CFA - theory

The theory of Confirmatory Factor Analysis

Intro

Today's goal:

Teach the idea behind Confirmatory Factor Analysis.

Outline:

- Rationale behind CFA
- Model specification and identification
- Estimation
- Item, Factor, and Model Fit

CFA

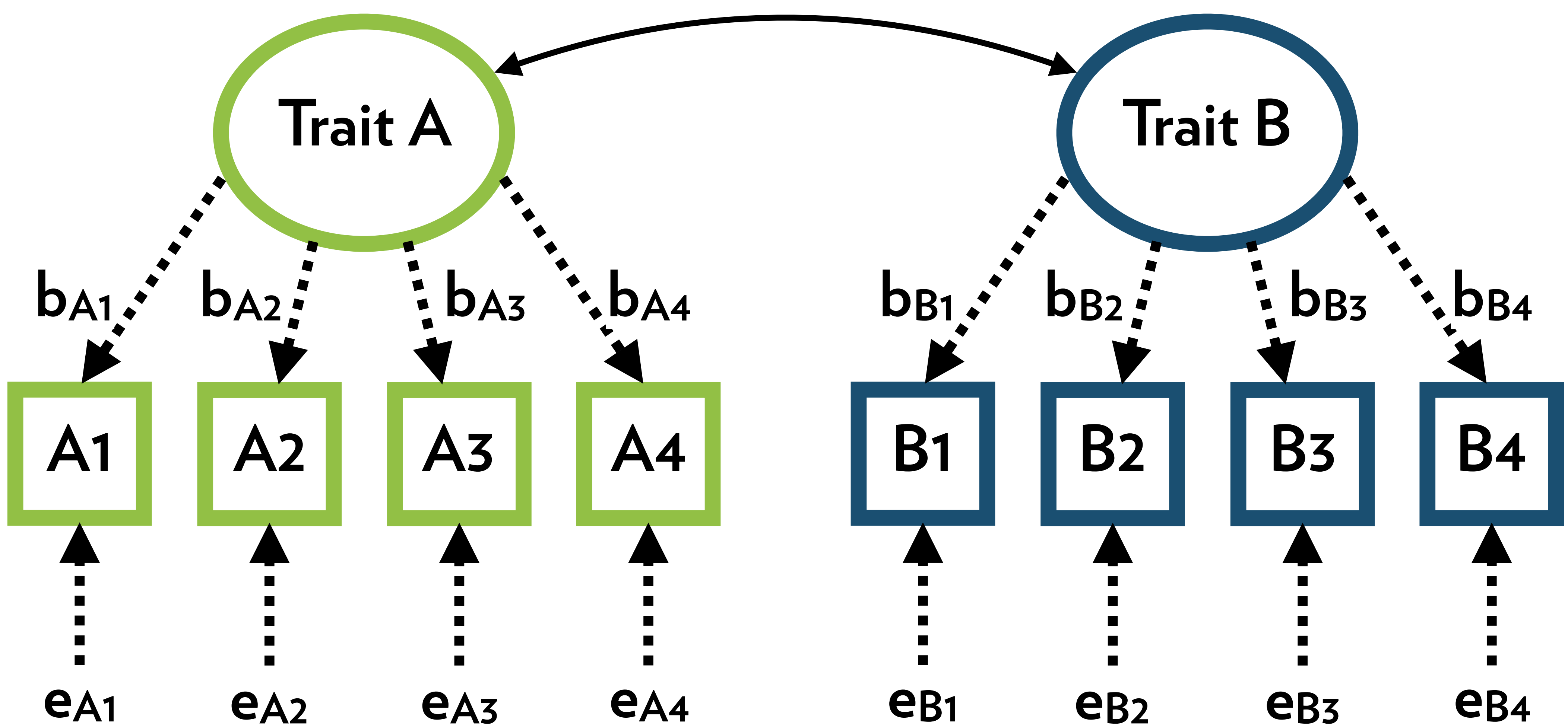
Confirmatory Factor Analysis

CFA

What if the items do not have an equal correlation?

$b_{A1} \neq b_{A2} \neq b_{A3} \neq b_{A4}$

And what if we want to measure multiple traits?



Why CFA?

Establish convergent and discriminant validity

- CFA can suggest ways to remedy problems with the scale

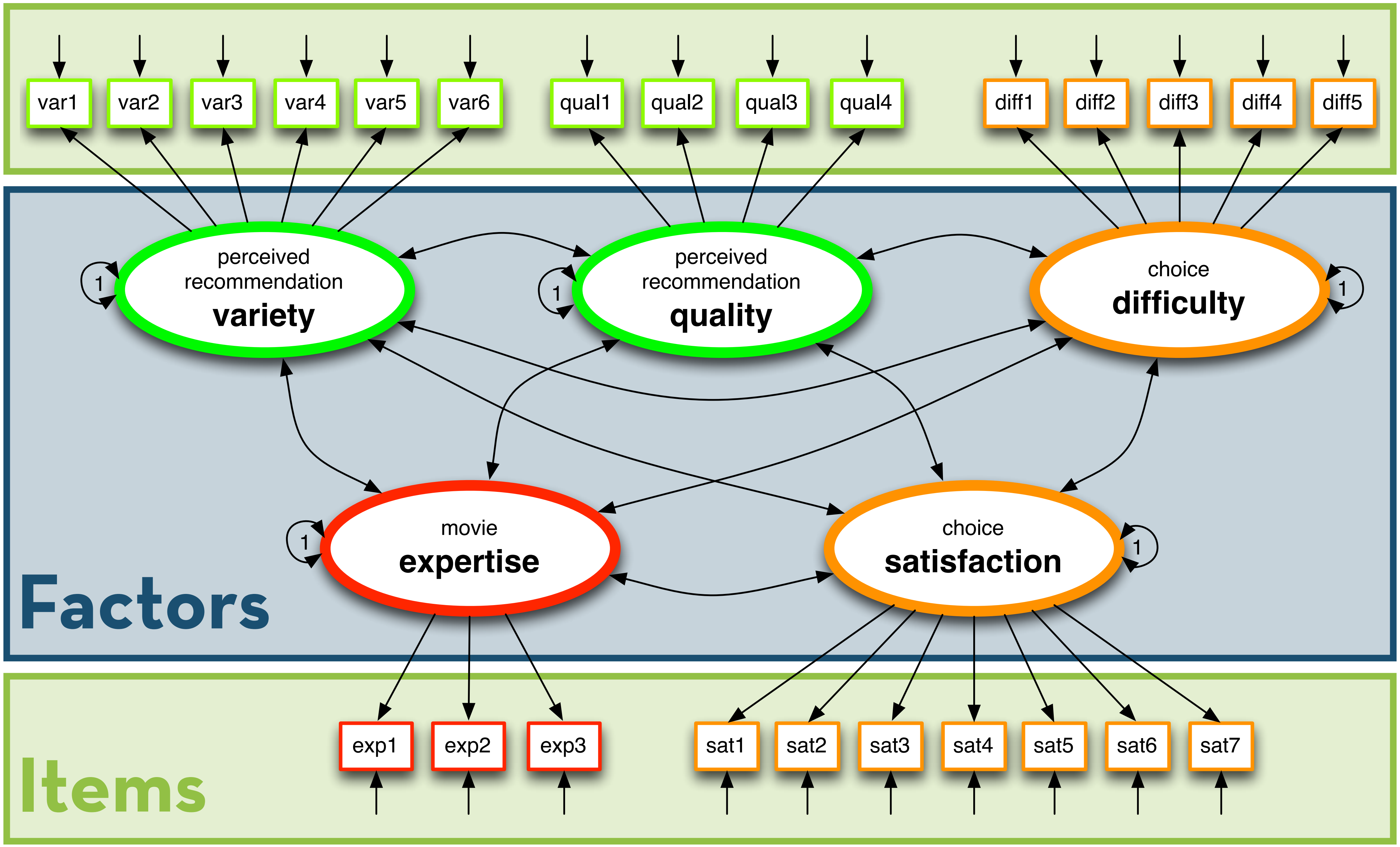
Outcome is a normally distributed measurement scale

- Even when the items are yes/no, 5- or 7-point scales!

