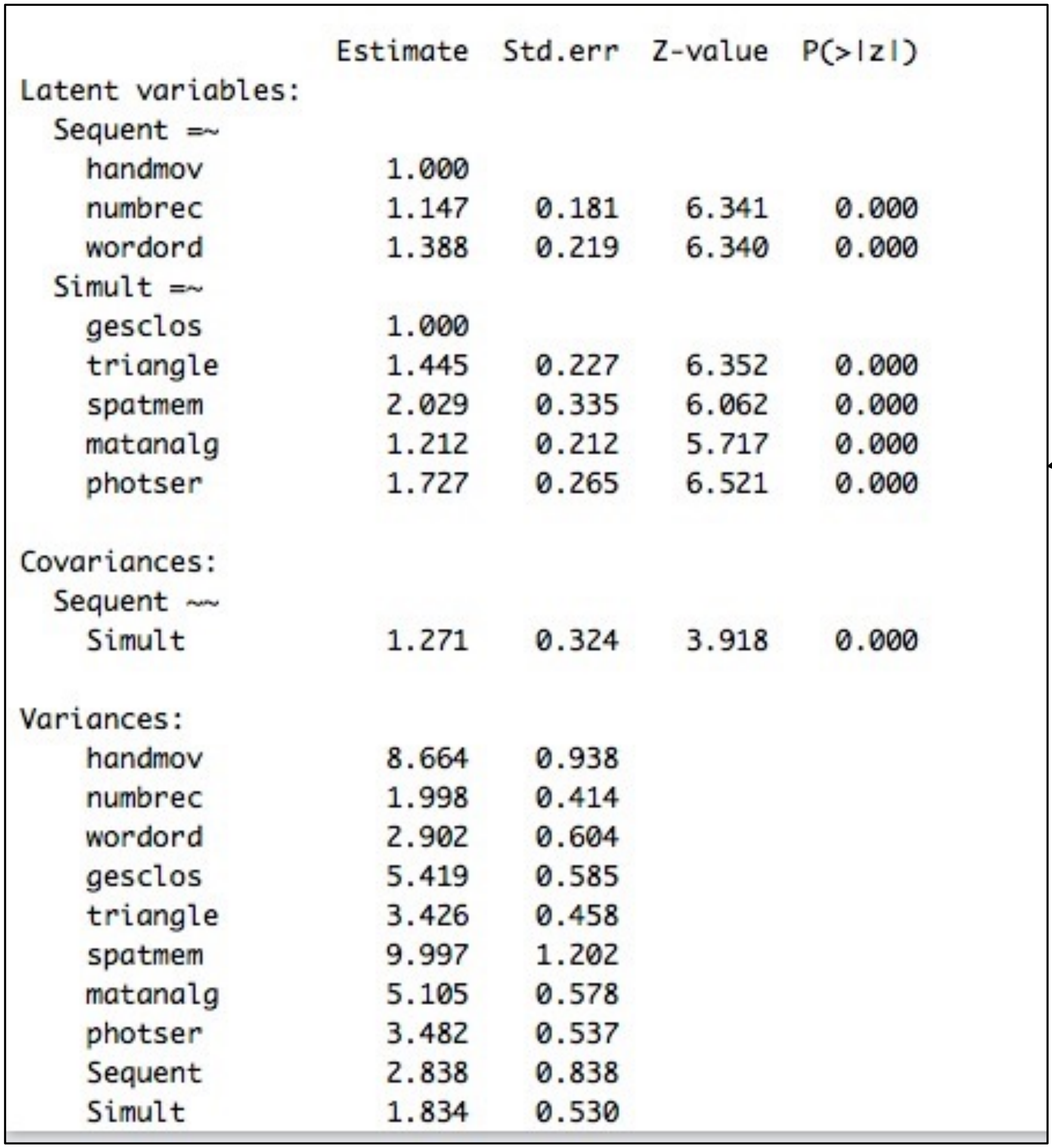


```
lavaan (0.5-16) converged normally after  43 iterations

  Number of observations                           200

  Estimator                                         ML
  Minimum Function Test Statistic               38.325
  Degrees of freedom                                19
  P-value (Chi-square)                           0.005

Model test baseline model:

  Minimum Function Test Statistic              498.336
  Degrees of freedom                                28
  P-value                                        0.000

User model versus baseline model:

  Comparative Fit Index (CFI)                    0.959
  Tucker-Lewis Index (TLI)                       0.939

Loglikelihood and Information Criteria:

  Loglikelihood user model (H0)              -3779.041
  Loglikelihood unrestricted model (H1)      -3759.878

  Number of free parameters                         17
  Akaike (AIC)                                7592.082
  Bayesian (BIC)                              7648.153
  Sample-size adjusted Bayesian (BIC)         7594.295
```

