Introduction to Structural Equation Models

Kimmo Vehkalahti

Adjunct Professor, University Lecturer
University of Helsinki, Department of Social Research, Statistics
http://www.helsinki.fi/people/Kimmo.Vehkalahti

HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Analyzing and Interpreting Data – Annual Seminar of FiDPEL:
The Finnish Doctoral Programme in Education and Learning
15 May 2013

Outline

- **Models, Modeling and Model-fitting**
- **Basic Concepts and Ideas of SEM**
- **Analyzing and Interpreting Data**

Models, Modeling and Model-fitting

Models are simplifications of real-world phenomena.

Models are grounded on some **substantial theory**.

Statistical modeling:

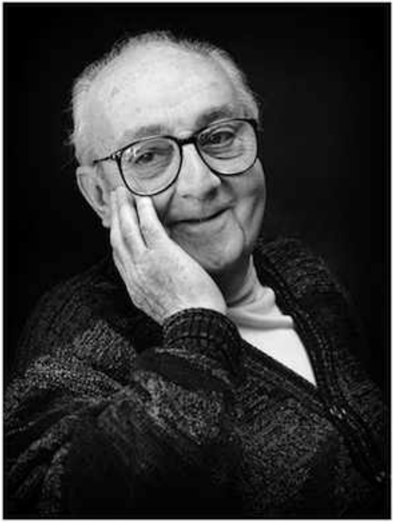
- primary task is to determine the **goodness-of-fit** between the hypothesized **model** and the sample **data**
- process of model-fitting can be summarized as

  Data = Model + Residual, where

  Data represent measurements of the observed variables,
  Model represents the hypothesized structure, and
  Residual represents the discrepancy between them
- **ultimate objective:** to find a model that is both
  1) substantive meaningful and 2) statistically well fitting

Models, Modeling and Model-fitting

However, remember the general truth:



**"All models are wrong,
but some are useful."**

George E. P. Box (1919–2013)

Professor of Statistics
University of Wisconsin–Madison

Models – what about the Structural Equations?

Equations are used for defining any **details of models.**

Example of a simple equation (linear regression model):

$$y = \beta_0 + \beta_1 x + \varepsilon,$$

where we are interested in the relationship of y and x:

- y is the dependent variable (in the data),
- x is the independent variable (in the data),
- β_0 and β_1 are the **regression coefficients**, and
- ε represents the random variation.

Structural equations are used for defining **causal processes:**

- much more **complicated** structures than the one above
- **simultaneous** analysis of an entire system of variables
- can be modeled **pictorially:** clearer conceptualization
- **hypothesized** model can be tested **statistically**
- **but:** causality is not a statistical concept: **theory required!**

Basic Concepts and Ideas of SEM

HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Latent variables and observed variables

Latent variables

- latent: not observed directly, cannot be measured directly
- operationally defined in terms of the believed behavior
- indirectly measured via observed variables

Observed variables

- responses to scales, scores in tests, coded responses etc.
- indicators of the **underlying construct** they represent
- choice of psychometrically sound measurement instruments crucial for assessing the underlying constructs

Basic Concepts and Ideas of SEM

(Sort of) combination of factor analysis and regression analysis

Factor analysis

- links the observed variables with the latent variables (factors)
- primary interest: the strength of the regression paths from the factors to the observed variables (i.e. the *factor loadings*)
- this part of SEM is called a **measurement model**
- exploratory vs confirmatory factor analysis (EFA, CFA)

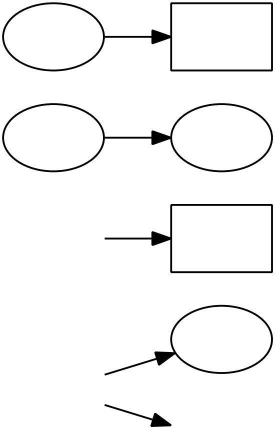
The full latent variable (SEM) model

- regression structure among the latent variables (factors)
- hypothesized causal impacts of one factor on another
- measurement (CFA) model and a **structural model**

Generalizations: growth curve models, multilevel models, non-linear models, time-series models, longitudinal models etc.