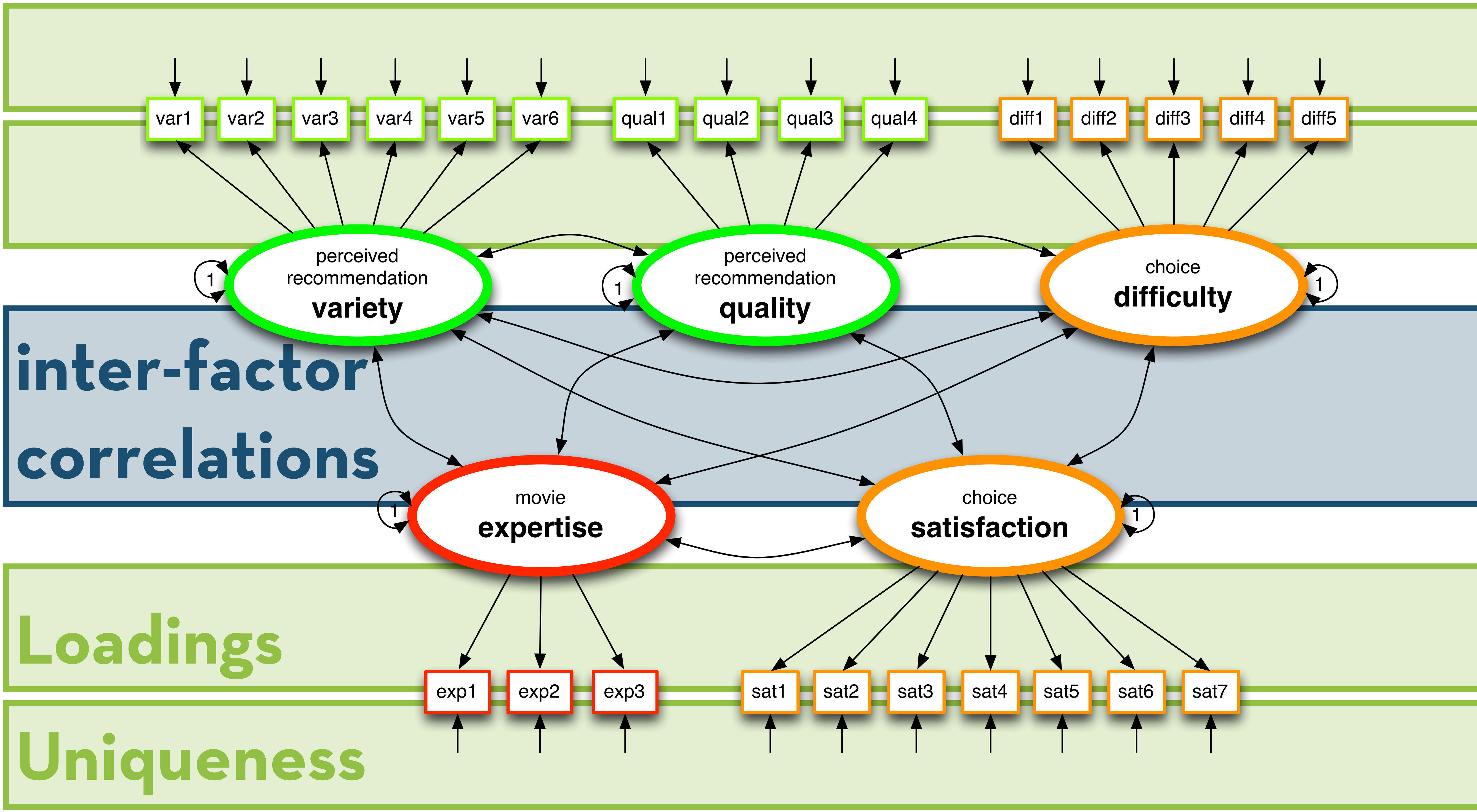
The scale captures the "shared essence" of the items

- You can remove the influence of measurement error in your statistical tests!

CFA: the concept



CFA: the concept



CFA: the concept

Factors are **latent constructs** that represent the trait or concept to be measured

- The latent construct cannot be measured directly

The latent construct "**causes**" users' answers to items

- Items are therefore also called **indicators**

Like any measurement, indicators are not perfect measurements

- They depend on the true score (loading) as well as some measurement error (uniqueness)

