Part 2: Measurement

Quantitative Research Methods Seminar

Slides

Feel free to share these slides with anyone

This is version 1.1. For the most recent version of these slides, visit www.usabart.nl/QRMS

If you want to use these slides in your own lectures, use the above link for attribution

BRENICS
INFORMATION AND COMPUTER SCIENCES

Measurement

In this part I discuss the following:

Scale selection and construction

Establishing validity

Confirmatory Factor Analysis

Bonus: Exploratory Factor Analysis

BRENICS
INFORMATION AND COMPUTER SCIENCES

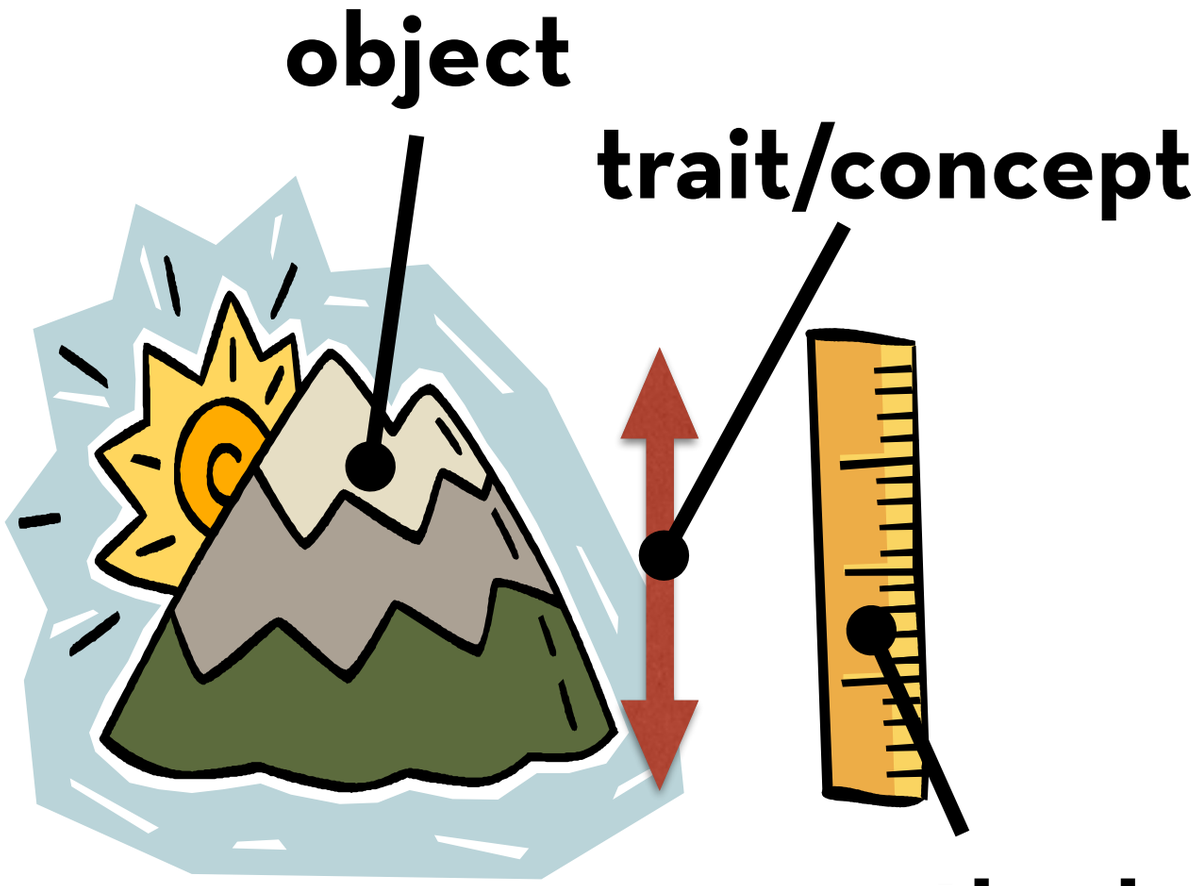
Measurement

The quantification of a trait of an object

Using a method

On a scale

Usually direct or indirect observation



BRENICS
INFORMATION AND COMPUTER SCIENCES

Psychometrics

The measurement of social and psychological concepts or traits

Rooted in the belief that these can be measured by asking questions (method)

Answers are an indirect observation on the concept/trait

Today: how to construct a proper scale

Why use a scale?

Objective traits can usually be measured with a single question

(e.g. age, income)

For subjective traits, single-item measurements lack **content validity**

Each participant may interpret the item differently

This reduces precision and conceptual clarity

Accurate measurement requires a **shared conceptual understanding** between all participants and researcher

Selection & construction

of measurement scales

Use existing scales

Why?

- Constructing your own scale is a lot of work
- "Famous" scales have undergone extensive validity tests
- Ascertains that two related papers measure exactly the same thing

Finding existing scales:

- In related work (especially if they tested them)
- The Inter-Nomological Network (INN) at inn.theorizeit.org

Create new scales

When?

- Existing scales do not hold up
- Nobody has measured what you want to measure before
- Scale relates to the specific context of measurement

How:

- Adapt existing scales to your purpose
- Develop a brand new scale (see next slides!)

Adapting scales

Information collection concerns:	System-specific concerns:
It usually bothers me when websites ask me for personal information.	It bothered me that [system] asked me for my personal information.
When websites ask me for personal information, I sometimes think twice before providing it.	I had to think twice before providing my personal information to [system].
It bothers me to give personal information to so many websites.	n/a
I am concerned that websites are collecting too much personal information about me.	I am concerned that [system] is collecting too much personal information about me.

Concept definition

Start by writing a good concept definition!

A concept definition is a careful explanation of what you want to measure

Examples: leadership

"Leadership is power, influence, and control" (objective)

"Leadership is status, respect, and authority" (subjective)

"Leadership is woolliness, foldability, and grayness" (nonsensical, but valid!)

BREN:ICS
INFORMATION AND COMPUTER SCIENCES

Concept definition

Note: They need to be more detailed than this!

A good definition makes it unambiguously clear what the concept is supposed to mean

The foundation for a shared conceptual understanding

Note 2: A concept definition is an equality relation, not a causal relation

Power, influence, control == leadership

Not: power, influence, control —> leadership

Concept definition

If a concept becomes “too broad”, split it up!

e.g. you could create separate concept definitions for power, influence, and control

If two concepts are too similar, try to differentiate them, but otherwise integrate them!

e.g. “attitude towards the system” and “satisfaction with the system” are often very similar

Creating items

E.g. Concept: “Leadership = status, respect, authority”

Find a way to measure these aspects in a leader

The respondent does not have to be the measured object!

E.g. one could ask employees to rate their supervisor

Example items:

“My supervisor is an admirable person.” (status, respect)

“I am more important than my supervisor.” (status, authority)

Creating items

Note: For objective concepts, you need to ask objective questions

E.g. behavior: “I do X” rather than “I like X”

Otherwise an exam could ask a single question:

> Do you believe that your understanding of the course materials is sufficient to pass this course?
>
> ( ) yes ( ) no

Answer categories

Most common types of items: binary, 5- or 7-point scale

Why? We want to measure the **extent** of the concept:

- Agreement (completely disagree - - - completely agree) or (no - yes)
- Frequency (never - - - very frequently)
- Importance (unimportant - - - very important)
- Quality (very poor - - - very good)
- Likelihood (almost never true - - - almost always true) or (false - true)

Answer categories

Sometimes, the answer categories represent the item

Based on what I have seen, FormFiller makes it _____ to fill out online forms.