Basic Concepts and Ideas of SEM

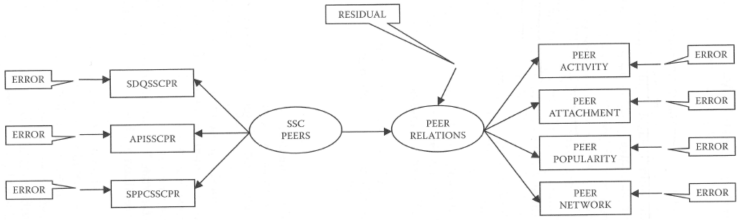
Typical configurations with simple graphical symbols:



- path coefficient for regression from a factor to an observed variable
- path coefficient for regression of one factor onto another
- **measurement error** associated with an observed variable
- **residual error** in the prediction of an unobserved variable

Models are drawn using various combinations of these symbols.

Schematic example of a SEM (Byrne 2012)

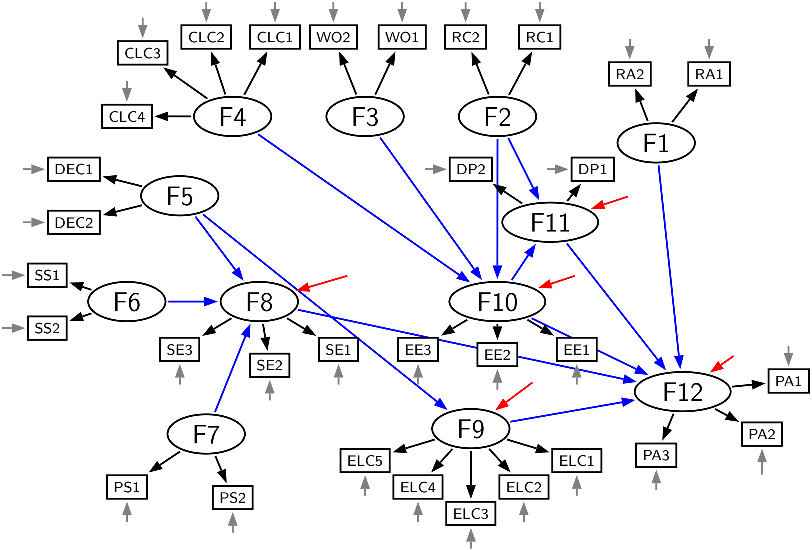


- two factors: (SSC = social self-concept, PR = peer relations)
- PR: dependent, SSC: independent, causal relation
- seven observed variables with measurement errors
- residual of predicting PR from SSC

The corresponding eight **structural equations** *(follow the paths!)*:

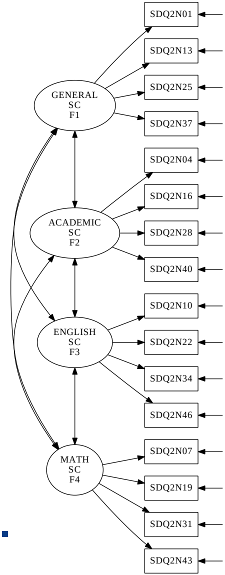
- PR = SSC + residual
- SDQSSCPR = SSC + error
- APISSCPR = SSC + error
- SPPCSCPR = SSC + error
- peer activity = PR + error
- peer attachment = PR + error
- peer popularity = PR + error
- peer network = PR + error

More complicated example of SEM (Byrne 2012)



Analyzing and Interpreting Data

Let us now focus on Confirmatory Factor Analysis (Byrne 2012).



- **Hypothesis:** self-concept (SC) is a multidimensional construct composed of four factors:
  - General SC (GSC)
  - Academic SC (ASC)
  - English SC (ESC)
  - Mathematics SC (MSC)
- **Data:** 16 variables from grade 7 children (N=265)

The aim here is simply:
to show some typical steps of the modeling process from a statistician's viewpoint.

Analyzing and Interpreting Data

Assessment of the model

- primary interest: the extent to which a model fits the data
- several criteria for a) model as a whole, b) parameter estimates

Model fitting

- Null hypothesis H_0: the postulated model holds in the population
- We hope *not* to reject H_0 (cf. traditional procedures).

Statistical significance

- As always, merely a starting point, not a result as such!
- **Practical** significance is important: values of estimates, effect sizes, confidence intervals etc.