Various (overall) parameters and fit indices

Check that the correct number of model params appears here

Output from Lavaan

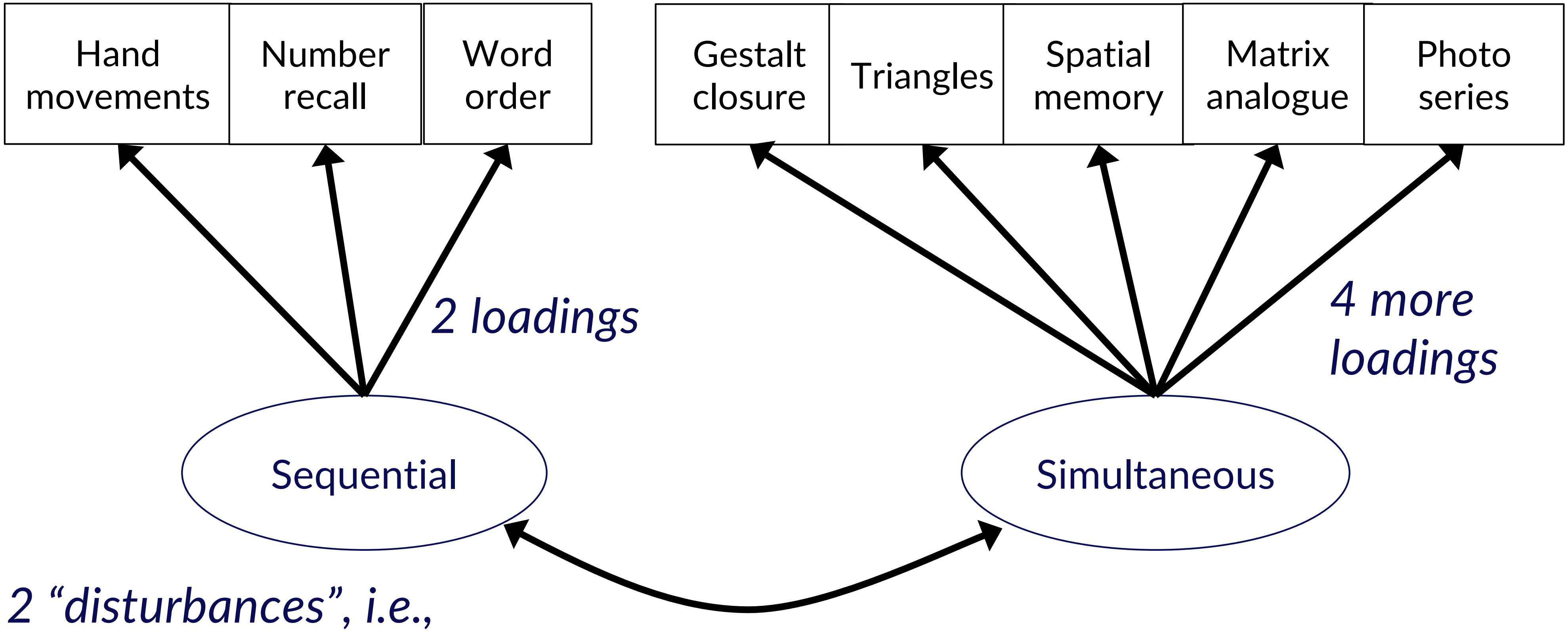


```
                   Estimate  Std.err  Z-value  P(>|z|)
Latent variables:
  Sequent =~
    handmov           1.000
    numbrec           1.147    0.181    6.341    0.000
    wordord           1.388    0.219    6.340    0.000
  Simult =~
    gesclos           1.000
    triangle          1.445    0.227    6.352    0.000
    spatmem           2.029    0.335    6.062    0.000
    matanalg          1.212    0.212    5.717    0.000
    photser           1.727    0.265    6.521    0.000

Covariances:
  Sequent ~~
    Simult            1.271    0.324    3.918    0.000

Variances:
    handmov           8.664    0.938
    numbrec           1.998    0.414
    wordord           2.902    0.604
    gesclos           5.419    0.585
    triangle          3.426    0.458
    spatmem           9.997    1.202
    matanalg          5.105    0.578
    photser           3.482    0.537
    Sequent           2.838    0.838
    Simult            1.834    0.530
```