- easy - - neutral - - difficult
- simple - - neutral - - complicated
- convenient - - neutral - - inconvenient
- effortless - - neutral - - daunting
- straightforward - - neutral - - burdensome

Answer categories

Examples:

http://www.gifted.uconn.edu/siegle/research/instrument%20reliability%20and%20validity/Likert.html

How many items?

One scale for each concept

At least 3 (but preferably 5 or more) items per scale

Developing items involves multiple iterations of testing and revising

- First develop 10–15 items
- Then reduce it to 5–7 through discussions with domain experts and comprehension pre-tests with test subjects
- You may remove 1-2 more items in the final analysis

Testing items

Experts discussion:

Card-sorting into concepts (with or without definition)

Let experts write the definition based on your items, then show them your definition and discuss difference

Comprehension pre-tests:

Also card-sorting

Think-aloud testing: ask users to 1) give an answer, 2) explain the question in their own words, and 3) explain their answer

Examples

Satisfaction:

- In most ways FormFiller is close to ideal.
- I would not change anything about FormFiller.
- I got the important things I wanted from FormFiller.
- FormFiller provides the precise functionality I need.
- FormFiller meets my exact needs.

(completely disagree - disagree - somewhat disagree - neutral - somewhat agree - agree - completely agree)

Examples

Satisfaction (alternative):

- Check-it-Out is useful.
- Using Check-it-Out makes me happy.
- Using Check-it-Out is annoying.
- Overall, I am satisfied with Check-it-Out.
- I would recommend Check-it-Out to others.

(completely disagree - disagree - somewhat disagree - neutral - somewhat agree - agree - completely agree)

BRENICS
INFORMATION AND COMPUTER SCIENCES

Examples

Satisfaction (another alternative):

I am _____ with FormFiller.

- very dissatisfied - - neutral - - very satisfied
- very displeased - - neutral - - very pleased
- very frustrated - - neutral - - very contended

Good items...

Use both positively and negatively phrased items

- They make the questionnaire less “leading”
- They help filtering out bad participants
- They explore the “flip-side” of the scale

The word “not” is easily overlooked

Bad: “The results were not very novel.”

Good: “The results felt outdated.”

BRENICS
INFORMATION AND COMPUTER SCIENCES

Good items...

Choose simple over specialized words

> Bad: “Do you find the illumination of your work environment sufficient to work in?”

Avoid double-barreled questions

> Bad: “The recommendations were relevant and fun.”

Avoid loaded or leading questions

> Bad: “Is it important to treat people fairly?”

Good items...

Avoid vague qualifiers or fuzzy words with an ambiguous meaning

Bad: “On the weekends I get down with my friends.”

Good: “I take the car for short distances (less than 7 miles).”

Avoid specificity that exceeds a respondent’s possibility for an accurate answer*

Bad: “How many minutes per day do you play games?”

Good: {give several answer categories}

Good items...

Avoid unnecessary calculations*

Bad: “How much are you willing to spend yearly on gas?”

Good: “How much are you willing to spend on a gallon of gas?”

Provide appropriate time referents*

Bad: “In the past five years, how often have you traveled for work?”

Good: “In the past three months, how often have you traveled for work?”

BREN:ICS
INFORMATION AND COMPUTER SCIENCES

Good items...

Use equal number of positive and negative response categories

Bad: yes, always - yes, sometimes - no

Good: never - very rarely - rarely - occasionally - frequently - very frequently - always

Develop mutually exclusive answer categories*

Bad: Age: 20-30, 30-40, 40-50, 50-60, 60+

Good: Age: 20-29, 30-39, 40-49, 50-59, 60+

Good items...

Avoid check-all-that-apply questions

Bad: “Which of the following cybercrimes have you been a victim of?” (check all that apply)

Good: “Have you been a victim of _____?” (yes - no)

“Undecided” and “neutral” are not the same thing

Bad: disagree - somewhat disagree - undecided - somewhat agree - agree

Good: disagree - somewhat disagree - neutral (or: neither agree nor disagree) - somewhat agree - agree

Good items...

Soften the impact of objectionable questions

Bad: "I do not care about the environment."

Good: "There are more important things than caring about the environment."

Avoid asking respondents to say "yes" in order to mean "no"

Bad: Do you favor or oppose not allowing the state to raise taxes without a 60% approval rate?

Good: Do you favor or oppose requiring a 60% approval rate in order to raise taxes?

Attention checks

Always begin with clear directions

Ask comprehension questions about the directions

Make sure your participants are paying attention!

"To make sure you are paying attention, please answer somewhat agree to this question."

"To make sure you are paying attention, please do not answer agree to this question."

Repeat certain questions

Test for non-reversals of reverse-coded questions

BRENICS
INFORMATION AND COMPUTER SCIENCES

Full example

www.uci-formfiller.com

INFORMATION AND COMPUTER SCIENCES

Learn more?

Learn it yourself:

Don Dillman, “Internet, Mail and Mixed-Mode Surveys”

Jelke Bethlehem & Silvia Biffignandi, “Handbook of Web Surveys”

BRENICS
INFORMATION AND COMPUTER SCIENCES

Establishing validity

of measurement scales

Validity in context

Note: validity is always assessed in **context**! It depends on:

- the specific **population** to be measured
- the **purpose** of the measure

BRENICS
INFORMATION AND COMPUTER SCIENCES

Types of validity

Content validity (face validity)

Criterion validity

- Predictive validity
- Concurrent validity

Construct validity

- Discriminant validity
- Convergent validity

Content validity

Content validity is assessed by specialists in the concept to be measured

Do the items cover the breath of the content area? (not too wide, not too narrow?)

Are they in an appropriate format?

Bad:

- A attitude scale that also has behavioral items
- A usability scale that only asks about learnability
- A relative measure of risk, trying to measure absolute risk

BRENICS
INFORMATION AND COMPUTER SCIENCES

Criterion validity

Predictive validity

Test how well a measure predicts a future outcome (e.g. behavioral intention —> future behavior)

Concurrent validity

Compare the measure with some other measure that is known to correlate with the concept (e.g. correlate a new scale for altruism with an existing scale for compassion)

Or, compare the measure between groups that are known to differ on the concept (e.g. compare altruism of nuns and homicidal maniacs)

Construct validity

Discriminant validity

Are two scales really measuring different things? (e.g. attitude and satisfaction may be too highly correlated)

Convergent validity

Is the scale really measuring a single thing? (e.g. a usability scale may actually consist of several sub-scales: learnability, effectiveness, efficiency, satisfaction, etc.)

Factor analysis helps you with construct validity

Other types you have to confirm yourself!

CFA

Confirmatory Factor Analysis

Why CFA?