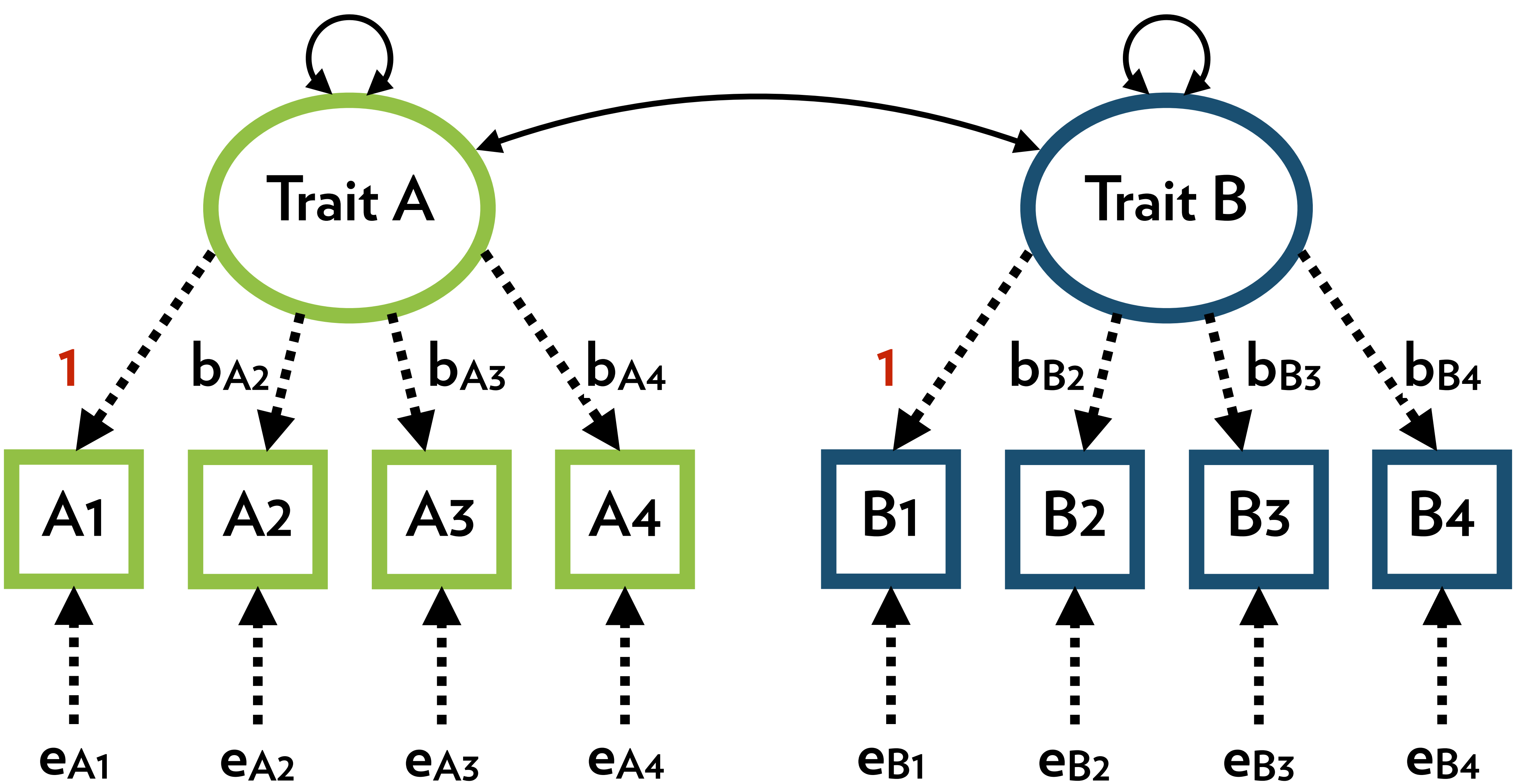
Identification

of CFA models

Identification

Factors need to be **scaled** in order to be identified

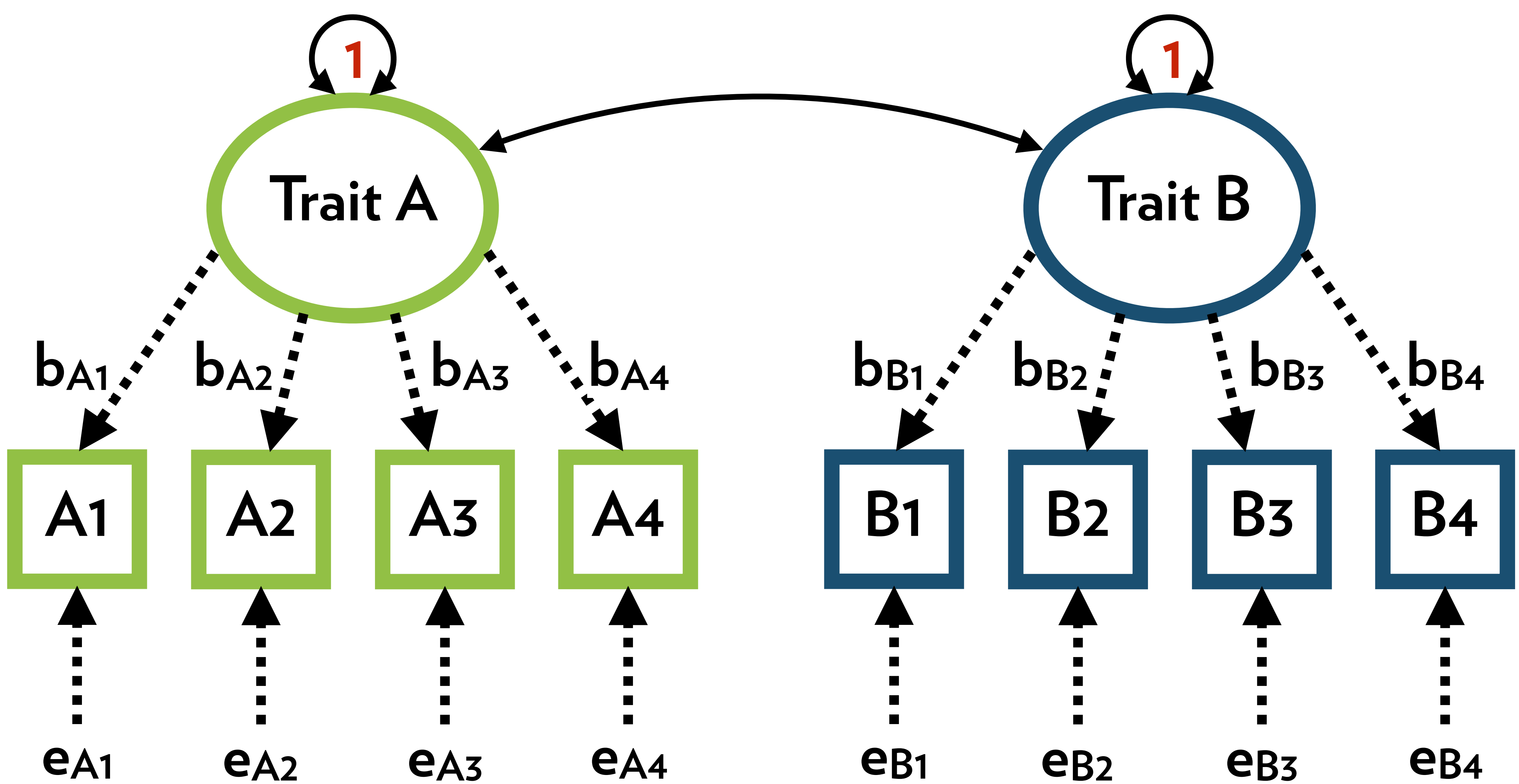
Unit Loading Identification (ULI)



Identification

Factors need to be **scaled** in order to be identified

Unit Variance Identification (UVI)



How it works

By looking at the **overlap** (covariance) between items, we can separate the measurement error from the true score!

- The scale captures the "shared essence" of the items

The factor is assumed to explain **all** of this overlap

- The partial correlation between A1 and A2 controlling for factor A is assumed to be zero!

There can be a **residual correlation**, but such a model will be more difficult to identify

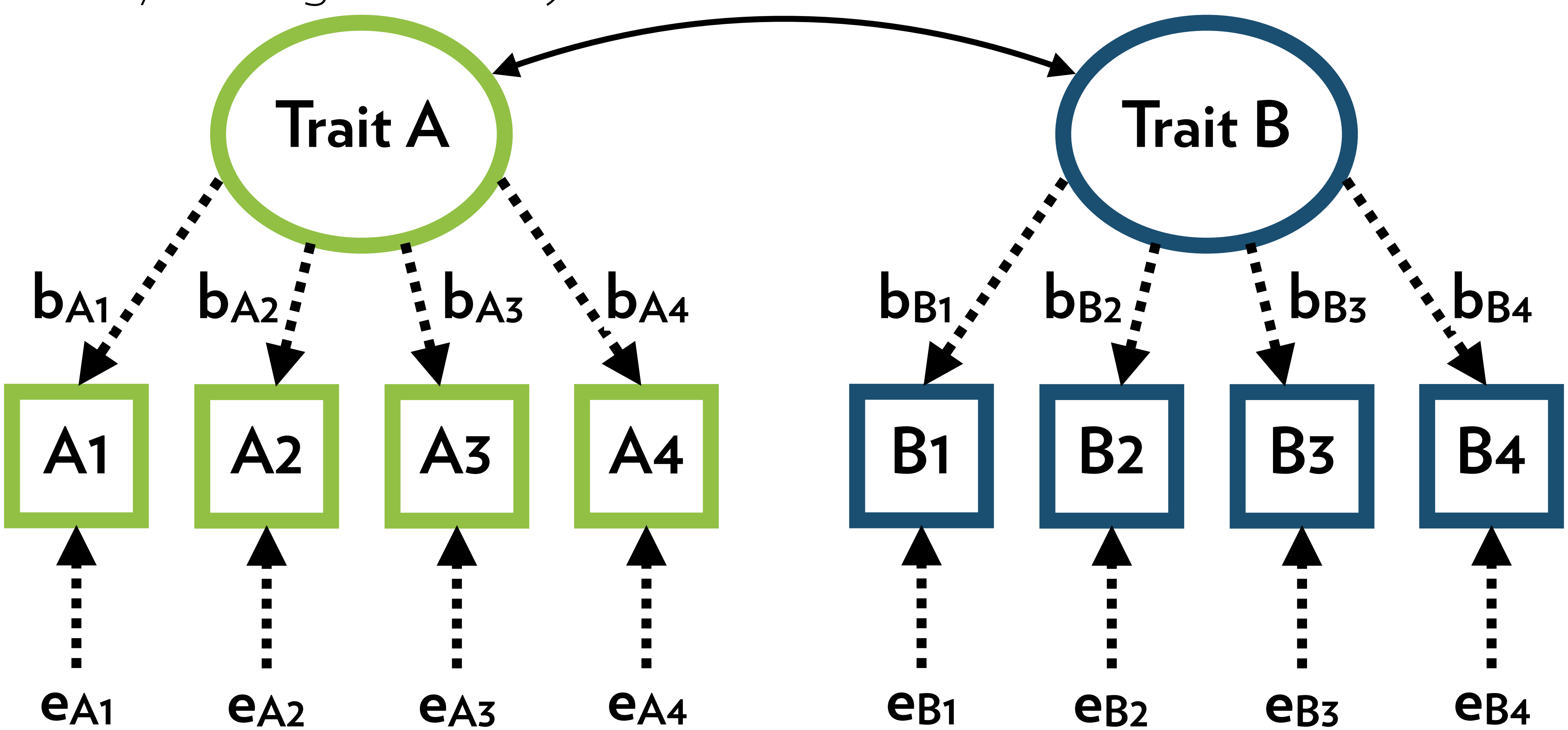
How it works

- Each item belongs to only a **single factor**
  - Again, **cross-loadings** are possible, but such a model will be more difficult to identify
- There still exist a relationship between A1 and factor B
  - Because factor B is correlated to factor A

