Confirmatory Factor Analysis

Theory Construction and Statistical Modeling

Utrecht University

Kyle M. Lang

Department of Methodology & Statistics
Utrecht University

Outline

South African Personality Inventory Project

SAPI©

Carin Hill
Leon Jackson
Deon Meiring
J. Aleweyn Nel

Ian Rothmann
Michael Temane
Velichko H. Valchev
Fons J. R. van de Vijver

Nel, J. A., Valchev, V. H., Rothmann, S., van de Vijver, F. J. R., Meiring, D., & de Bruin, G. P. (2012). Exploring the personality structure in the 11 languages of South Africa. *Journal of Personality*, *80*, 915–948.

SAPI details

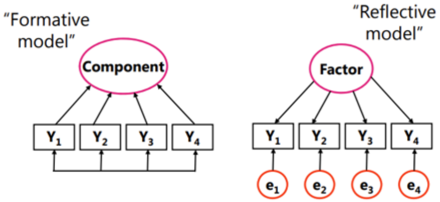
- 1216 participants from 11 official language groups
- From about 50,000 descriptive responses to 262 personality items
- Nine personality clusters:
  - Conscientiousness
  - Emotional Stability
  - Extraversion
  - Facilitating
  - Integrity
  - Intellect
  - Openness
  - Relationship Harmony
  - Soft-Heartedness (Ubuntu)
- Our data: selection of 1000 participants

Factor Analysis

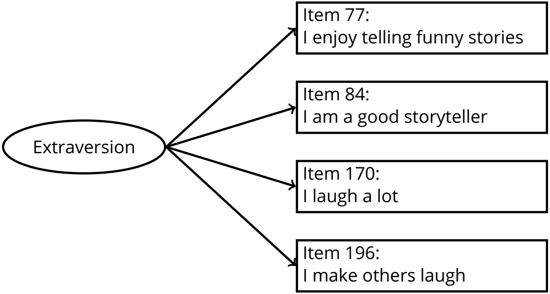
Factor Analysis: Modeling measurement of a latent variable

- EFA: Exploratory Factor Analysis.
- CFA: Confirmatory Factor Analysis.

Both EFA and CFA use a ”reflective” measurement model, not a ”formative” model.

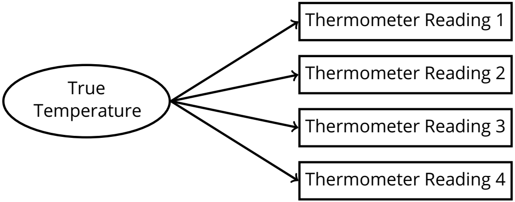


Reflective Constructs



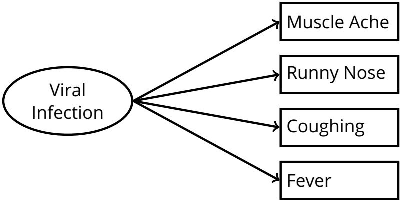
- Items are dependent variables, caused by the factor!
- Latent variable ‘extraversion’ explains item correlations: The factor is the reason for the covariances/correlations.

Reflective Constructs



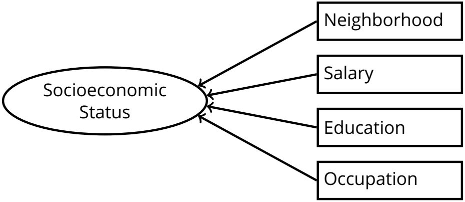
Thermometer readings are the dependent variables, caused by the temperature!

Reflective Constructs



Symptoms are the dependent variables, caused by the viral infection!

Formative Constructs



SES is an *index* defined as a (weighted) sum of the observed items.

- SES is the (latent) dependent variable, predicted by the items.
- This model is not empirically testable.

Interesting read

Interesting read on theory & latent variables:

Borsboom, D., Mellenbergh, G.J., & Van Heerden, J. (2003). The theoretical status of latent variables. *Psychological review, 110*(2), 203.

Confirmatory or Exploratory?