Establish convergent and discriminant validity

CFA can suggest ways to remedy problems with the scale

Outcome is a normally distributed measurement scale

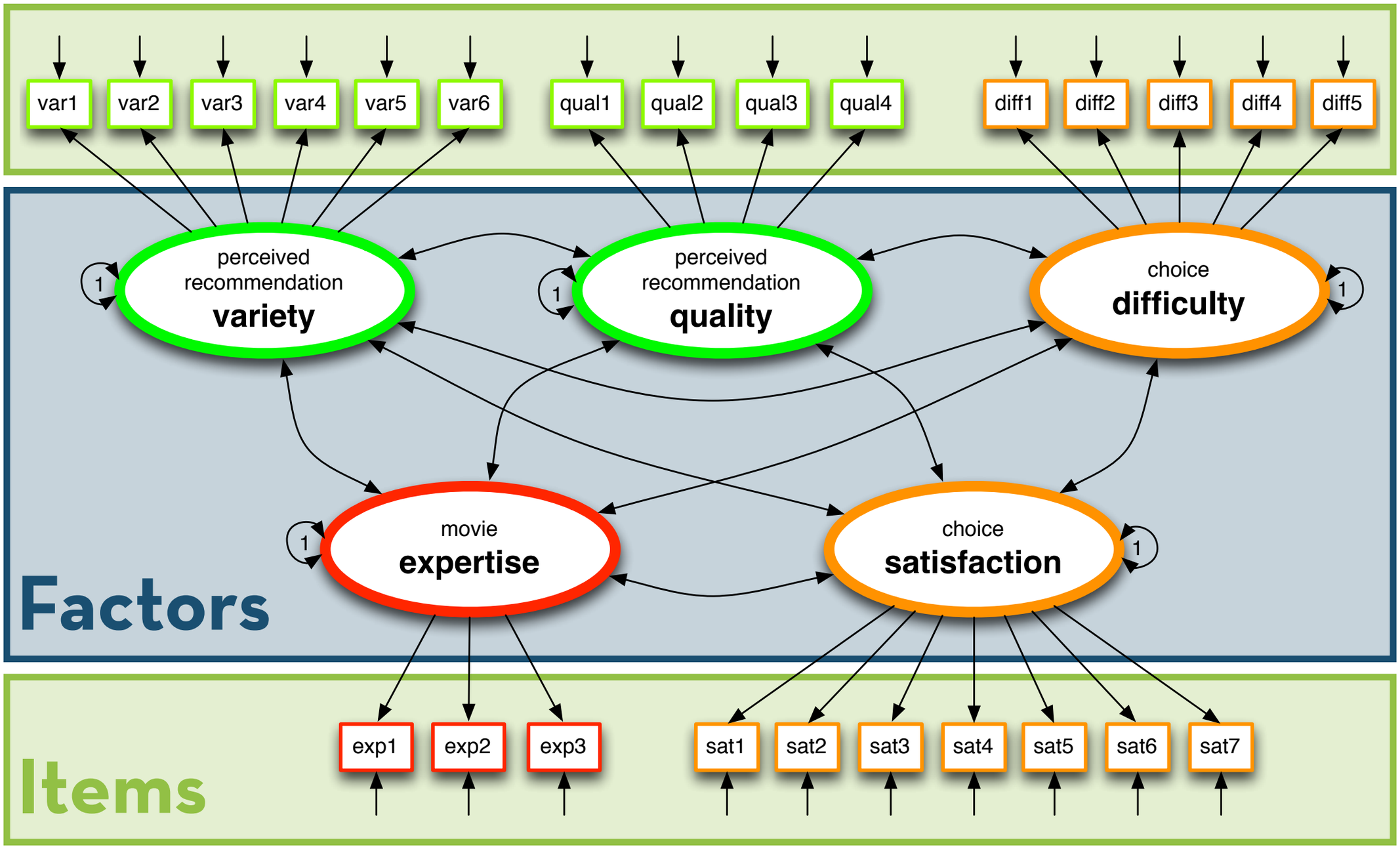
Even when the items are yes/no, 5- or 7-point scales!

The scale captures the “shared essence” of the items

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

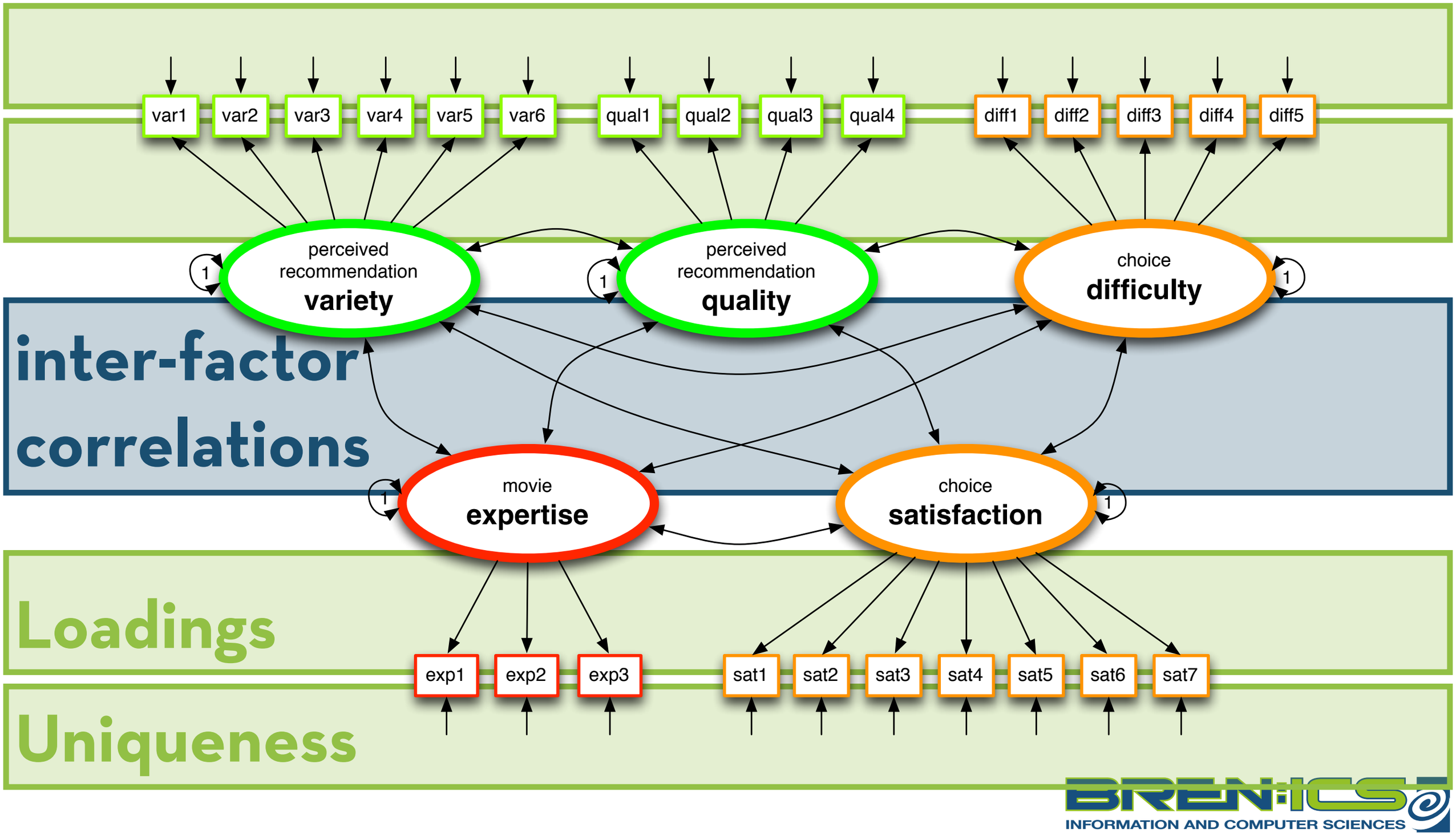
BRENICS
INFORMATION AND COMPUTER SCIENCES

CFA: the concept



BRENICS
INFORMATION AND COMPUTER SCIENCES

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)

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 basis for Factor Analysis is thus the item correlation matrix

How do we determine the loadings etc?