Identification

A model with no residual correlations and no cross-loadings is called a **unidimensional model**

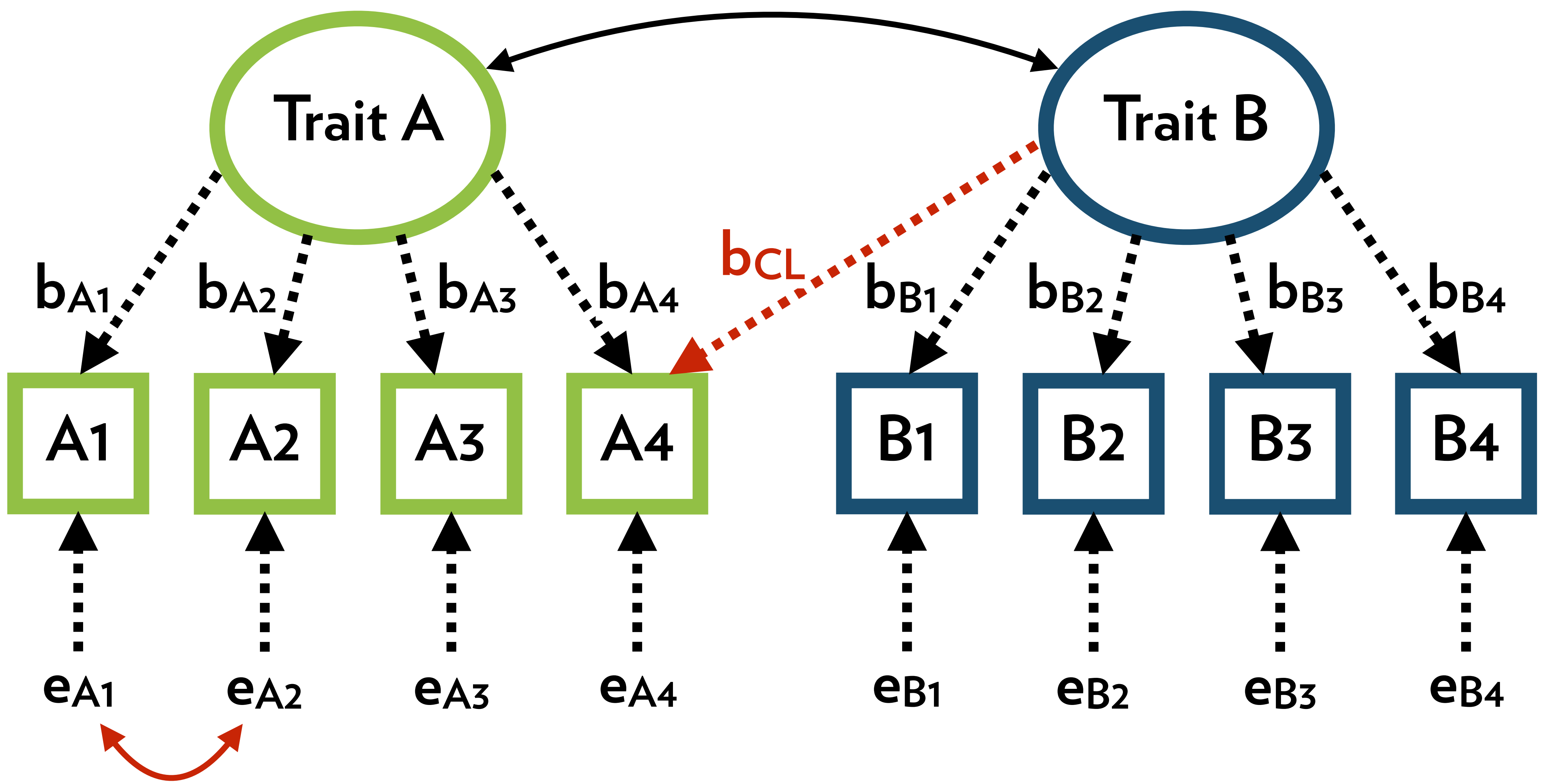
It is identified if each factor has at least 2 items (3 if it is only a single factor)



Identification

Example model with residual correlations and cross-loadings

Try to avoid this as much as possible



Estimation

of CFA models

How it works

The basis for Factor Analysis is the item correlation matrix

- Or the covariance matrix (which is a little more complex)

How do we determine the loadings etc?

- By **modeling** the correlation matrix as closely as possible!
- This is the job of that wizard in your computer...