Hypothesized four-factor CFA model (Mplus output)


```
SUMMARY OF ANALYSIS

Number of groups                                                 1
Number of observations                                         265

Number of dependent variables                                   16
Number of independent variables                                  0
Number of continuous latent variables                            4

Observed dependent variables

  Continuous
   SDQ2N01     SDQ2N13     SDQ2N25     SDQ2N37     SDQ2N04     SDQ2N16
   SDQ2N28     SDQ2N40     SDQ2N10     SDQ2N22     SDQ2N34     SDQ2N46
   SDQ2N07     SDQ2N19     SDQ2N31     SDQ2N43

Continuous latent variables
   F1          F2          F3          F4


Estimator                                                       ML
Information matrix                                        OBSERVED
Maximum number of iterations                                  1000
Convergence criterion                                    0.500D-04
Maximum number of steepest descent iterations                   20
```

Hypothesized four-factor CFA model (Mplus output)


```
TESTS OF MODEL FIT

Chi-Square Test of Model Fit
          Value                              159.112
          Degrees of Freedom                      98
          P-Value                             0.0001

Chi-Square Test of Model Fit for the Baseline Model
          Value                             1703.155
          Degrees of Freedom                     120
          P-Value                             0.0000

CFI/TLI
          CFI                                  0.961
          TLI                                  0.953

Loglikelihood
          H0 Value                         -6562.678
          H1 Value                         -6483.122

Information Criteria
          Number of Free Parameters               54
          Akaike (AIC)                     13233.356
          Bayesian (BIC)                   13426.661
          Sample-Size Adjusted BIC         13255.453
            (n* = (n + 2) / 24)

RMSEA (Root Mean Square Error Of Approximation)
          Estimate                             0.049
          90 Percent C.I.                      0.034  0.062
          Probability RMSEA <= .05             0.556

SRMR (Standardized Root Mean Square Residual)
          Value                                0.045
```

Analyzing and Interpreting Data

The goodness-of-fit statistics

Chi-Square Test of Model Fit

- traditional Likelihood Ratio Test statistic, expressed as a chi-square (χ^2) statistic
- *p*-value represents the likelihood of obtaining a χ^2 value that exceeds the χ^2 value when H_0 is true, so the higher the *p*, the closer the fit.
- always reported but rarely used as the sole index of model fit

Here, $\chi^2 = 159.112$, with 98 degrees of freedom (df) and a *p*-value of less than 0.0001 suggests that the fit of the data to the model is not adequate and H_0 should be rejected.

Analyzing and Interpreting Data

Alternative (subjective) indices of fit: more pragmatic approach, e.g. Comparative Fit Index (CFI), Tucker–Lewis Fit Index (TLI), Akaike's Information Criterion (AIC), Bayes Information Criterion (BIC).

First two represent the most typical incremental indices measuring the proportionate improvement in fit with *nested* models. The latter two are predictive or parsimony-corrected criteria for *non-nested* models.

A careful consideration of these is essential when fitting models. Use of multiple indices highly recommended. Here, only four, perhaps the most typical ones. **Warning:** model may fit well, and still be incorrectly specified! These are merely giving information on the model's lack of fit. **Researcher must know if the model is plausible or not.**

Analyzing and Interpreting Data

The assessment of model adequacy must be based on multiple criteria that take into account
a) **theoretical**, b) **statistical**, and c) **practical** considerations.

- CFI (Comparative Fit Index):
  - compares the hypothesized (H) and the baseline model (B)
  - range: $[0, 1]$, well-fitting models have CFI > 0.95
- TLI (Tucker–Lewis Fit Index):
  - quite similar as CFI, but nonnormed (may extend $[0, 1]$)
  - includes a penalty function for overly complex models (H)
  - well-fitting models have TLI close to 1

Analyzing and Interpreting Data

The last two fit indices are RMSEA and SRMR, both absolute ones, sometimes termed "absolute misfit indices". They do not compare models, but depend only on determining how well the model fits the data. Therefore these decrease as the fit improves.

Values of RMSEA less than 0.05 indicate good fit, 0.08 to 0.10 mediocre/reasonable fit, and greater than 0.10 poor fit.

RMSEA: Root Mean Square Error Of Approximation
SRMR: Standardized Root Mean Square Residual

Analyzing and Interpreting Data

RMSEA is sensitive to the number of estimated parameters (model complexity). Routine use of RMSEA is strongly recommended in literature. Here: 0.049 (with a 90 % C.I. [0.034, 0.062]), indicating a good precision.