By **modeling** the correlation matrix as closely as possible!

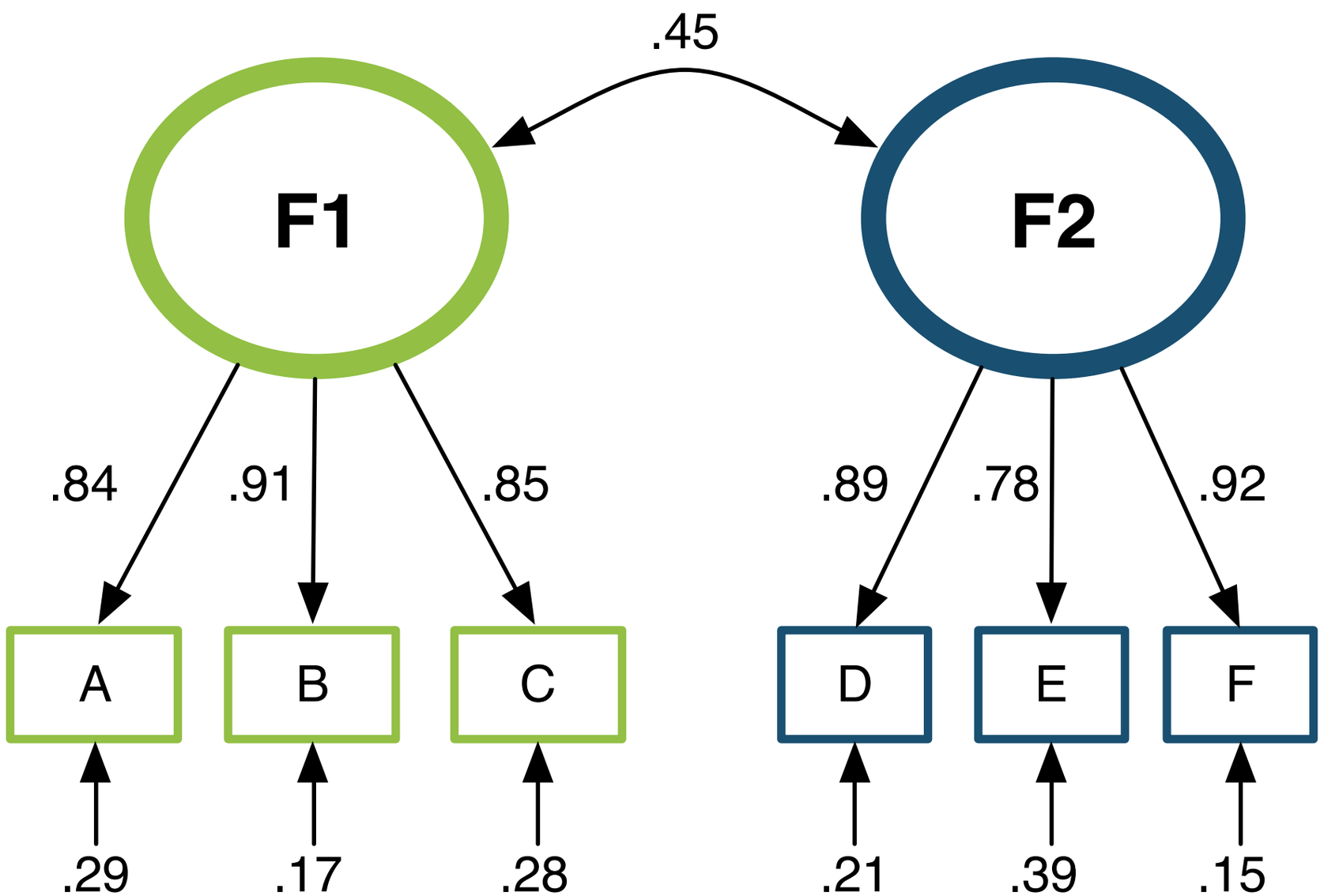
Observed

	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



BRENICS
INFORMATION AND COMPUTER SCIENCES

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

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

BREN:ICS
INFORMATION AND COMPUTER SCIENCES

Item-fit metrics

Variance extracted (squared loading):

- The amount of variance explained by the factor (1-uniqueness)
- Should be > 0.50 (although some argue 0.40 is okay)

Residual correlations:

- The observed correlation between two items is significantly higher (or lower) than predicted
- Might mean that factors should be split up

Item-fit metrics

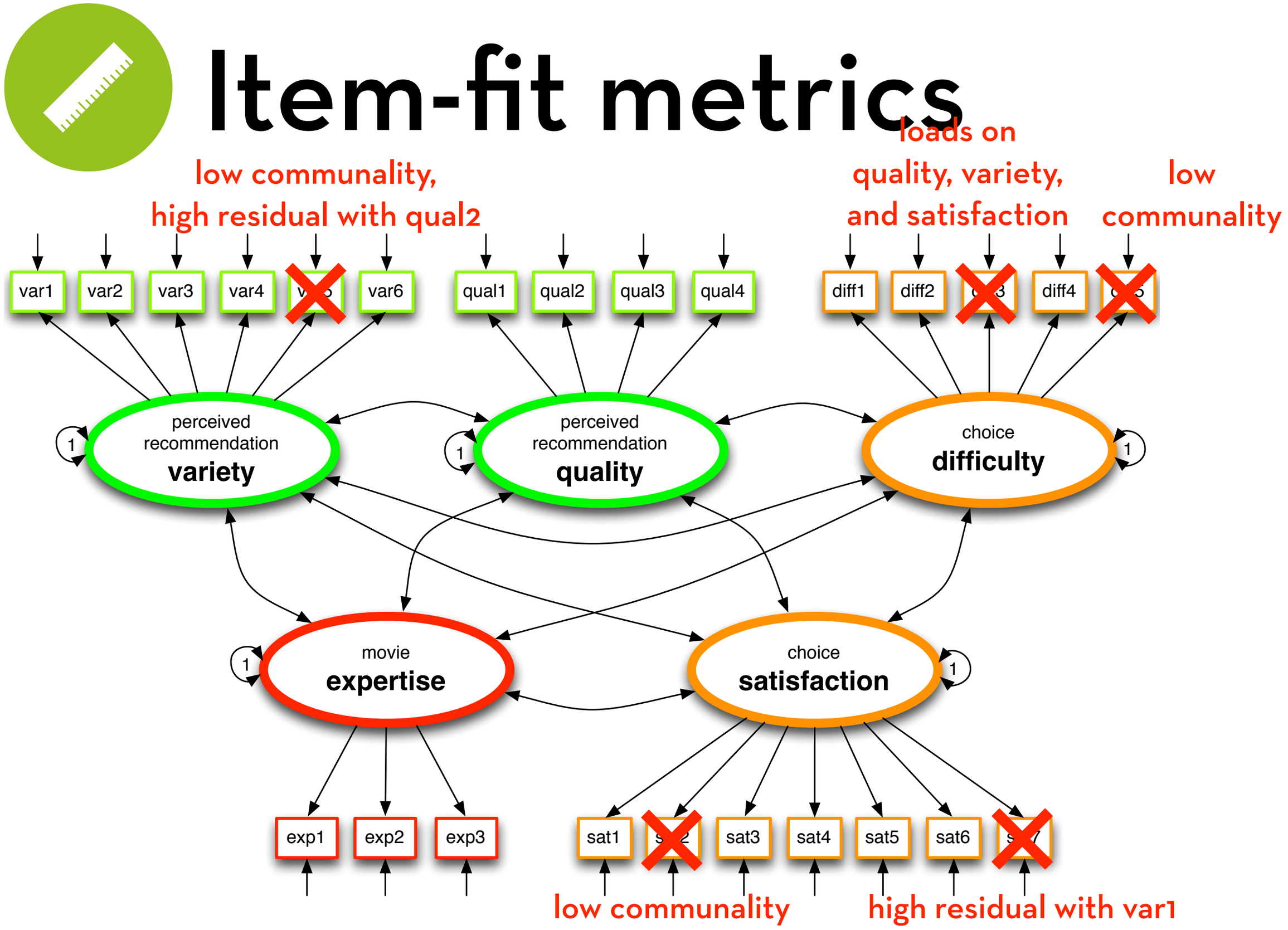
Cross-loadings:

- 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

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

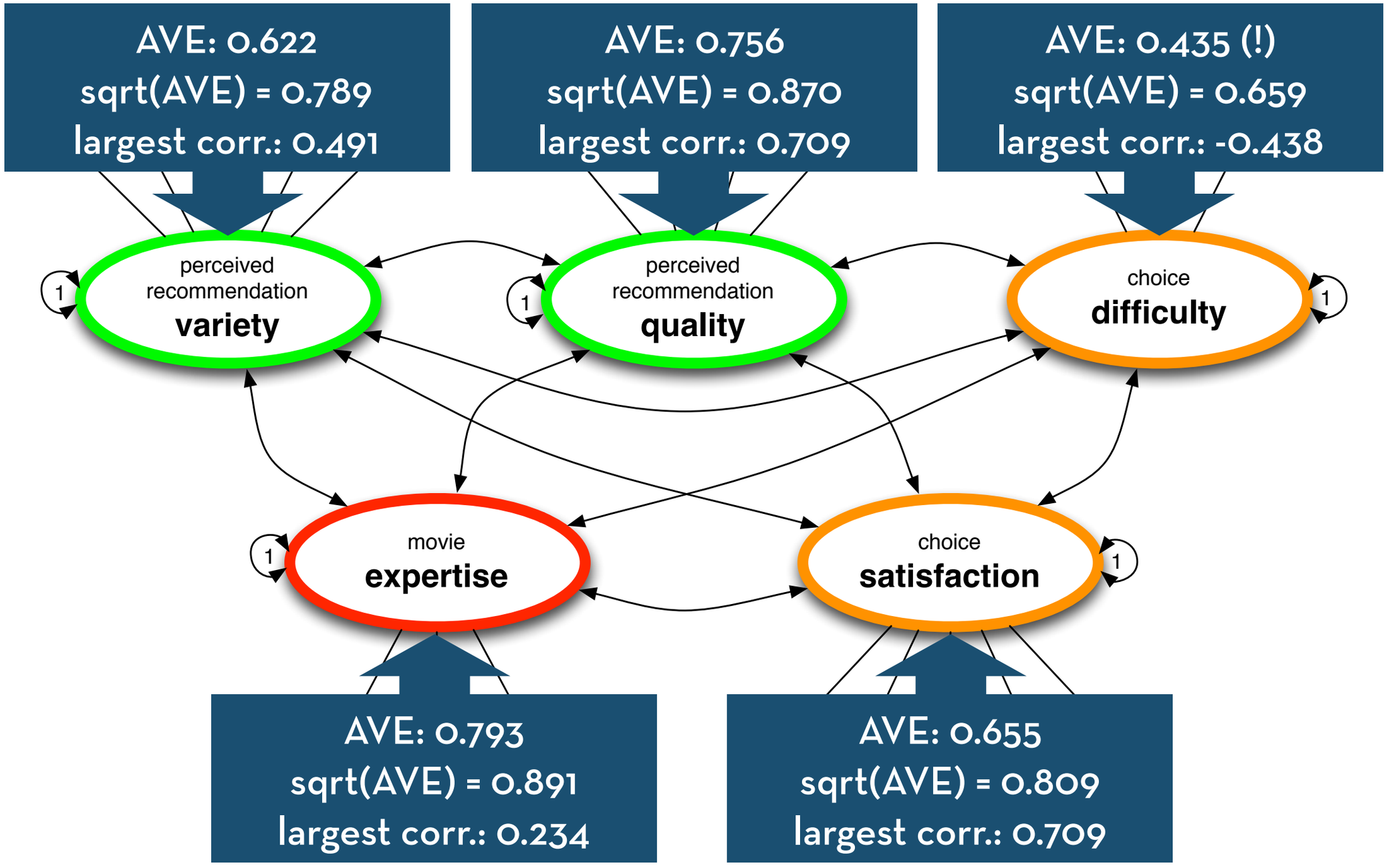


Factor-fit metrics

AVE:

- Average variance extracted (over all items per factor)
- Indicates convergent validity
- Should be > 0.50
- 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 ($p < .05$)... models are rarely perfect!
- Alternative metric: chi-squared / df < 3 (good fit) or < 2 (great fit)