Two Subscales of Extraversion

Having Fun

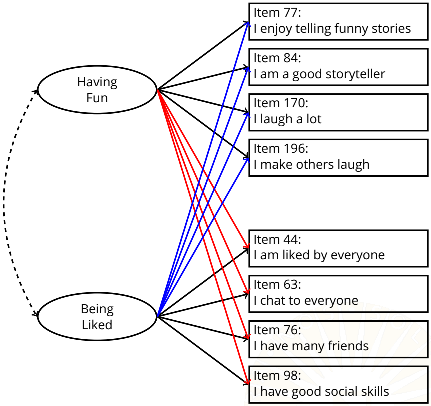
- Item 77: I enjoy telling funny stories
- Item 84: I am a good storyteller
- Item 170: I laugh a lot
- Item 196: I make others laugh

Being Liked

- Item 44: I am liked by everyone
- Item 63: I chat to everyone
- Item 76: I have many friends
- Item 98: I have good social skills

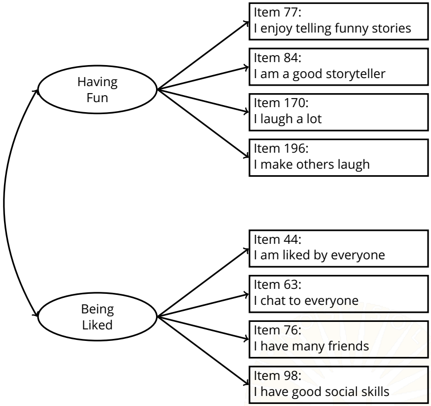
EFA

- All items load onto all factors
- No hypothesized measurement model
- Estimating latent covariances is optional
    - Oblique factors → Estimated
    - Orthogonal factors → Fixed
- Solution is not unique
- Use rotation to improve interpretability



CFA

- The statistical model represents the hypothesized measurement model
- No cross-loadings unless they're predicted by theory
- Almost always estimate the latent covariances
- A unique solution exists



CFA IN R

SAPI CFA in R

Load the SAPI data.

```
dataDir <- "../data/"
sapi <- read.table(paste0(dataDir, "sapi.txt"), header = TRUE, na.strings = "-999
```

Specify the lavaan model syntax.

```
mod1 <- '
fun   =~ Q77 + Q84 + Q170 + Q196
liked =~ Q44 + Q63 + Q76  + Q98
'
```

Use the `cfa()` function to estimate the model.

```
out1 <- cfa(mod1, data = sapi)
```

SAPI CFA in R

Visualize the fitted model.

```
library(lavaanPlot)
lavaanPlot(model = out1,
           node_options = list(shape = "box",
                                fontname = "Helvetica"),
           edge_options = list(color = "grey"),
           coefs = TRUE,
           stand = TRUE,
           covs = TRUE)
```

SAPI CFA in R

```
Error in path.expand(path):  invalid 'path' argument
```

Step 9: Acquiring the summary

```
summary(out1)

parameterEstimates(out1)

fitMeasures(out1, c("chisq", "df", "pvalue",
                         "cfi", "tli",
                         "rmsea","srmr"))
# As an example, there are more.
```

```
#The factors scores for each subject can be required via:
predict(out1)
```

CFA: modification indices - lavaan commands

Remark: Blending confirmatory and exploratory!
Make sure it makes sense!

In lavaan, modification indices can be requested

- within the summary call:

```
summary(out1, modindices = TRUE)
```

- directly:

```
modindices(out1, sort = TRUE)
```

- for specific parameters, say, factor loadings:

```
mi <- modindices(fit)
mi[mi$op == "=~",]
```

Also, have a look at the lavTestScore() function.

CFA: modification indices - interpretation

```
#modindices(fit, sort = TRUE, maximum.number = 7) # or:
modindices(out1, sort = TRUE)[1:7,] # first 7 rows

     lhs op  rhs     mi    epc sepc.lv sepc.all sepc.nox
28   Q77 ~~  Q84 37.984  0.198   0.198    0.314    0.314
41  Q170 ~~ Q196 36.341  0.139   0.139    0.279    0.279
36   Q84 ~~ Q196 30.175 -0.141  -0.141   -0.275   -0.275
35   Q84 ~~ Q170 22.466 -0.129  -0.129   -0.182   -0.182
40   Q84 ~~  Q98 15.821  0.090   0.090    0.154    0.154
25 liked =~  Q84 12.916  0.550   0.234    0.224    0.224
24 liked =~  Q77 12.862 -0.595  -0.253   -0.234   -0.234
```

- mi: If parameter freely estimated, overall Chi-square statistic could decrease by approximately this amount.
- epc (= expected parameter change): Approximate value that a parameter is expected to attain.