SRMR represents the average residual value of the fit (standardized, i.e., range $[0, 1]$). In well-fitting models, it will be small, less than 0.05. Here: 0.045, that is, model explains the correlations to within an average error of 0.045.

RMSEA: Root Mean Square Error Of Approximation
SRMR: Standardized Root Mean Square Residual

Analyzing and Interpreting Data

Assessment of parameter estimates

1) appropriateness, 2) statistical significance

- Parameter estimates should exhibit the correct sign and size, and be **consistent with the underlying theory.**
- Any estimates falling outside the admissible range indicate that a) model is wrong or b) input matrix lacks sufficient information.
- Standard errors of parameters should not be excessively large or small.
- Assuming that the sample size is adequate, nonsignificant parameters (except the error variances) should be deleted from the model.

Hypothesized four-factor CFA model (Mplus output)

```
MODEL RESULTS
                                                         Two-Tailed
                      Estimate       S.E.  Est./S.E.    P-Value
                  (factor loadings)
 F1       BY
    SDQ2N01     fixed: 1.000      0.000    999.000    999.000
    SDQ2N13            1.083      0.156      6.939      0.000
    SDQ2N25            0.851      0.125      6.806      0.000
    SDQ2N37            0.934      0.141      6.607      0.000

 F2       BY
    SDQ2N04            1.000      0.000    999.000    999.000
    SDQ2N16            1.279      0.151      8.489      0.000
    SDQ2N28            1.247      0.155      8.022      0.000
    SDQ2N40            1.259      0.159      7.913      0.000

 F3       BY
    SDQ2N10            1.000      0.000    999.000    999.000
    SDQ2N22            0.889      0.104      8.561      0.000
    SDQ2N34            0.670      0.148      4.541      0.000
    SDQ2N46            0.843      0.118      7.160      0.000

 F4       BY
    SDQ2N07            1.000      0.000    999.000    999.000
    SDQ2N19            0.841      0.058     14.447      0.000
    SDQ2N31            0.952      0.048     19.905      0.000
    SDQ2N43            0.655      0.050     13.182      0.000
                  (factor covariances)
 F2       WITH
    F1                 0.415      0.078      5.325      0.000

 F3       WITH
    F1                 0.355      0.072      4.928      0.000
    F2                 0.464      0.080      5.825      0.000

 F4       WITH
    F1                 0.635      0.117      5.437      0.000
    F2                 0.873      0.134      6.507      0.000
    F3                 0.331      0.100      3.309      0.001
```

Hypothesized four-factor CFA model (Mplus output)

```
Intercepts
   SDQ2N01          4.408      0.083     53.312      0.000
   SDQ2N13          5.004      0.083     60.087      0.000
   SDQ2N25          5.098      0.075     67.766      0.000
   SDQ2N37          4.826      0.070     68.758      0.000
   SDQ2N04          4.521      0.086     52.630      0.000
   SDQ2N16          4.649      0.076     61.116      0.000
   SDQ2N28          4.691      0.082     57.419      0.000
   SDQ2N40          4.977      0.083     59.717      0.000
   SDQ2N10          4.623      0.071     65.454      0.000
   SDQ2N22          5.377      0.067     80.384      0.000
   SDQ2N34          3.891      0.104     37.256      0.000
   SDQ2N46          5.268      0.080     66.253      0.000
   SDQ2N07          4.321      0.109     39.560      0.000
   SDQ2N19          4.543      0.104     43.738      0.000
   SDQ2N31          4.740      0.096     49.225      0.000
   SDQ2N43          4.977      0.086     57.961      0.000

Variances
   F1               0.613      0.141      4.342      0.000
   F2               0.561      0.126      4.449      0.000
   F3               0.668      0.116      5.744      0.000
   F4               2.307      0.272      8.483      0.000

Residual Variances
   SDQ2N01          1.198      0.130      9.228      0.000
   SDQ2N13          1.119      0.124      9.000      0.000
   SDQ2N25          1.056      0.109      9.675      0.000
   SDQ2N37          0.771      0.089      8.621      0.000
   SDQ2N04          1.394      0.128     10.890      0.000
   SDQ2N16          0.616      0.070      8.856      0.000
   SDQ2N28          0.896      0.092      9.739      0.000
   SDQ2N40          0.952      0.095     10.061      0.000
   SDQ2N10          0.653      0.082      7.926      0.000
   SDQ2N22          0.657      0.076      8.703      0.000
   SDQ2N34          2.590      0.233     11.093      0.000
   SDQ2N46          1.201      0.118     10.164      0.000
   SDQ2N07          0.854      0.098      8.729      0.000
   SDQ2N19          1.228      0.125      9.808      0.000
   SDQ2N31          0.365      0.065      5.581      0.000
   SDQ2N43          0.964      0.093     10.410      0.000
```