Model parameters

Sanity check: How many parameters?

8 error variances



Hand movements

Number recall

Word order

Gestalt closure

Triangles

Spatial memory

Matrix analogue

Photo series

2 loadings

4 more loadings

Sequential

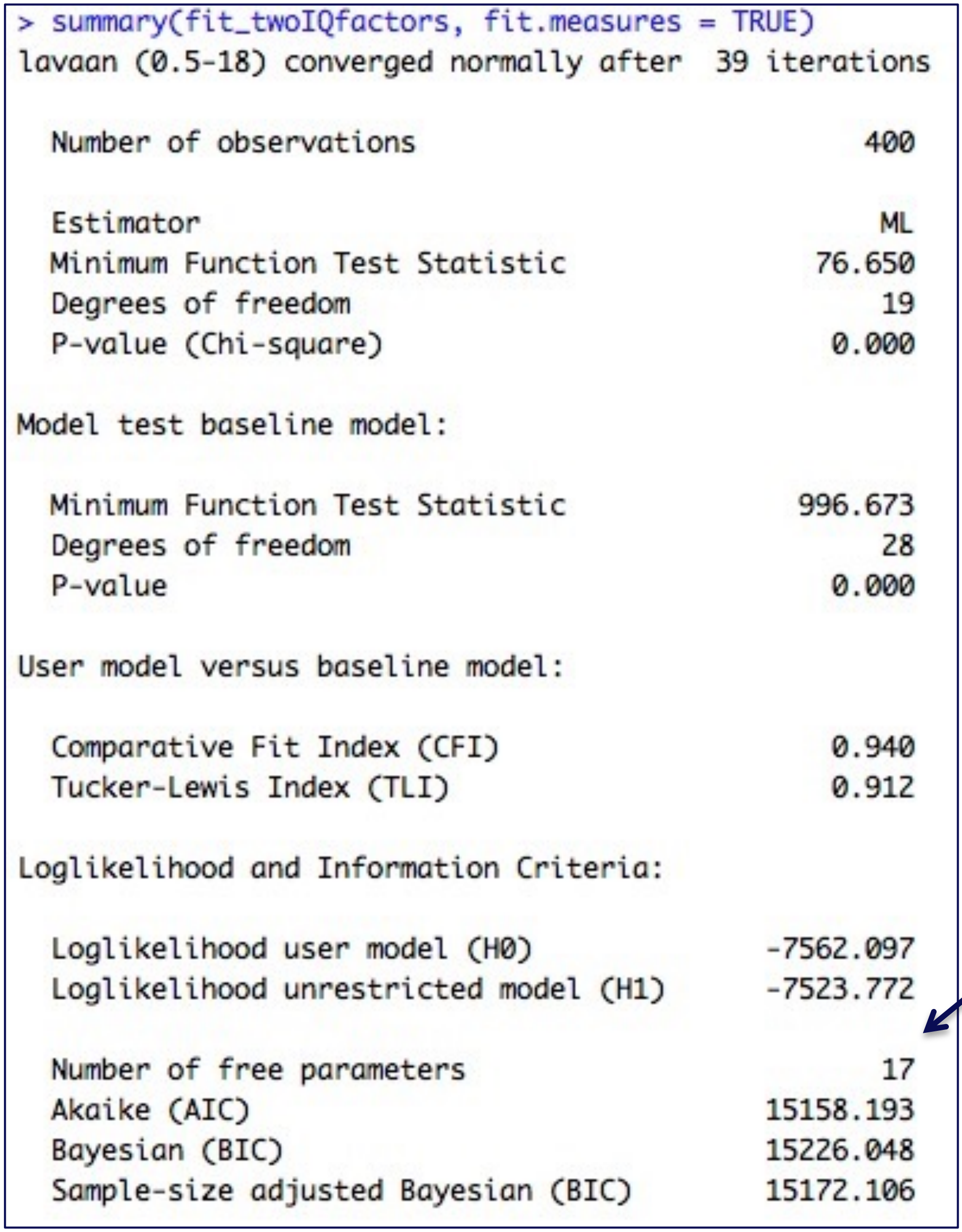
Simultaneous

2 “disturbances”, i.e., latent factor variances

1 covariance

17 parameters

Output

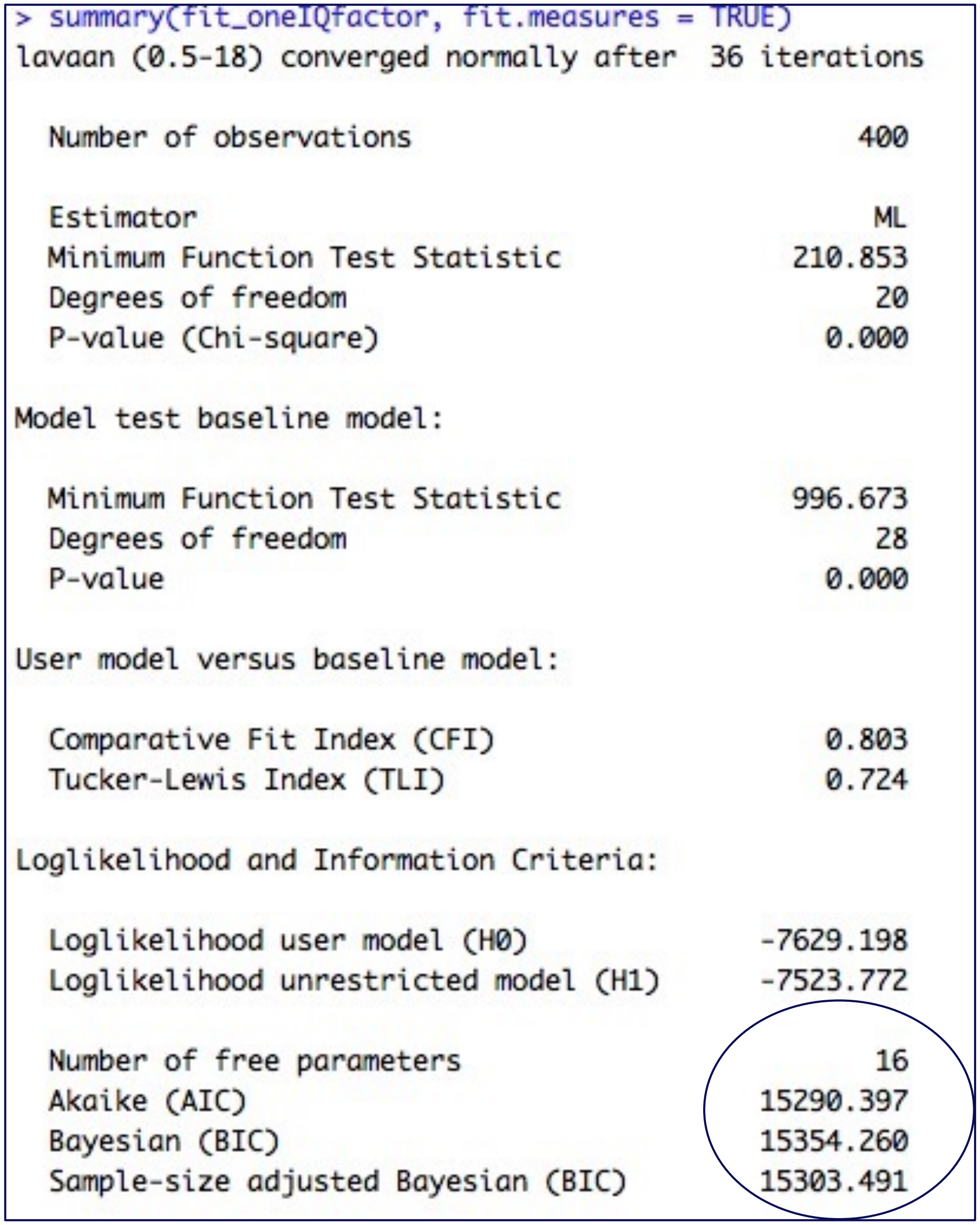


```
> summary(fit_twoIQfactors, fit.measures = TRUE)
lavaan (0.5-18) converged normally after  39 iterations

  Number of observations                           400

  Estimator                                         ML
  Minimum Function Test Statistic               76.650
  Degrees of freedom                                19
  P-value (Chi-square)                           0.000

Model test baseline model:

  Minimum Function Test Statistic              996.673
  Degrees of freedom                                28
  P-value                                        0.000

User model versus baseline model:

  Comparative Fit Index (CFI)                    0.940
  Tucker-Lewis Index (TLI)                       0.912

Loglikelihood and Information Criteria:

  Loglikelihood user model (H0)              -7562.097
  Loglikelihood unrestricted model (H1)      -7523.772

  Number of free parameters                         17
  Akaike (AIC)                               15158.193
  Bayesian (BIC)                             15226.048
  Sample-size adjusted Bayesian (BIC)        15172.106
```