CFA: modification indices - cross-loadings

Cross-loadings:

lhs	op	rhs	mi	epc	sepc.lv	sepc.all	sepc.nox
Q170	~~	Q196	36.273	0.138	0.138	0.271	0.271
Q77	~~	Q84	35.431	0.194	0.194	0.305	0.305
Q84	~~	Q196	31.126	-0.143	-0.143	-0.276	-0.276
Q84	~~	Q170	20.426	-0.123	-0.123	-0.170	-0.170
Q84	~~	Q98	16.530	0.092	0.092	0.156	0.156
Beingliked	=~	Q84	13.085	0.552	0.234	0.222	0.222
Q77	~~	Q170	12.876	-0.104	-0.104	-0.166	-0.166
Beingliked	=~	Q77	12.564	-0.586	-0.249	-0.229	-0.229
Havingfun	=~	Q44	11.853	-0.255	-0.203	-0.221	-0.221
Q77	~~	Q44	10.621	-0.078	-0.078	-0.129	-0.129

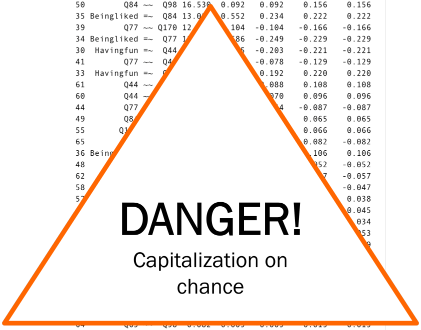
CFA: modification indices – residual variances

Residual covariances:

	lhs	op	rhs	mi	epc	sepc.lv	sepc.all	sepc.nox
	Q170	~~	Q196	36.273	0.138	0.138	0.271	0.271
	Q77	~~	Q84	35.431	0.194	0.194	0.305	0.305
	Q84	~~	Q196	31.126	-0.143	-0.143	-0.276	-0.276
	Q84	~~	Q170	20.426	-0.123	-0.123	-0.170	-0.170
	Q84	~~	Q98	16.530	0.092	0.092	0.156	0.156
	Beingliked	=~	Q84	13.085	0.552	0.234	0.222	0.222
	Q77	~~	Q170	12.876	-0.104	-0.104	-0.166	-0.166
	Beingliked	=~	Q77	12.564	-0.586	-0.249	-0.229	-0.229
	Havingfun	=~	Q44	11.853	-0.255	-0.203	-0.221	-0.221
	Q77	~~	Q44	10.621	-0.078	-0.078	-0.129	-0.129

CFA: modification indices – be aware!

	lhs	op	rhs	mi	epc	sepc.lv	sepc.all	sepc.nox
51	Q170	~~	Q196	36.273	0.138	0.138	0.271	0.271
38	Q77	~~	Q84	35.431	0.194	0.194	0.305	0.305
46	Q84	~~	Q196	31.126	-0.143	-0.143	-0.276	-0.276
45	Q84	~~	Q170	20.426	-0.123	-0.123	-0.170	-0.170
50	Q84	~~	Q98	16.530	0.092	0.092	0.156	0.156
35	Beingliked	=~	Q84	[illegible]	0.552	0.234	0.222	0.222
39	Q77	~~	Q170	[illegible]	[illegible]	-0.104	-0.166	-0.166
34	Beingliked	=~	Q77	[illegible]	[illegible]	-0.249	-0.229	-0.229
30	Havingfun	=~	Q44	[illegible]	[illegible]	-0.203	-0.221	-0.221
41	Q77	~~	[illegible]	[illegible]	[illegible]	-0.078	-0.129	-0.129
33	Havingfun	=~	[illegible]	[illegible]	[illegible]	0.192	0.220	0.220
61	Q44	~~	[illegible]	[illegible]	[illegible]	[illegible]	0.108	0.108
60	Q44	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.096	0.096
44	Q77	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	-0.087	-0.087
49	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.065	0.065
55	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.066	0.066
65	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	-0.082
36	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.106
48	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	-0.052
62	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	-0.057
58	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	-0.047
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.038
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.045
43	Q77	~~	Q76	0.068	0.008	0.008	0.011	0.011
47	Q84	~~	Q44	0.042	0.005	0.005	0.007	0.007
54	Q170	~~	Q76	0.011	0.003	0.003	0.004	0.004