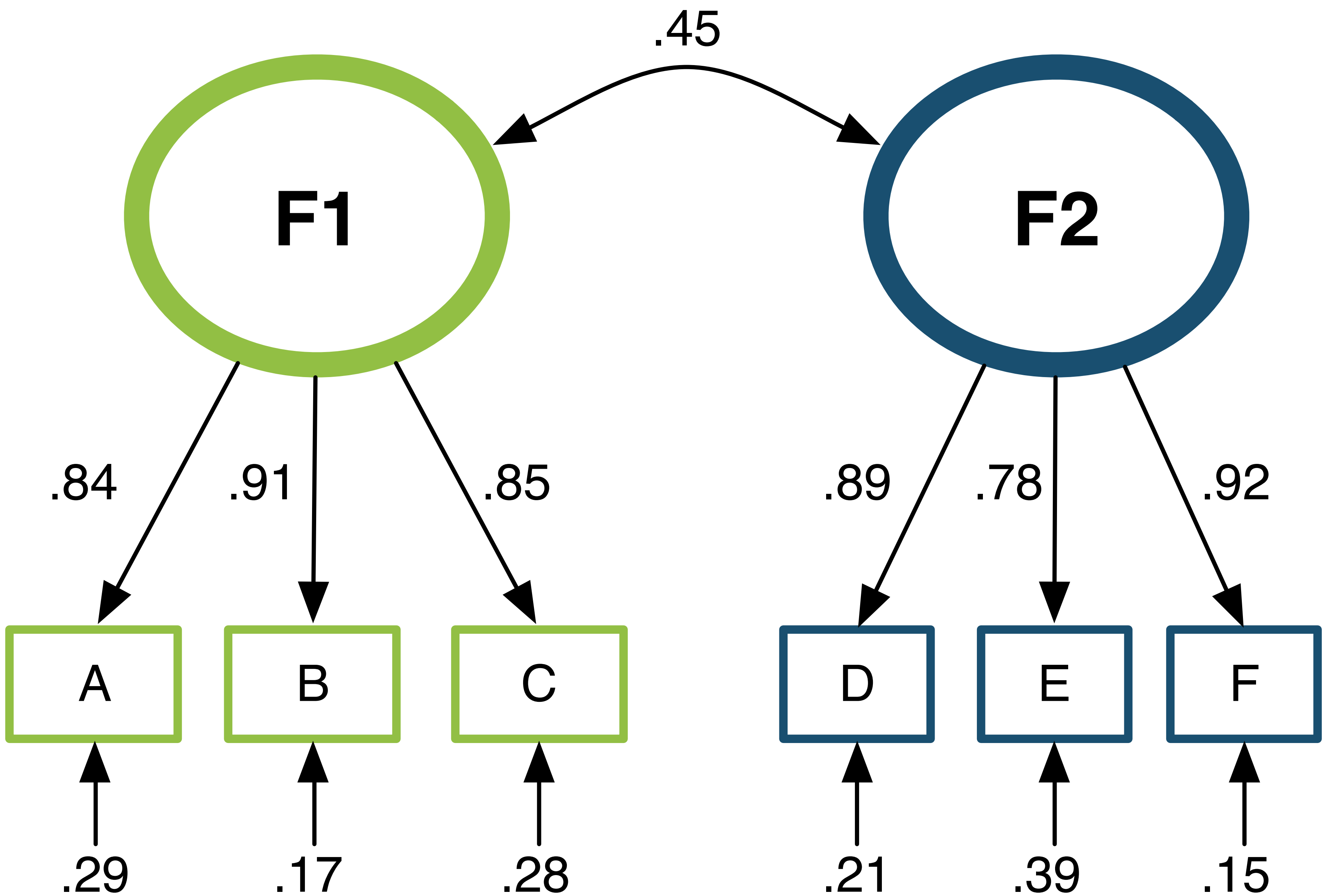
Observed

	A	B	C	D	E	F
A	1.00	0.73	0.71	0.34	0.49	0.34
B	0.73	1.00	0.79	0.35	0.32	0.32
C	0.71	0.79	1.00	0.29	0.33	0.35
D	0.34	0.35	0.29	1.00	0.74	0.81
E	0.49	0.32	0.33	0.74	1.00	0.75
F	0.34	0.32	0.35	0.81	0.75	1.00

Observed

	A	B	C	D	E	F
A	1.00	0.73	0.71	0.34	0.49	0.34
B	0.73	1.00	0.79	0.35	0.32	0.32
C	0.71	0.79	1.00	0.29	0.33	0.35
D	0.34	0.35	0.29	1.00	0.74	0.81
E	0.49	0.32	0.33	0.74	1.00	0.75
F	0.34	0.32	0.35	0.81	0.75	1.00

Model



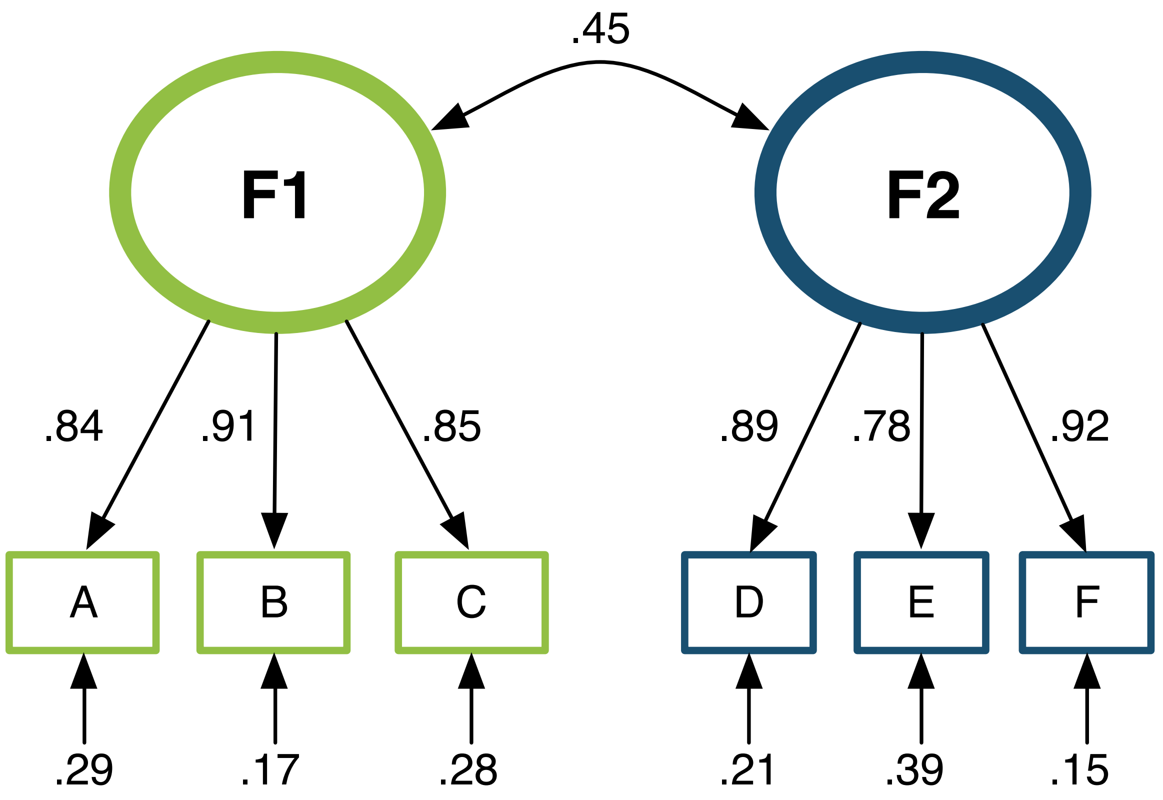
Estimation

Estimate the correlation matrix:

r_{AB} = .84*.91 = .76

r_{AD} = .84*.45*.89 = .34

etc.



Estimated

	A	B	C	D	E	F
A	0.71	0.76	0.71	0.34	0.29	0.35
B	0.76	0.83	0.77	0.36	0.32	0.38
C	0.71	0.77	0.72	0.34	0.30	0.35
D	0.34	0.36	0.34	0.79	0.69	0.82
E	0.29	0.32	0.30	0.69	0.61	0.72
F	0.35	0.38	0.35	0.82	0.72	0.85

Estimated

	A	B	C	D	E	F
A	1.00	0.73	0.71	0.34	0.49	0.34
B	0.73	1.00	0.79	0.35	0.32	0.32
C	0.71	0.79	1.00	0.29	0.33	0.35
D	0.34	0.35	0.29	1.00	0.74	0.81
E	0.49	0.32	0.33	0.74	1.00	0.75
F	0.34	0.32	0.35	0.81	0.75	1.00

–

	A	B	C	D	E	F
A	0.71	0.76	0.71	0.34	0.29	0.35
B	0.76	0.83	0.77	0.36	0.32	0.38
C	0.71	0.77	0.72	0.34	0.30	0.35
D	0.34	0.36	0.34	0.79	0.69	0.82
E	0.29	0.32	0.30	0.69	0.61	0.72
F	0.35	0.38	0.35	0.82	0.72	0.85

=

	A	B	C	D	E	F
A	0.29	–0.03	0.00	0.00	0.20	–0.01
B	–0.03	0.17	0.02	–0.01	0.00	–0.06
C	0.00	0.02	0.28	–0.05	0.03	0.00
D	0.00	–0.01	–0.05	0.21	0.05	–0.01
E	0.20	0.00	0.03	0.05	0.39	0.03
F	–0.01	–0.06	0.00	–0.01	0.03	0.15

Residual

	A	B	C	D	E	F
A	0.29	–0.03	0.00	0.00	0.20	–0.01
B	–0.03	0.17	0.02	–0.01	0.00	–0.06
C	0.00	0.02	0.28	–0.05	0.03	0.00
D	0.00	–0.01	–0.05	0.21	0.05	–0.01
E	0.20	0.00	0.03	0.05	0.39	0.03
F	–0.01	–0.06	0.00	–0.01	0.03	0.15

How it works

Covariance matrix, estimate variables to fit

- ML, WLS

Use estimates and misfit in item-, factor-, and model-fit metrics

- Item-fit: Loadings, communality, residuals
- Factor-fit: Average Variance Extracted
- Model-fit: Chi-square test, CFI, TLI, RMSEA