Model-fit metrics

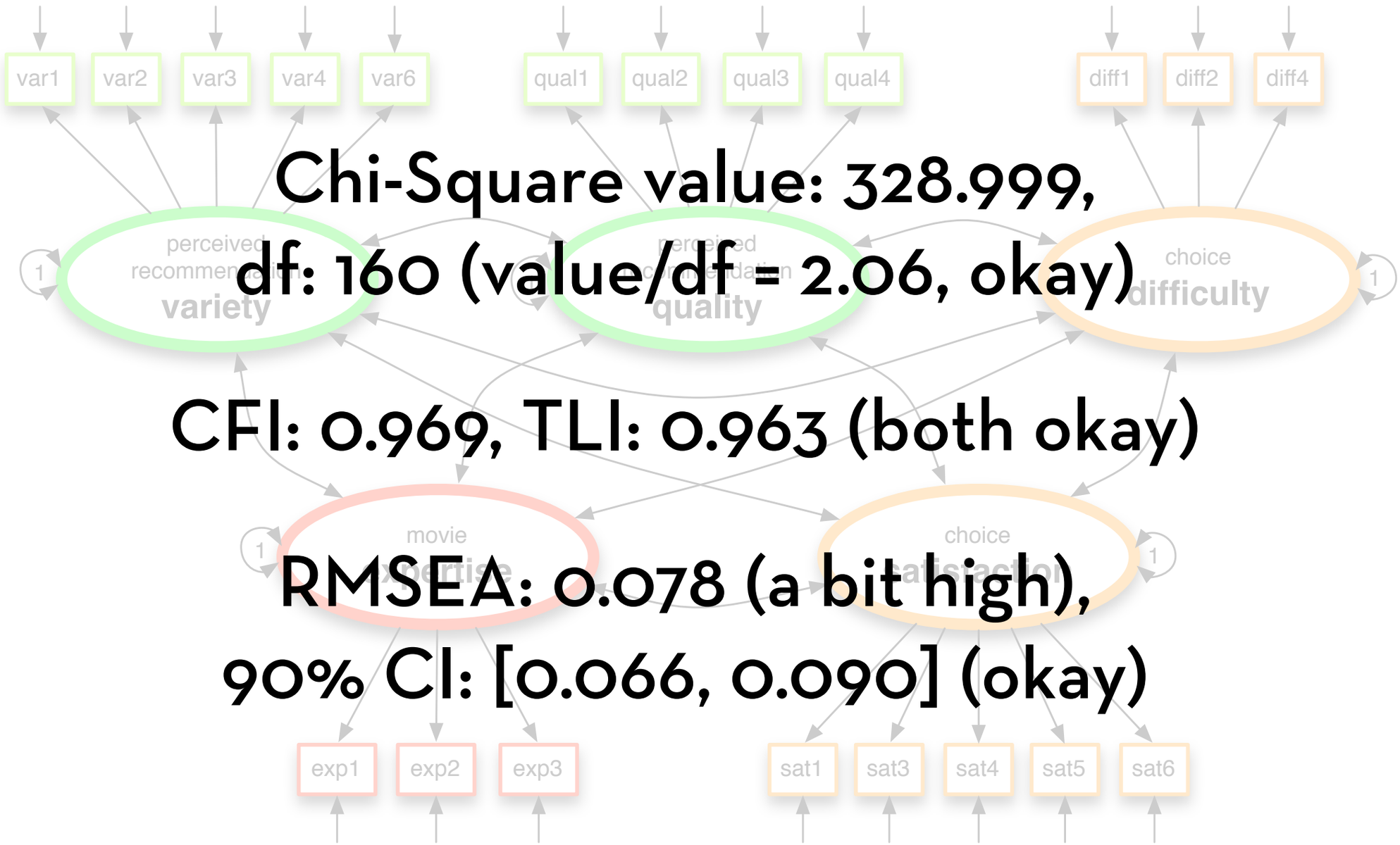
CFI and TLI:

- 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 intervall 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)

BRENICS
INFORMATION AND COMPUTER SCIENCES

Example

Confirmatory Factor Analysis in R and MPlus

Example

Effect of inspectability and control on a social recommender system

3 control conditions:

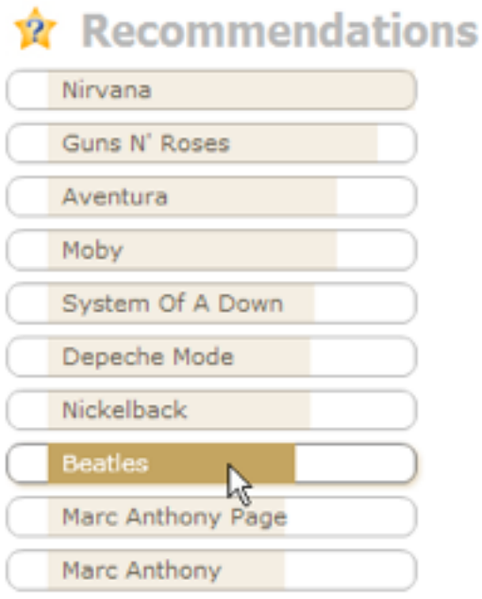
- No control (just use likes)
- Item control (weigh likes)
- Friend control (weigh friends)

BRENICS
INFORMATION AND COMPUTER SCIENCES

Example

2 inspectability conditions:

- List of recommendations vs. recommendation graph



Example

Dataset:

- s1-s7: satisfaction with the system
- q1-q6: perceived recommendation quality
- c1-c5: perceived control
- u1-u5: understandability
- cgraph: inspectability (0: list, 1: graph)
- citem-cfriend: control (baseline: no control)

Example

Construct	Item
System satisfaction	I would recommend TasteWeights to others. TasteWeights is useless. TasteWeights makes me more aware of my choice options. I can make better music choices with TasteWeights. I can find better music using TasteWeights. Using TasteWeights is a pleasant experience. TasteWeights has no real benefit for me.
Perceived Recommendation Quality	I liked the artists/bands recommended by the TasteWeights system. The recommended artists/bands fitted my preference. The recommended artists/bands were well chosen. The recommended artists/bands were relevant. TasteWeights recommended too many bad artists/bands. I didn't like any of the recommended artists/bands.
Perceived Control	I had limited control over the way TasteWeights made recommendations. TasteWeights restricted me in my choice of music. Compared to how I normally get recommendations, TasteWeights was very limited. I would like to have more control over the recommendations. I decided which information was used for recommendations.
Understandability	The recommendation process is not transparent. I understand how TasteWeights came up with the recommendations. TasteWeights explained the reasoning behind the recommendations. I am unsure how the recommendations were generated. The recommendation process is clear to me.

Example

Prepare the data (csv, space separated, …)

In RStudio:

- Import the dataset
- Install and load package ‘lavaan’
- Write model definition: model <- ‘[definition]’
- Run model: fit <- cfa(model, [params])
- Inspect model output: summary(fit, [params])

Example

Write model definition:

```
model <- 'satisf =~ s1+s2+s3+s4+s5+s6+s7
quality =~ q1+q2+q3+q4+q5+q6
control =~ c1+c2+c3+c4+c5
underst =~ u1+u2+u3+u4+u5'
```

Run model:

```
fit <- cfa(model, data=twq, ordered=names(twq))
```

Inspect model output:

```
summary(fit, rsquare=TRUE, fit.measures=TRUE)
```

Example

In MPlus:

- Remove heading row from data file
- Make a new file in MPlus with the dataset and model definition
- Save file as model.inp
- Run the model, this will create and open model.out
- Inspect model output file

Example

Write dataset and model definition:

```
DATA: FILE IS twq.datm;
VARIABLE:
 names are s1 s2 s3 s4 s5 s6 s7 q1 q2 q3 q4 q5 q6
  c1 c2 c3 c4 c5 u1 u2 u3 u4 u5 cgraph citem cfriend;

 usevariables are s1-u5;
 categorical are s1-u5;

MODEL:
 satisf  by s1-s7;
 quality by q1-q6;
 control by c1-c5;
 underst by u1-u5;
```

Scaling a factor

Factors are **latent** variables

based on a linear combination of their indicators

They have no “scale”

Their mean and variance are **arbitrary**

We don’t care about means

We only make comparisons anyway

We have to choose a variance

There are two methods for this...