- Model fit statistics including dfs, deviance, and some others…

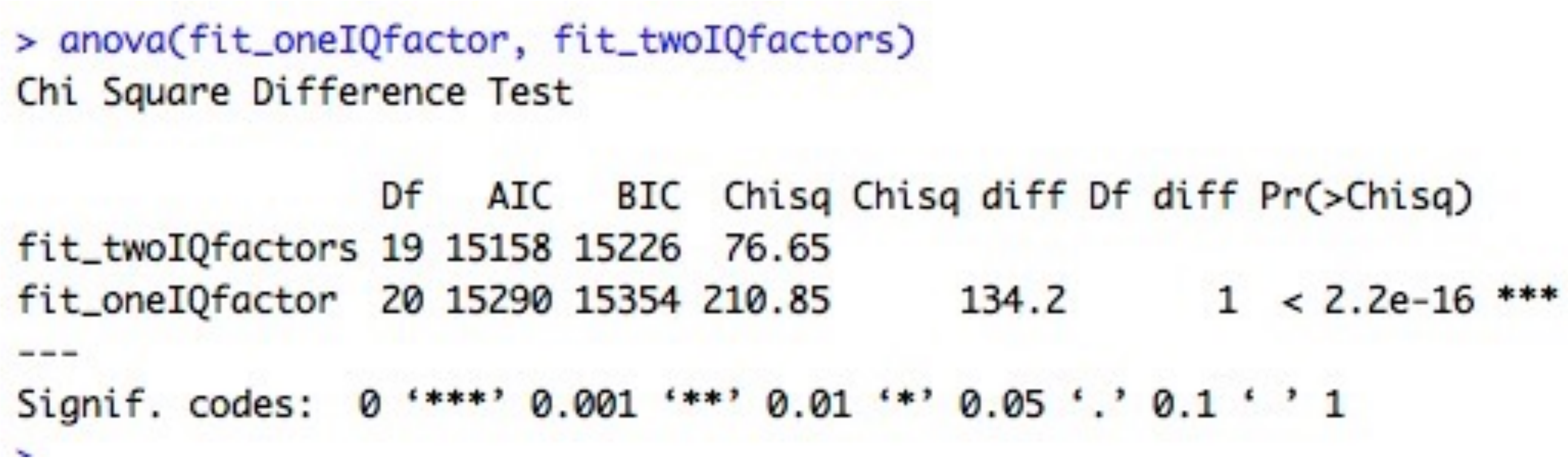
Compare nested models based on chi-squared change test using these two numbers. They are *deviances*