CFA: modification indices - modification

```
modindices(out1, sort = TRUE)[1,]

   lhs op rhs     mi   epc sepc.lv sepc.all sepc.nox
28 Q77 ~~ Q84 37.984 0.198   0.198    0.314    0.314

# Allow residuals of Q170 and Q196 to covary
```

```
# Modified model:
# two-factor CFA + residual covariance Q170 and Q196
model.2CFA_mod <- "
 Having fun  =~ Q77 + Q84 + Q170 + Q196
 Being liked =~ Q44 + Q63 + Q76  + Q98
 Q170 ~~ Q196
"

# Fit model
out1_mod <- cfa(model.2CFA_mod, data=data_sapi,
                    missing='fiml', fixed.x=F)  # use FIML

Error in eval(sc, parent.frame()):  object 'data_sapi' not found
```

CFA: modification indices - test modification

```
anova(out1, fit_2CFA_mod)[,-c(2,3)] # without AIC & BIC

Error in eval(expr, envir, enclos): object 'fit_2CFA_mod' not found
```

```
modindices(out1, sort = TRUE)[1,]

   lhs op rhs     mi   epc sepc.lv sepc.all sepc.nox
28 Q77 ~~ Q84 37.984 0.198   0.198    0.314    0.314
```

CFA: modification indices - new parameter value

```
parameterEstimates(out1_mod)[9,-c(5,6,7)] # no se, z, and p

Error in eval(expr, envir, enclos):  object 'out1_mod' not found
```

```
modindices(out1, sort = TRUE)[1,1:5] # no sepc

   lhs op rhs     mi   epc
28 Q77 ~~ Q84 37.984 0.198
```

CFA: cross-loadings approximately zero

Problem:

- Restricting cross-loadings to exactly zero can be too strict.
- Consequence: rejection of the model, model modifications that capitalise on chance.

(Possible) solution in Bayesian SEM (BSEM) blavaan:

- Replace exact zero restrictions with approximate ones.
- Using Bayesian small-variance priors.

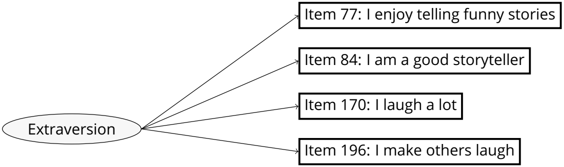
Interesting reading:

– Merkle, E. C., & Rosseel, Y. (2018). blavaan: Bayesian Structural Equation Models via Parameter Expansion. Journal of Statistical Software, 85(4), 1–30. https://doi.org/10.18637/jss.v085.i04

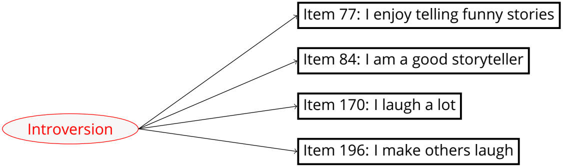
– Muthén, B., & Asparouhov, T. (2012). Bayesian structural equation modelling: A more flexible representation of substantive theory. Psychological Methods, 17(3), 313-335.

Latent variable scaling

Latent variables are not observed, thus no inherent scale.



Latent variable scaling Ctd.



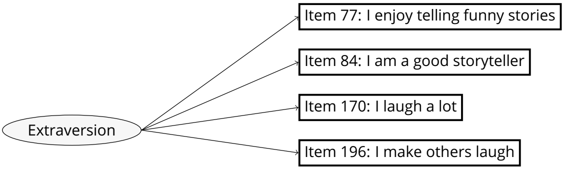
Therefore, set up model such that scale of latent variable is clear.

Three common ways

1. Marker-variable method
Constrain one of the factor loadings (default).

2. Reference group method:
Constrain the factor variance.

3. Effect coding:
Constrain the average of the loadings.



1. Marker-variable method (default)

Default parameterization:

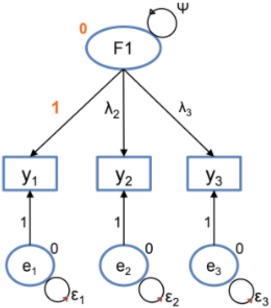
- First factor loading constrained at 1.
- Factor mean constrained at 0.

Other defaults:

- Mean of residuals is by definition 0.
- Residuals have a loading of 1.

Estimated:

- factor variance (Ψ),
- 'other' factor loadings (λ_2, λ_3),
- all item intercepts (ν_1, ν_2, ν_3),
- all residual variances (ϵ_1, ϵ_2, ϵ_3).