BRENICS
INFORMATION AND COMPUTER SCIENCES

Scaling a factor

Method 1: set one factor loading to 1.00

- All other loadings are relative to this one
- This is useful for between-dataset variance comparisons
- Regression coefficients are harder to interpret

Method 2: standardize the factor variance to 1.00

- Regression coefficients are then standardized effects

Scaling a factor

In R, change:

```
fit <- cfa(model, data=twq, ordered=names(twq), std.lv=TRUE)
```

In MPlus, add:

```
OUTPUT:
 standardized;
```

BRENICS
INFORMATION AND COMPUTER SCIENCES

Modification indices

	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

Modification indices

With high residuals, two things can happen:

1. Items may significantly load on other factors
2. There may be significant cross-correlation

MPlus/R can automatically detect these

In R, run:

```
modindices(fit,power=TRUE)
```

In MPlus, add to the output section:

```
modindices(3.84);
```

BREN:ICS
INFORMATION AND COMPUTER SCIENCES

Improve model

Let's start with item-fit

Look at r-squared for each item (should be > 0.40)

Look at modification indices (no "large" values)

Based on r-squared, iteratively remove items:

c5 (r-squared = 0.180)

u1 (r-squared = 0.324)

Based on modification indices, remove item:

u3 loads on control (modification index = 15.287)

BRENICS
INFORMATION AND COMPUTER SCIENCES

Factor-fit

Satisfaction:

AVE = 0.709, $\sqrt{(AVE)}$ = 0.842, largest correlation = 0.762

Quality:

AVE = 0.737, $\sqrt{(AVE)}$ = 0.859, largest correlation = 0.687

Control:

AVE = 0.643, $\sqrt{(AVE)}$ = 0.802, largest correlation = 0.762

Understandability:

AVE = 0.874, $\sqrt{(AVE)}$ = 0.935, largest correlation = 0.341

BRENICS
INFORMATION AND COMPUTER SCIENCES

Model-fit

Use the “robust” column in R:

- Chi-Square value: 288.517, df: 164 (value/df = 1.76, good)
- CFI: 0.990, TLI: 0.989 (both good)
- RMSEA: 0.053 (slightly high), 90% CI: [0.043, 0.063] (ok)

BRENICS
INFORMATION AND COMPUTER SCIENCES

Summary

Specify and run your CFA

Alter the model until all remaining items fit

Make sure you have at least 3 items per factor!

Report final loadings, factor fit, and model fit

Summary

We conducted a CFA and examined the validity and reliability scores of the constructs measured in our study.

Upon inspection of the CFA model, we removed items c5 (communality: 0.180) and u1 (communality: 0.324), as well as item u3 (high cross-loadings with several other factors). The remaining items shared at least 48% of their variance with their designated construct.

Summary

To ensure the convergent validity of constructs, we examined the average variance extracted (AVE) of each construct. The AVEs were all higher than the recommended value of 0.50, indicating adequate convergent validity.

To ensure discriminant validity, we ascertained that the square root of the AVE for each construct was higher than the correlations of the construct with other constructs.

Finally, to confirm scale reliability we calculated Cronbach's alpha for each factor. Alpha scores were higher than 0.84, indicating excellent scale reliability.

BRENICS
INFORMATION AND COMPUTER SCIENCES

Summary

Construct	Item	Loading
System satisfaction Alpha: 0.92 AVE: 0.709	I would recommend TasteWeights to others. TasteWeights is useless. TasteWeights makes me more aware of my choice options. I can make better music choices with TasteWeights. I can find better music using TasteWeights. Using TasteWeights is a pleasant experience. TasteWeights has no real benefit for me.	0.888 -0.885 0.768 0.822 0.889 0.786 -0.845
Perceived Recommendation Quality Alpha: 0.90 AVE: 0.737	I liked the artists/bands recommended by the TasteWeights system. The recommended artists/bands fitted my preference. The recommended artists/bands were well chosen. The recommended artists/bands were relevant. TasteWeights recommended too many bad artists/bands. I didn't like any of the recommended artists/bands.	0.950 0.950 0.942 0.804 -0.697 -0.775
Perceived Control Alpha: 0.84 AVE: 0.643	I had limited control over the way TasteWeights made recommendations. TasteWeights restricted me in my choice of music. Compared to how I normally get recommendations, TasteWeights was very limited. I would like to have more control over the recommendations. I decided which information was used for recommendations.	0.700 0.859 0.911 0.716
Understandability Alpha: 0.92 AVE: 0.874	The recommendation process is not transparent. I understand how TasteWeights came up with the recommendations. TasteWeights explained the reasoning behind the recommendations. I am unsure how the recommendations were generated. The recommendation process is clear to me.	0.893 -0.923 0.987

Summary

Construct	Item	Loading	Response Frequencies				
			-2	-1	0	1	2
System satisfaction Alpha: 0.92 AVE: 0.709	I would recommend TasteWeights to others.	0.888	9	32	47	128	51
	TasteWeights is useless.	-0.885	99	106	29	27	6
	TasteWeights makes me more aware of my choice options.	0.768	11	43	56	125	32
	I can make better music choices with TasteWeights.	0.822	12	50	70	95	40
	I can find better music using TasteWeights.	0.889	14	45	62	109	37
	Using TasteWeights is a pleasant experience.	0.786	0	11	38	130	88
	TasteWeights has no real benefit for me.	-0.845	56	91	49	53	18
Perceived Recommendation Quality Alpha: 0.90 AVE: 0.737	I liked the artists/bands recommended by the TasteWeights system.	0.950	6	30	27	125	79
	The recommended artists/bands fitted my preference.	0.950	10	30	24	123	80
	The recommended artists/bands were well chosen.	0.942	10	35	26	101	95
	The recommended artists/bands were relevant.	0.804	4	18	14	120	111
	TasteWeights recommended too many bad artists/bands.	-0.697	104	88	45	20	10
	I didn't like any of the recommended artists/bands.	-0.775	174	61	16	14	2
Perceived Control Alpha: 0.84 AVE: 0.643	I had limited control over the way TasteWeights made recommendations.	0.700	13	52	48	112	42
	TasteWeights restricted me in my choice of music.	0.859	40	90	45	76	16
	Compared to how I normally get recommendations, TasteWeights was very limited.	0.911	36	86	53	68	24
	I would like to have more control over the recommendations.	0.716	8	27	38	130	64
	I decided which information was used for recommendations.		42	82	50	79	14
Understandability Alpha: 0.92 AVE: 0.874	The recommendation process is not transparent.		24	77	76	68	22
	I understand how TasteWeights came up with the recommendations.	0.893	8	41	17	127	74
	TasteWeights explained the reasoning behind the recommendations.		28	59	46	91	43
	I am unsure how the recommendations were generated.	-0.923	71	90	28	62	16
	The recommendation process is clear to me.	0.987	14	65	23	101	64

Summary

	Alpha	AVE	Satisfaction	Quality	Control	Underst.
Satisfaction	0.92	0.709	**0.842**	0.687	–0.762	0.336
Quality	0.90	0.737	0.687	**0.859**	–0.646	0.282
Control	0.84	0.643	–0.762	–0.646	**0.802**	–0.341
Underst.	0.92	0.874	0.336	0.282	–0.341	**0.935**

diagonal: √(AVE)

off-diagonal: correlations

BRENICS
INFORMATION AND COMPUTER SCIENCES

Learn more?

Learn it yourself:

Sections on CFA in Rex Kline, “Principles and Practice of Structural Equation Modeling”, 3rd ed.

MPlus: check the video tutorials at www.statmodel.com

BRENICS
INFORMATION AND COMPUTER SCIENCES

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

T H A N K S !

George Bernard Shaw

Bonus: EFA

Exploratory Factor Analysis

Why EFA?

In CFA, we specify the factor structure

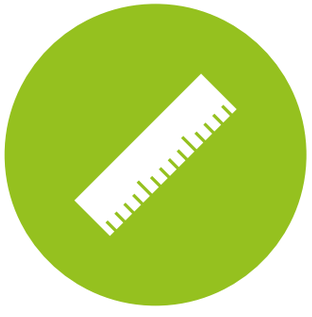
CFA will tell you how well this structure fits

CFA will give you suggestions on how to improve fit

In EFA, the factor structure is “free”

EFA will “extract” factors and then “rotate” them to fit

Effectively, it infers the structure from the data



Why EFA?