Analyzing and Interpreting Data

Standardized estimates

Mplus offers three types of standardization: STDYX, STDY and STD. These correspond to various ways of conceptualizing standardization (there is no one right choice!). In this respect, SEM programs vary. Hence: check, compare and verify that particular parameter values are consistent with the literature, if you want to replicate some known results! (Here, I follow Byrne and consider only the STDYX option).

Two aspects of standardized values (compared with the unstandardized solution):

- parameters reported earlier as 1.0 have now new values
- factor variances are now reported as 1.0 (no matter which STD option is used)

Hypothesized four-factor CFA model (Mplus output)

```
STANDARDIZED MODEL RESULTS - STDYX Standardization
                                                    Two-Tailed
                    Estimate       S.E.  Est./S.E.    P-Value

 F1       BY
    SDQ2N01            0.582      0.055     10.613      0.000
    SDQ2N13            0.626      0.050     12.426      0.000
    SDQ2N25            0.544      0.056      9.644      0.000
    SDQ2N37            0.640      0.051     12.608      0.000

 F2       BY
    SDQ2N04            0.536      0.048     11.143      0.000
    SDQ2N16            0.774      0.031     24.983      0.000
    SDQ2N28            0.703      0.037     19.071      0.000
    SDQ2N40            0.695      0.036     19.043      0.000

 F3       BY
    SDQ2N10            0.711      0.044     16.167      0.000
    SDQ2N22            0.668      0.046     14.447      0.000
    SDQ2N34            0.322      0.064      5.002      0.000
    SDQ2N46            0.532      0.054      9.873      0.000

 F4       BY
    SDQ2N07            0.854      0.020     41.953      0.000
    SDQ2N19            0.755      0.030     24.825      0.000
    SDQ2N31            0.923      0.016     59.292      0.000
    SDQ2N43            0.712      0.033     21.287      0.000

 F2       WITH
    F1                 0.707      0.057     12.421      0.000

 F3       WITH
    F1                 0.555      0.072      7.733      0.000
    F2                 0.758      0.052     14.652      0.000

 F4       WITH
    F1                 0.534      0.061      8.756      0.000
    F2                 0.767      0.038     20.291      0.000
    F3                 0.266      0.073      3.652      0.000
```

Hypothesized four-factor CFA model (Mplus output)


```
Intercepts
   SDQ2N01            3.275      0.155     21.135      0.000
   SDQ2N13            3.691      0.172     21.498      0.000
   SDQ2N25            4.163      0.191     21.798      0.000
   SDQ2N37            4.224      0.193     21.831      0.000
   SDQ2N04            3.233      0.153     21.092      0.000
   SDQ2N16            3.754      0.174     21.544      0.000
   SDQ2N28            3.527      0.165     21.368      0.000
   SDQ2N40            3.668      0.171     21.481      0.000
   SDQ2N10            4.021      0.185     21.718      0.000
   SDQ2N22            4.938      0.223     22.132      0.000
   SDQ2N34            2.289      0.117     19.584      0.000
   SDQ2N46            4.070      0.187     21.746      0.000
   SDQ2N07            2.430      0.122     19.898      0.000
   SDQ2N19            2.687      0.132     20.372      0.000
   SDQ2N31            3.024      0.145     20.854      0.000
   SDQ2N43            3.561      0.166     21.396      0.000

Variances
   F1          fixed: 1.000      0.000    999.000    999.000
   F2                 1.000      0.000    999.000    999.000
   F3                 1.000      0.000    999.000    999.000
   F4                 1.000      0.000    999.000    999.000

Residual Variances
   SDQ2N01            0.661      0.064     10.366      0.000
   SDQ2N13            0.609      0.063      9.665      0.000
   SDQ2N25            0.704      0.061     11.467      0.000
   SDQ2N37            0.591      0.065      9.096      0.000
   SDQ2N04            0.713      0.051     13.847      0.000
   SDQ2N16            0.402      0.048      8.381      0.000
   SDQ2N28            0.506      0.052      9.784      0.000
   SDQ2N40            0.517      0.051     10.195      0.000
   SDQ2N10            0.494      0.063      7.905      0.000
   SDQ2N22            0.554      0.062      8.987      0.000
   SDQ2N34            0.896      0.042     21.583      0.000
   SDQ2N46            0.717      0.057     12.493      0.000
   SDQ2N07            0.270      0.035      7.763      0.000
   SDQ2N19            0.429      0.046      9.341      0.000
   SDQ2N31            0.148      0.029      5.170      0.000
   SDQ2N43            0.493      0.048     10.351      0.000
```