1. Default marker-variable method - lavaan

```
# Model
model.1CFA <- '
 Extraversion =~ Q77 + Q84 + Q170 + Q196
'

# Fit model
fit_1CFA <- cfa(model.1CFA, data=data_sapi,
                missing='fiml', fixed.x=F)  # use FIML

Error in eval(sc, parent.frame()):  object 'data_sapi' not found
```

- First factor loading constrained at 1:

```
Extraversion =~
  Q77                    1.000
```

- Factor mean constrained at 0:

```
Extraversion       0.000
```

1. Default marker-variable method - lavaan Ctd

```
parameterEstimates(fit_1CFA)[1:4,-c(5,6,7)]

Error in eval(expr, envir, enclos):  object 'fit_1CFA' not found
```

Factor loading of first indicator fixed to 1.
all other loadings are relative to that.

If reference category changed, other loadings also change.

2. Reference-group method

Parameterization:

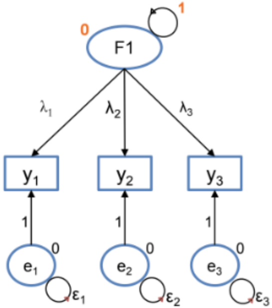
- Factor variance constrained at 1.
- Factor mean constrained at 0.

Defaults:

- Mean of residuals is by definition 0.
- Residuals have a loading of 1.

Estimated:

- all factor loadings (λ_1, λ_2, λ_3),
- all item intercepts (ν_1, ν_2, ν_3),
- all residual variances (ϵ_1, ϵ_2, ϵ_3).



2. Reference-group method - lavaan

```
# Model
model.1CFA_RefGr <- '
  # Free first factor loading, using: NA*
  Extraversion =~ NA*Q77 + Q84 + Q170 + Q196

  # Set factor variance to 1, using: 1*
  Extraversion ~~ 1*Extraversion
 '


# Fit model
fit_1CFA_RefGr <- cfa(model.1CFA_RefGr, data=data_sapi,
                missing='fiml', fixed.x=F)  # use FIML

Error in eval(sc, parent.frame()):  object 'data_sapi' not found
```

- Factor variance constrained at 1:

```
Extraversion        1.000
```

- Factor mean constrained at 0:

```
Extraversion        0.000
```

2. Reference-group method - lavaan Ctd

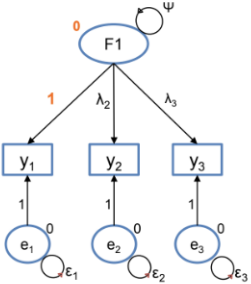
```
parameterEstimates(fit_1CFA_RefGr)[1:4,-c(5,6,7)]

Error in eval(expr, envir, enclos):  object 'fit_1CFA_RefGr' not found
```

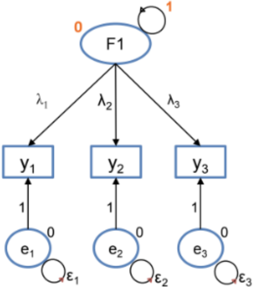
Advantage:
All factor loadings and scores on standardized metric.

Which method to choose?

1. Marker-variable method



2. Reference-group method



Does not matter for substantive conclusions.
Sometimes, pragmatic reasons.

3. Effects-coding method

Parameterization:

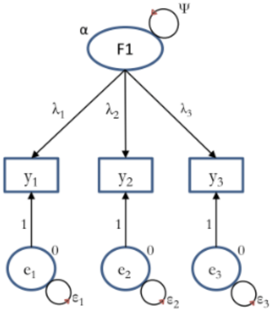
- Constrain the average of the factor loadings to 1: $\frac{1}{3}\sum_{i=1}^{3}\lambda_i = 1.$
- Constrain the average of the item intercepts to 0: $\frac{1}{3}\sum_{i=1}^{3}\nu_i = 0.$

Defaults:

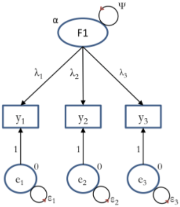
- Mean of residuals is by definition 0.
- Residuals have a loading of 1.

Estimated (subject to the constraints):

- factor variance (Ψ),
- factor mean (α),
- all factor loadings (λ_1, λ_2, λ_3),
- all item intercepts (ν_1, ν_2, ν_3),
- all residual variances (ϵ_1, ϵ_2, ϵ_3).