Comparison model: One factor



```
> summary(fit_oneIQfactor, fit.measures = TRUE)
lavaan (0.5-18) converged normally after  36 iterations

  Number of observations                           400

  Estimator                                         ML
  Minimum Function Test Statistic              210.853
  Degrees of freedom                                20
  P-value (Chi-square)                           0.000

Model test baseline model:

  Minimum Function Test Statistic              996.673
  Degrees of freedom                                28
  P-value                                        0.000

User model versus baseline model:

  Comparative Fit Index (CFI)                    0.803
  Tucker-Lewis Index (TLI)                       0.724

Loglikelihood and Information Criteria:

  Loglikelihood user model (H0)              -7629.198
  Loglikelihood unrestricted model (H1)      -7523.772

  Number of free parameters                         16
  Akaike (AIC)                               15290.397
  Bayesian (BIC)                             15354.260
  Sample-size adjusted Bayesian (BIC)        15303.491
```

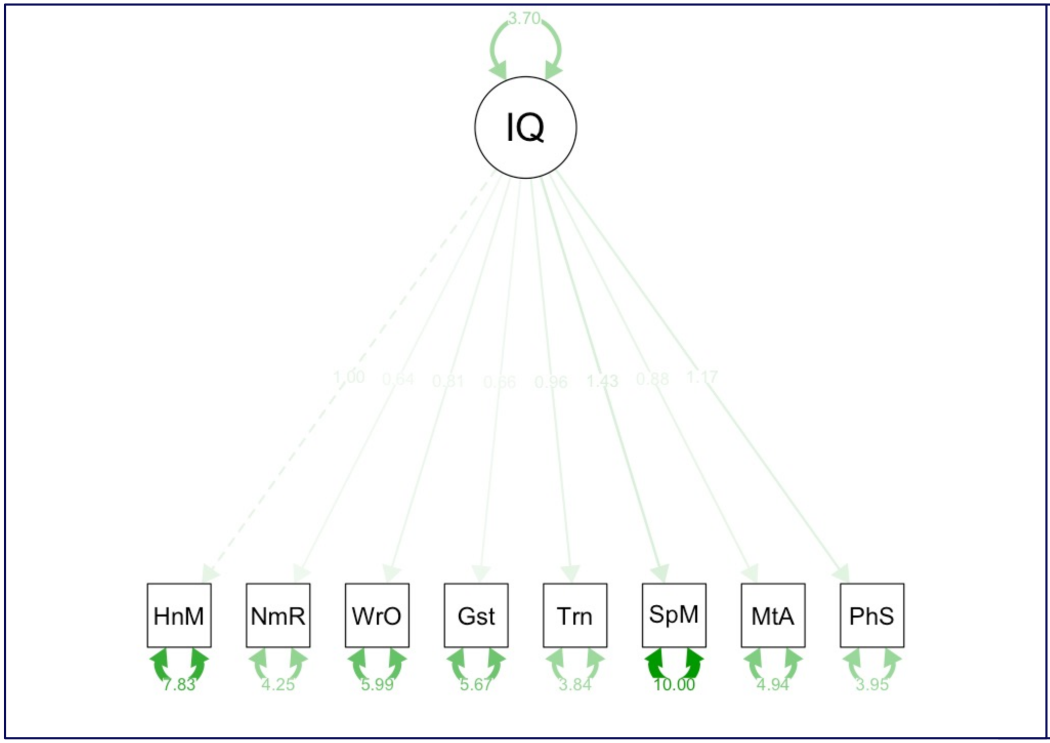
Model comparison test

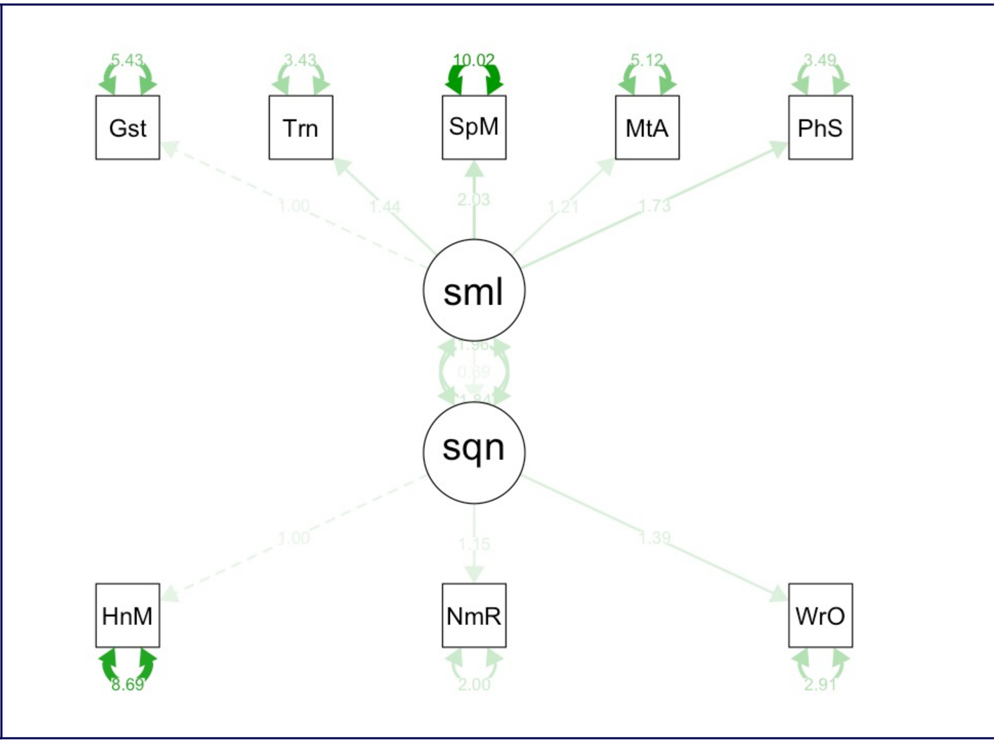


```
> anova(fit_oneIQfactor, fit_twoIQfactors)
Chi Square Difference Test

                 Df   AIC   BIC  Chisq Chisq diff Df diff Pr(>Chisq)    
fit_twoIQfactors 19 15158 15226  76.65                                  
fit_oneIQfactor  20 15290 15354 210.85      134.2       1  < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
```

Plots

> semPaths(fit_twoIQfactors, what = “est”)


3.70
IQ
1.00
0.64
0.81
0.66
0.96
1.43
0.88
1.17
HnM
NmR
WrO
Gst
Trn
SpM
MtA
PhS
7.83
4.25
5.99
5.67
3.84
10.00
4.94
3.95


5.43
3.43
10.02
5.12
3.49
Gst
Trn
SpM
MtA
PhS
1.00
1.44
2.03
1.21
1.73
sml
1.96
0.69
1.84
sqn
1.00
1.15
1.39
HnM
NmR
WrO
8.69
2.00
2.91