Analyzing and Interpreting Data

R^2 values

When standardized estimates are requested, R^2 values for the dependent variables are reported.

- R^2 is the proportion of variance accounted for by its related factor (*communality* in factor analytic terms).
- SDQ2N34 is the weakest indicator ($R^2 = 1 - 0.896 = 0.104$).

R-SQUARE

Observed Variable	Estimate	S.E.	Est./S.E.	Two-Tailed P-Value
SDQ2N01	0.339	0.064	5.307	0.000
SDQ2N13	0.391	0.063	6.213	0.000
SDQ2N25	0.296	0.061	4.822	0.000
SDQ2N37	0.409	0.065	6.304	0.000
SDQ2N04	0.287	0.051	5.572	0.000
SDQ2N16	0.598	0.048	12.492	0.000
SDQ2N28	0.494	0.052	9.535	0.000
SDQ2N40	0.483	0.051	9.522	0.000
SDQ2N10	0.506	0.063	8.083	0.000
SDQ2N22	0.446	0.062	7.224	0.000
SDQ2N34	**0.104**	0.042	2.501	0.012
SDQ2N46	0.283	0.057	4.936	0.000
SDQ2N07	0.730	0.035	20.976	0.000
SDQ2N19	0.571	0.046	12.412	0.000
SDQ2N31	0.852	0.029	29.646	0.000
SDQ2N43	0.507	0.048	10.643	0.000

Analyzing and Interpreting Data

Model misspecification and Modification Indices (MI)

The function of MIs is to identify badly chosen parameter constraints (e.g. those fixed to a value of 0.00).

MIs are used to help answering to these questions:

- "What if a parameter would be freely estimated?"
- "How much would χ^2 value of the model decrease?"
- "Would the drop be significant?"
- "Would it lead to a better fitting model?"

A clue is given by the corresponding EPC (Expected Parameter Change) values. **Again substantive knowledge and reasoning is required.**

Analyzing and Interpreting Data

Here, we have four suggestions: one factor loading and three *residual* covariances:

```
MODEL MODIFICATION INDICES
                                  M.I.     E.P.C.  Std E.P.C.  StdYX E.P.C.
BY Statements

F2        BY SDQ2N07        11.251    -0.563     -0.422       -0.237

WITH Statements

SDQ2N25  WITH SDQ2N01      17.054     0.359      0.359        0.319
SDQ2N31  WITH SDQ2N07      10.696     0.305      0.305        0.546
SDQ2N31  WITH SDQ2N19      17.819    -0.331     -0.331       -0.495
```

The factor loading (F2 BY SDQ2N07) represents a **cross-loading** that could be added. The problem is that from a **substantive** perspective, the EPC value has a wrong sign! (The relation should be positive, not negative.) Thus it would be questionable to free this parameter for estimation, no matter the MI and EPC.

Analyzing and Interpreting Data

Overall: the residual covariances seem to be rather small and not worthy of inclusion in a subsequently specified model. Remember the topic of scientific parsimony: avoid too many parameters and hence too complex models! **First of all: model must be substantively meaningful.**

In general, model respecification is commonly conducted in SEM. It is important to realize that when we move to these *Post Hoc* analyses, they will then be framed within *exploratory*, not anymore confirmatory modeling approach! Combination of 1) substantive and 2) statistical aspects is required **(always in this order!)**.

"When to stop fitting a model?" – a good question...