3. Effects-coding method Ctd



Interpretations can be intuitive:

- Factor on similar scale as the indicators.
- Factor variance (Ψ): average variance of each indicator that can be explained by the factor.
- Factor mean (α): weighted mean of the indicator means

3. Effects-coding method - lavaan model

```
# Model
model.1CFA_EffC <- '
  # Label parameters, such that they can be constrained
  Extraversion =~ lambda1*Q77 + lambda2*Q84 +
                  lambda3*Q170 + lambda4*Q196
  # intercepts
  Q77  ~ nu1*1
  Q84  ~ nu2*1
  Q170 ~ nu3*1
  Q196 ~ nu4*1

  # Constrain average of loadings to 1, i.e., set sum to 4
  lambda1 == 4 - lambda2 - lambda3 - lambda4
  # Constrain average of item intercepts to 0,
  # i.e., set sum to 0
  nu1 == 0 - nu2 - nu3 - nu4
 '
```

3. Effects-coding method - fit lavaan model

Now, use the lavaan() function:

```
# Fit model: Now, use the lavaan() function!
fit_1CFA_EffC <- lavaan(model.1CFA_EffC, data=data_sapi,
                        missing='fiml', fixed.x=F,
                        auto.var = TRUE,
                        auto.fix.first = FALSE,
                        auto.cov.lv.x = TRUE,
                        int.ov.free = TRUE)

Error in eval(expr, envir, enclos):  object 'data_sapi' not found
```

3. Effects-coding method - lavaan outpu

- Constrain the average of the factor loadings to 1: $\frac{1}{4}\sum_{i=1}^{4}\lambda_i = 1$.

```
parameterEstimates(fit_1CFA_EffC)[1:4,1:5]
Error in eval(expr, envir, enclos):  object 'fit_1CFA_EffC' not found
```

- Constrain the average of the item intercepts to 0: $\frac{1}{4}\sum_{i=1}^{4}\nu_i = 0$.

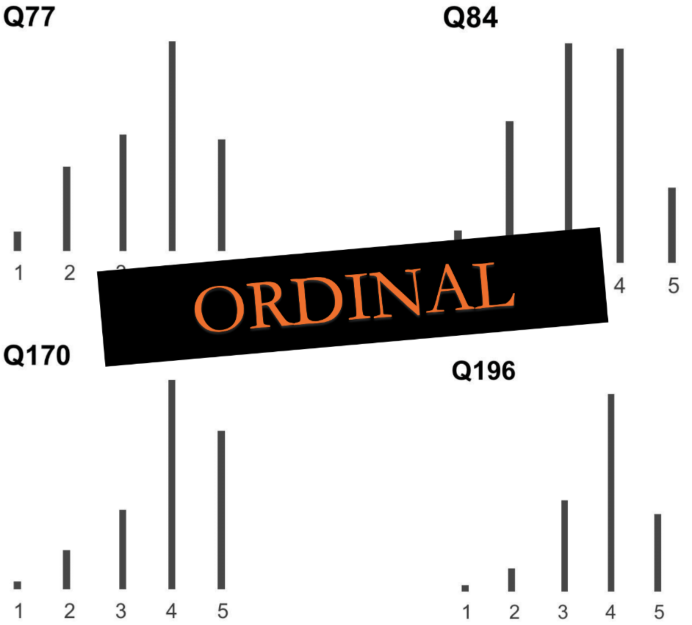
```
parameterEstimates(fit_1CFA_EffC)[5:8,1:5]
Error in eval(expr, envir, enclos):  object 'fit_1CFA_EffC' not found
```

Extra:

Categorical/Ordinal or continuous indicators?

Note: in cfa() you can, for example use ‘ordered = TRUE’ for endogenous variable.
Default then: estimator = ”WLSMV”.

More information on: https://lavaan.ugent.be/tutorial/cat.html


Q77
1 2
Q84
4 5
ORDINAL
Q170
1 2 3 4 5
Q196
1 2 3 4 5

Remark!

Do NOT use a χ^2 test or IC (AIC or BIC)
to compare categorical and continuous models:

- Obviously not nested (so, no χ^2 test anyway).
- AND likelihoods of categorical and continuous indicator models are incomparable!

Note: χ^2 test and IC are based on (log) likelihood (= fit).

Interesting Reading

https://lavaan.ugent.be/tutorial/cat.html

Muthén, B. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. Psychometrica, 49(1), 115-132.

Rhemtulla, M., Brosseau-Liard, P.É., & Savalei, V. (2012). When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. Psychological Methods, 17(3), 354-373.

Sventina, D., Rutkowski, D. (2020). Multiple group invariance with categorical outcomes using updated guidelines: an illustration using Mplus and the lavaan/semtools packages. Structural Equational Modelling: A Multidisciplinary Journal, 27(1), 111-130