Item, factor & model fit

of CFA models

Item-fit metrics

Variance extracted (squared standardized loading):

- Regression R^2, which is the amount of variance explained by the factor (1-uniqueness)
- Should be > 0.50 (although some argue 0.40 is okay)

In lavaan output: r-squared

Variance extracted

	A	B	C	D	E	F
A	0.29	–0.03	0.00	0.00	0.20	–0.01
B	–0.03	0.17	0.02	–0.01	0.00	–0.06
C	0.00	0.02	0.28	–0.05	0.03	0.00
D	0.00	–0.01	–0.05	0.21	0.05	–0.01
E	0.20	0.00	0.03	0.05	0.59	0.03
F	–0.01	–0.06	0.00	–0.01	0.03	0.15

Item-fit metrics

Residual correlations:

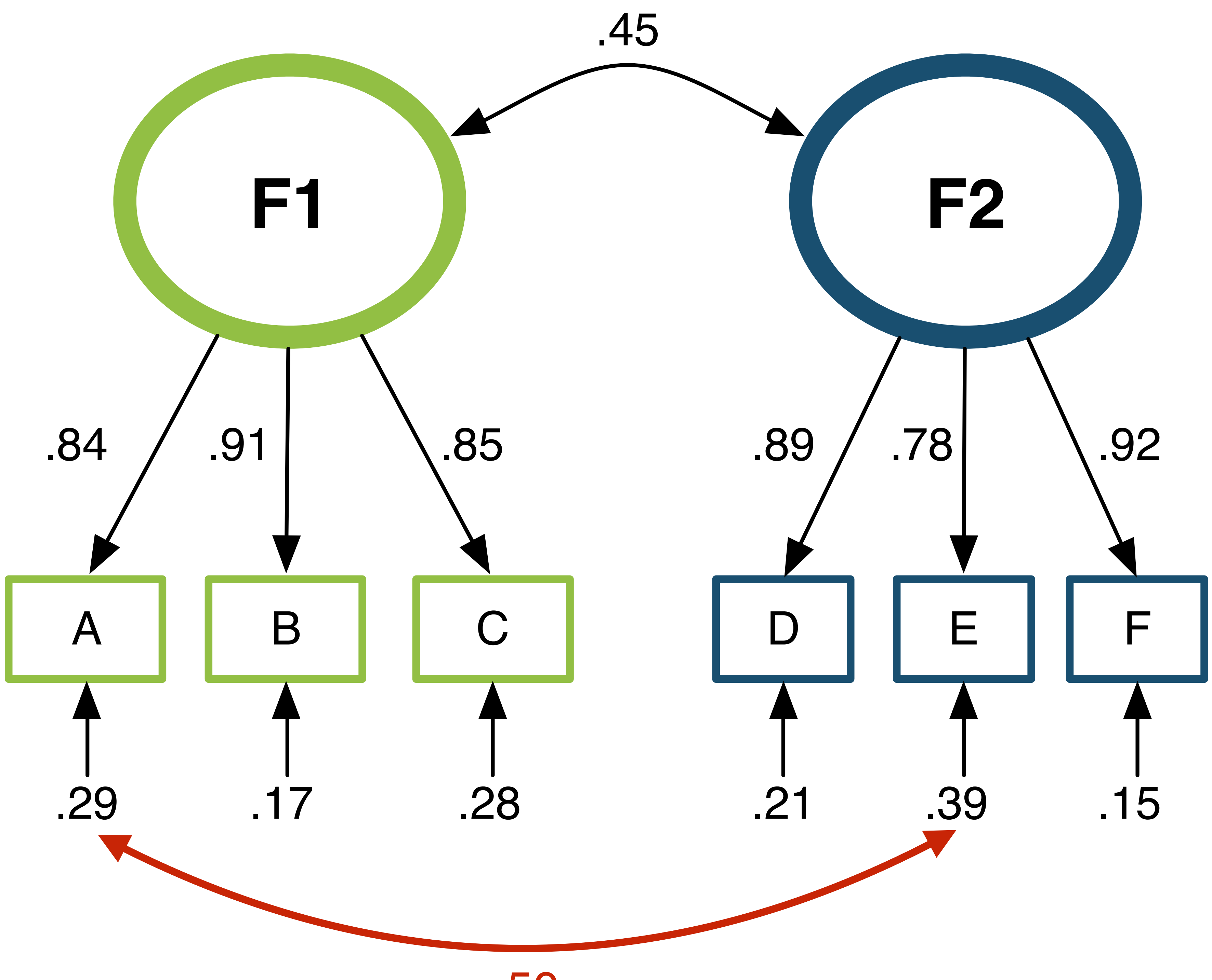
- The observed correlation between two items is significantly higher (or lower) than predicted

In lavaan output: modification indices

Residual...

	A	B	C	D	E	F
A	0.29	–0.03	0.00	0.00	0.20	–0.01
B	–0.03	0.17	0.02	–0.01	0.00	–0.06
C	0.00	0.02	0.28	–0.05	0.03	0.00
D	0.00	–0.01	–0.05	0.21	0.05	–0.01
E	0.50	0.00	0.03	0.05	0.39	0.03
F	–0.01	–0.06	0.00	–0.01	0.03	0.15

Solution



Split factors…

If you have lots of residuals, it might mean that factors should be split up!

- E.g., "satisfaction" turns out to be "satisfaction" and "intention to use"

In lavaan: low r-squared, many high modification indices

Split factors…

	A	B	C	D	E	F
A	0.59	0.29	0.24	–0.27	–0.23	–0.30
B	0.29	0.47	0.19	–0.20	–0.19	–0.31
C	0.24	0.19	0.58	–0.31	–0.25	–0.28
D	–0.27	–0.23	–0.30	0.51	0.30	0.18
E	–0.20	–0.19	–0.31	0.30	0.69	0.25
F	–0.31	–0.25	–0.28	0.18	0.25	0.45

Item-fit metrics

Cross-loadings:

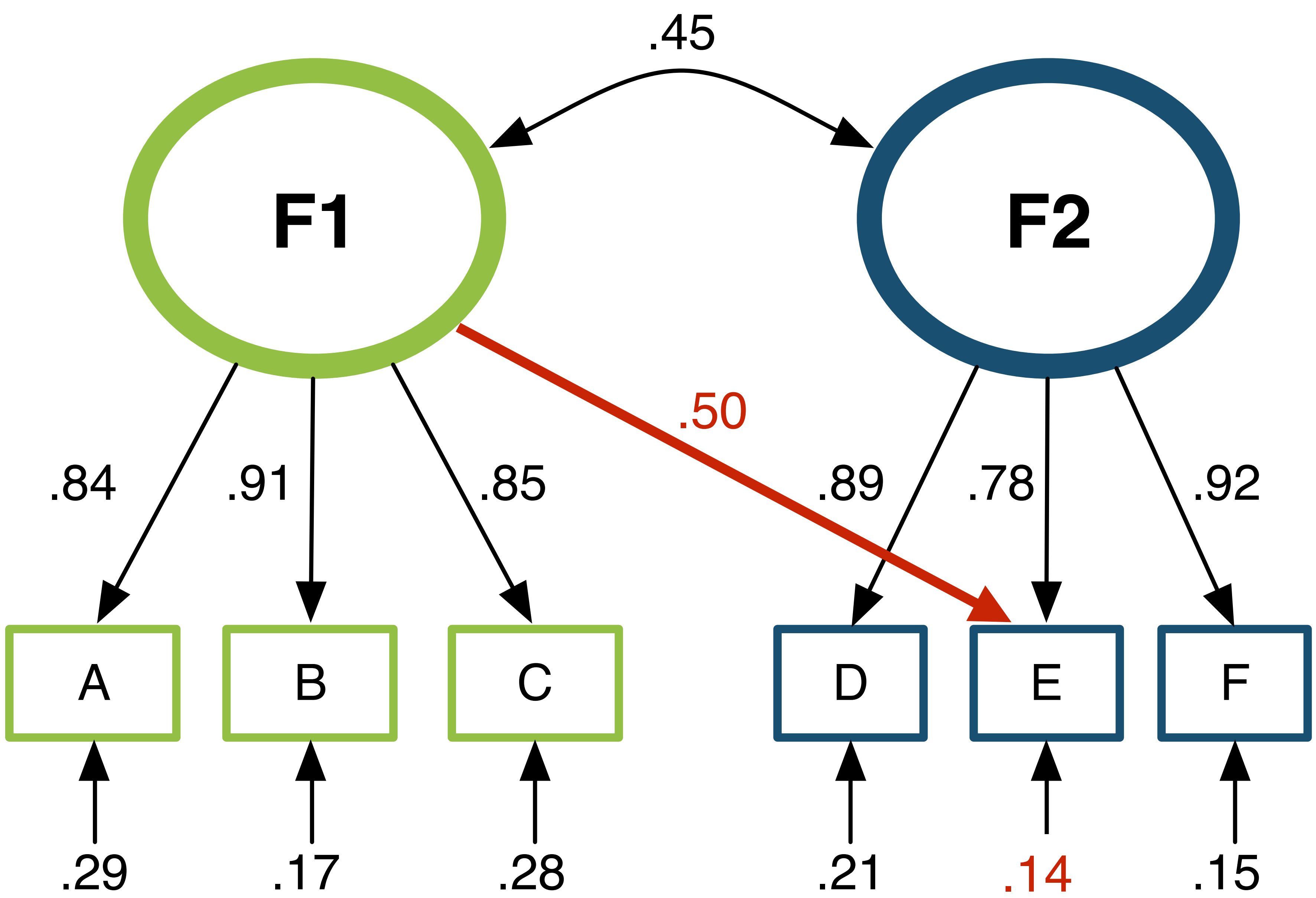
- When the model suggest that the model fits significantly better if an item also loads on an additional factor
- Could mean that an item actually measures two things

In R: modification indices

Cross-loadings…

	A	B	C	D	E	F
A	0.29	–0.03	0.00	0.00	0.20	–0.01
B	–0.03	0.17	0.02	–0.01	0.00	–0.06
C	0.00	0.02	0.28	–0.05	0.03	0.00
D	0.00	–0.01	–0.05	0.21	0.05	–0.01
E	0.34	0.36	0.36	0.05	0.39	0.03
F	–0.01	–0.06	0.00	–0.01	0.03	0.15

Solution

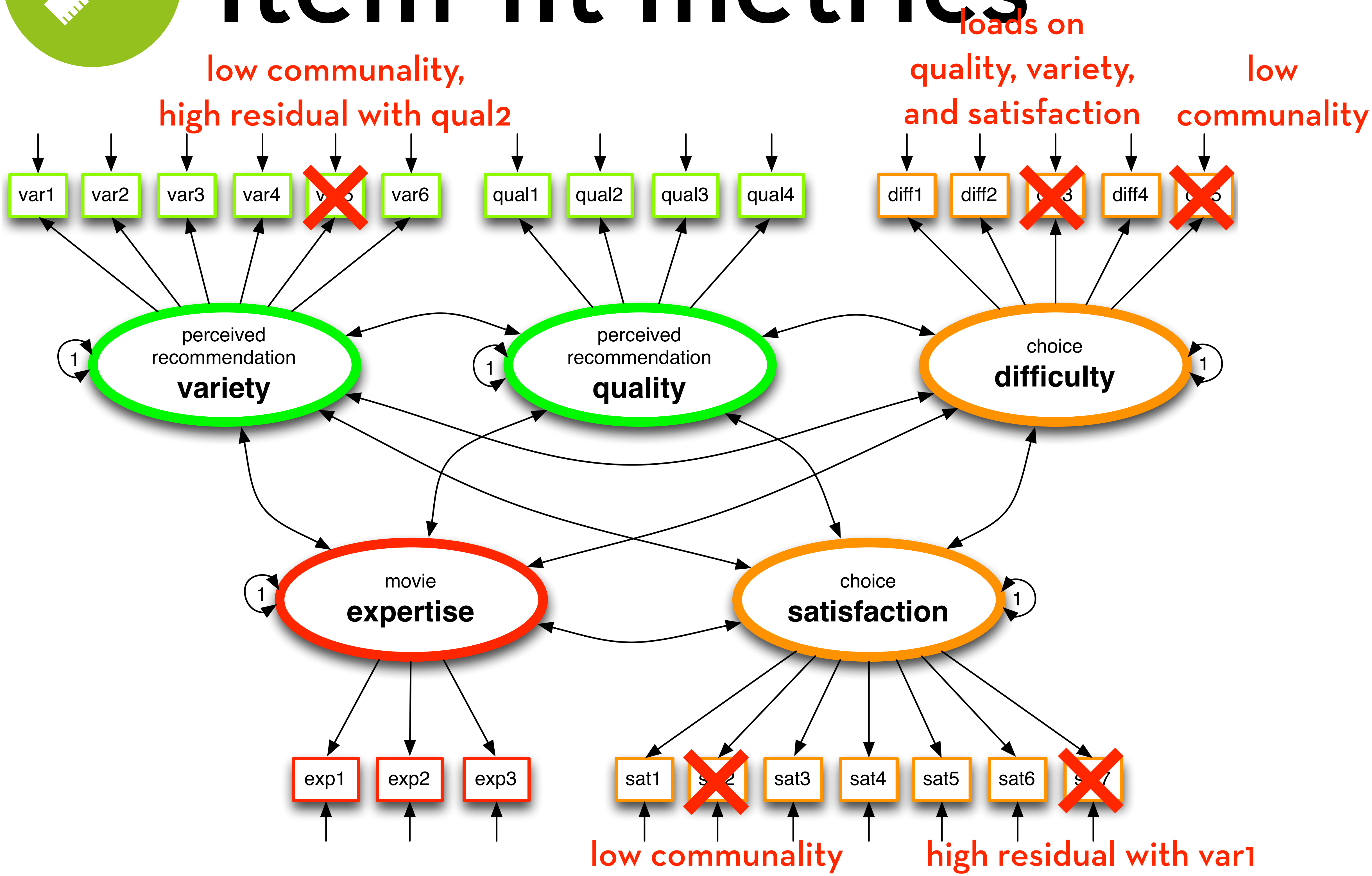


Item-fit metrics

For all these metrics:

- Remove items that do not meet the criteria, but be careful to keep at least 3 items per factor
- One may remove an item that has values much lower than other items, even if it meets the criteria

Item-fit metrics



Factor-fit metrics

AVE (average variance extracted, over all items per factor)

- In lavaan: average r-squared of items per factor

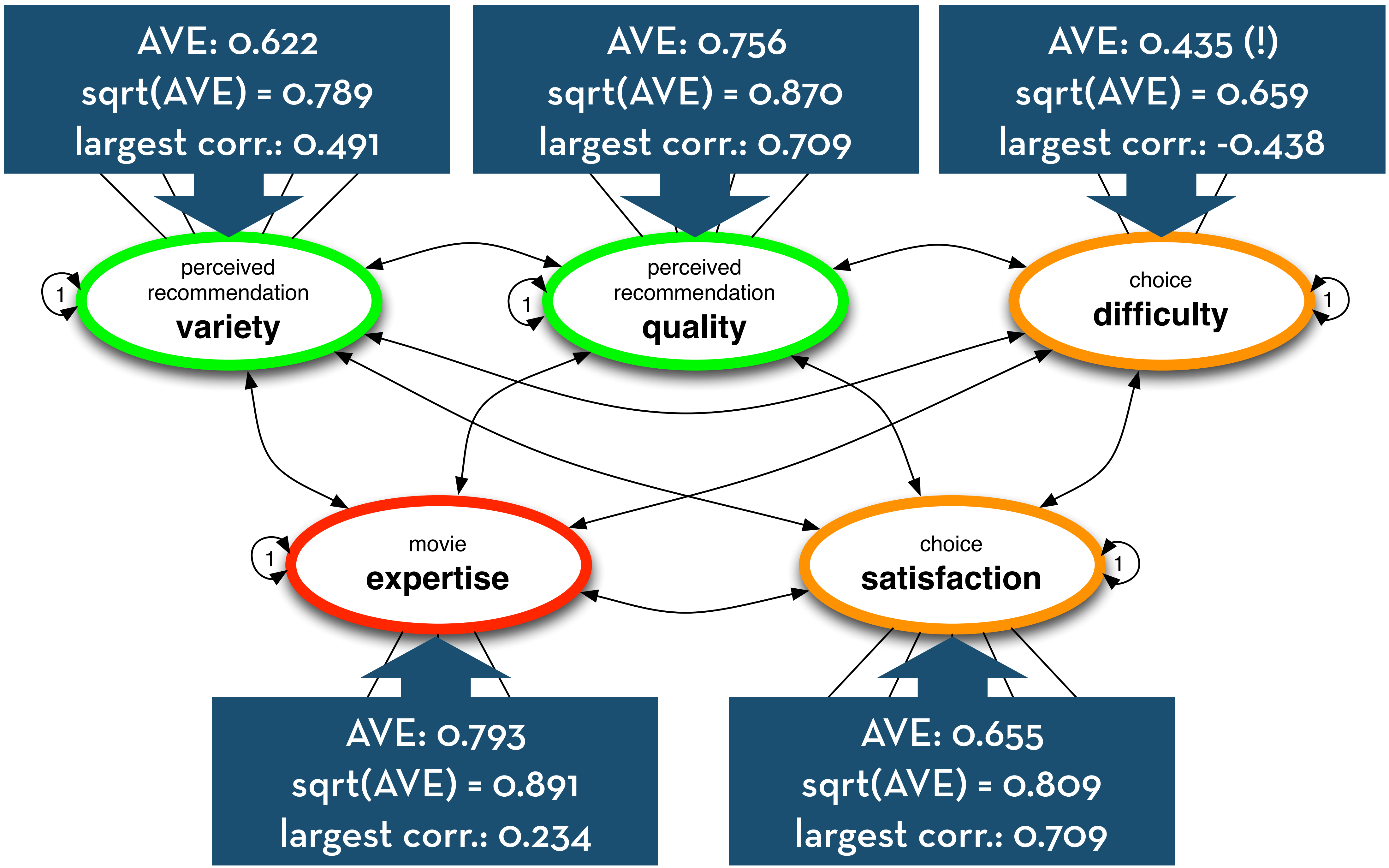
A value > 0.50 indicates convergent validity

- Otherwise, remove worst-fitting items

Also, the square root of the AVE of a factor should be higher than its highest correlation with other factors

- This indicates discriminant validity
- Otherwise, the factors may as well be combined

Factor-fit metrics



Model-fit metrics

Chi-square test of model fit:

- Tests whether there any significant misfit between estimated and observed correlation matrix
- Often this is true ($p < .05$); factor models are rarely perfect!
- Alternative metric: chi-squared / df < 3 (good fit) or < 2 (great fit)

Model-fit metrics

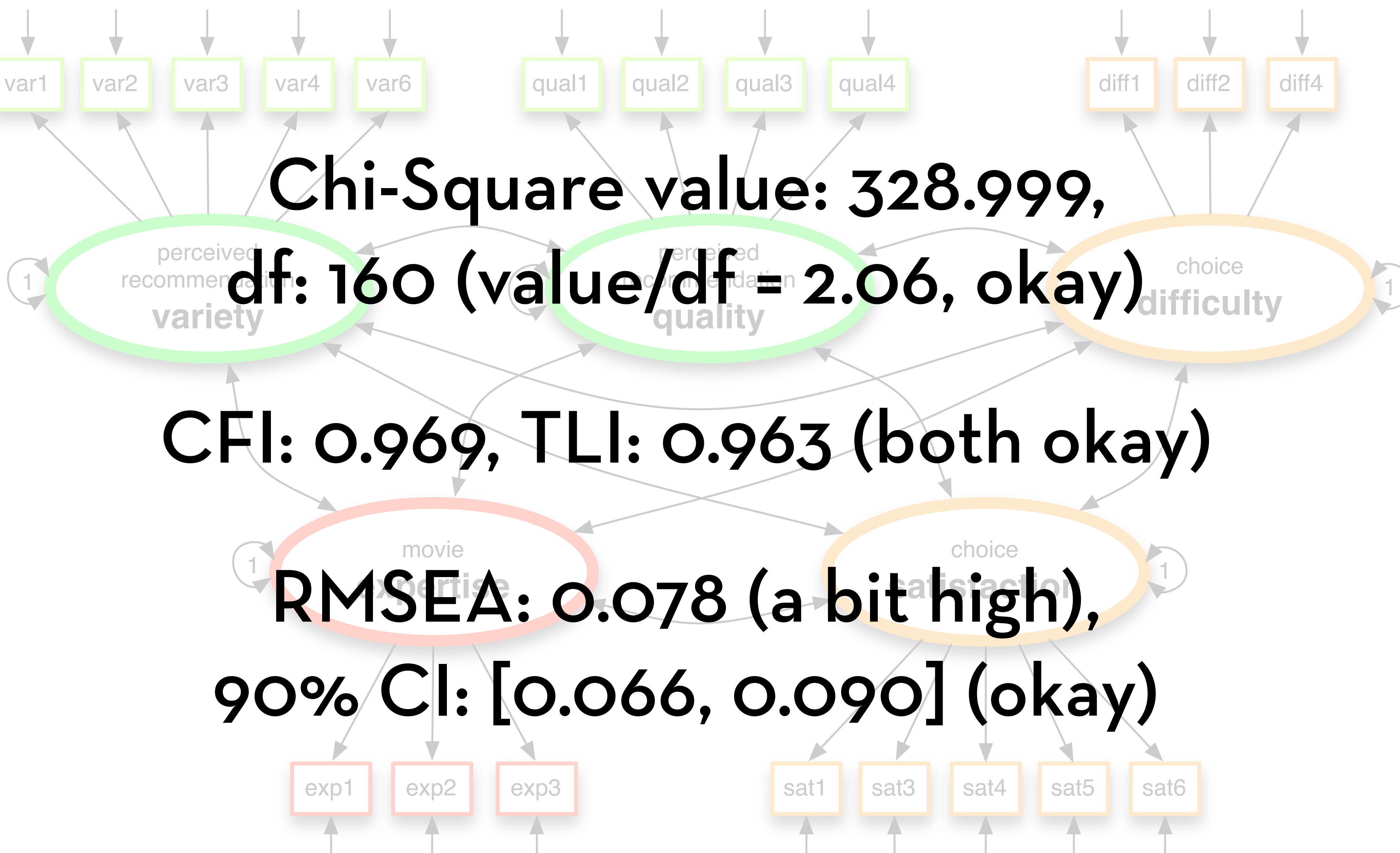
CFI and TLI:

- Relative improvement over baseline model; ranging from 0.00 to 1.00
- CFI should be > 0.96 and TLI should be > 0.95

RMSEA:

- Root mean square error of approximation
- Overall measure of misfit
- Should be < 0.05, and its confidence interval should not exceed 0.10.

Model-fit metrics



Chi-Square value: 328.999,
df: 160 (value/df = 2.06, okay)

CFI: 0.969, TLI: 0.963 (both okay)

RMSEA: 0.078 (a bit high),
90% CI: [0.066, 0.090] (okay)

“It is the mark of a truly intelligent person to be moved by statistics.”

T H A N K S !

George Bernard Shaw