Use EFA when you have no idea about the factor structure

- E.g. semi-related behaviors (see example at the end)
- E.g. A (large) factor that didn't fit and might consist of multiple dimensions instead

Many HCI researchers use EFA instead of CFA

- Why? Because it is available in SPSS...
- Using EFA instead of CFA is a crutch
- Moreover, the default EFA settings of SPSS are almost always wrong!

BRENICS
INFORMATION AND COMPUTER SCIENCES

EFA

Steps in EFA:

Factor Extraction

Factor Rotation

Determining the number of factors

Extraction

R	A	B	C	D	E	F
A	1.00	0.48	0.44	0.52	0.28	0.24
B	0.48	1.00	0.33	0.39	0.21	0.18
C	0.44	0.33	1.00	0.47	0.35	0.30
D	0.52	0.39	0.47	1.00	0.49	0.42
E	0.28	0.21	0.35	0.49	1.00	0.42
F	0.24	0.18	0.30	0.42	0.42	1.00

Communalities

Rr	A	B	C	D	E	F
A	0.64	0.48	0.44	0.52	0.28	0.24
B	0.48	0.36	0.33	0.39	0.21	0.18
C	0.44	0.33	0.37	0.47	0.35	0.30
D	0.52	0.39	0.47	0.61	0.49	0.42
E	0.28	0.21	0.35	0.49	0.49	0.42
F	0.24	0.18	0.30	0.42	0.42	0.36

Total shared variance = sum(diagonal) = 2.83

BRENICS
INFORMATION AND COMPUTER SCIENCES

Extract factor I

Try to match Rr and explain a lot of variance

How? Several methods possible...

Factor loadings: sqrt(diagonal)

Explained variance: sum(diagonal) = 2.36

impR1	A	B	C	D	E	F
A	0.50	0.37	0.43	0.55	0.42	0.36
B	0.37	0.28	0.32	0.41	0.31	0.27
C	0.43	0.32	0.37	0.47	0.36	0.31
D	0.55	0.41	0.47	0.61	0.46	0.40
E	0.42	0.31	0.36	0.46	0.36	0.30
F	0.36	0.27	0.31	0.40	0.30	0.26

	I
A	0.704
B	0.528
C	0.607
D	0.778
E	0.596
F	0.510

BRENICS
INFORMATION AND COMPUTER SCIENCES

Subtract from Rr

resR1	A	B	C	D	E	F
A	0.14	0.11	0.01	-0.03	-0.14	-0.12
B	0.11	0.08	0.01	-0.02	-0.11	-0.09
C	0.01	0.01	0.00	0.00	-0.01	-0.01
D	-0.03	-0.02	0.00	0.00	0.03	0.02
E	-0.14	-0.10	-0.01	0.03	0.13	0.12
F	-0.12	-0.09	-0.01	0.02	0.12	0.10

Extract factor II

Try to match resR1 and explain a lot of variance

Explained variance: sum(diagonal) = 0.465

	II
A	-0.379
B	-0.284
C	-0.032
D	0.073
E	0.368
F	0.315

impR2	A	B	C	D	E	F
A	0.14	0.11	0.01	-0.03	-0.14	-0.12
B	0.11	0.08	0.01	-0.02	-0.10	-0.09
C	0.01	0.01	0.00	0.00	-0.01	-0.01
D	-0.03	-0.02	0.00	0.01	0.03	0.02
E	-0.14	-0.10	-0.01	0.03	0.14	0.12
F	-0.12	-0.09	-0.01	0.02	0.12	0.10

Subtract from resR1

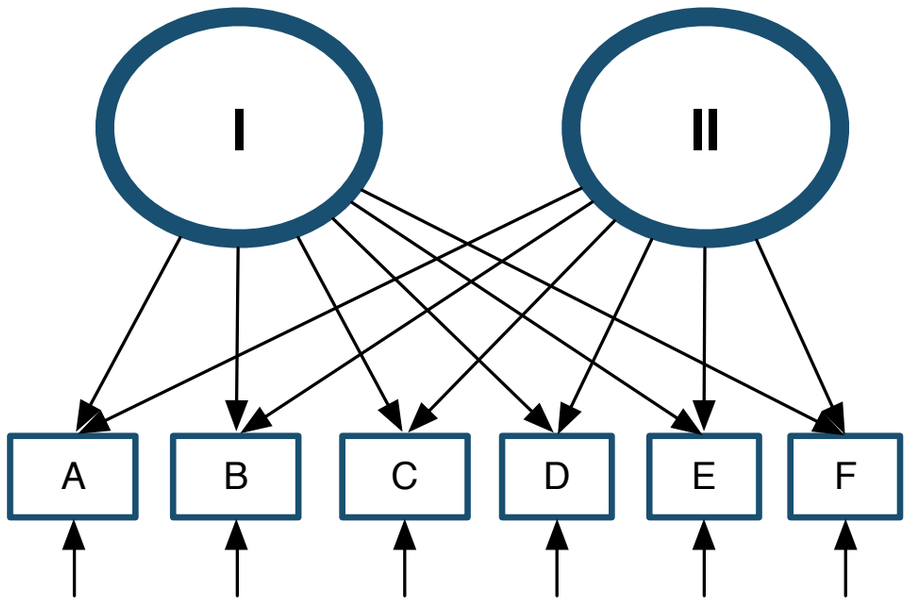
resR2	A	B	C	D	E	F
A	0.00	0.00	0.00	0.00	0.00	0.00
B	0.00	0.00	0.00	0.00	0.00	0.00
C	0.00	0.00	0.00	0.00	0.00	0.00
D	0.00	0.00	0.00	0.00	0.00	0.00
E	0.00	0.00	0.00	0.00	0.00	0.00
F	0.00	0.00	0.00	0.00	0.00	0.00

Rotation

Current solution:

Complicated! Can we simplify this?

P0	I	II
A	0.704	-0.379
B	0.528	-0.284
C	0.607	-0.032
D	0.778	0.073
E	0.596	0.368
F	0.510	0.315



Rotation

Make the solution more parsimonious by spreading the explained variance over the factors in a “smart” way

- So that each item loads only on one factor, as much as possible
- Solution does not improve, just becomes easier to interpret!

Two methods:

- Orthogonal (no correlations between factors allowed)
- Oblique (correlations allowed)

Orthogonal

Multiply P0 with a transformation matrix T

TT' = I, so the explained variance remains the same

How? Different methods exist

P0	I	II
A	0.704	-0.379
B	0.528	-0.284
C	0.607	-0.032
D	0.778	0.073
E	0.596	0.368
F	0.510	0.315

→

T	1	2
I	0.736	0.677
II	-0.677	0.736

Varimax

→

P0	1	2
A	0.78	0.20
B	0.58	0.15
C	0.47	0.39
D	0.52	0.58
E	0.19	0.67
F	0.16	0.58

Oblique

Multiply P0 with a transformation matrix T, and inter-factor correlation matrix F

$$TFT' = I$$

P0	I	II
A	0.704	-0.379
B	0.528	-0.284
C	0.607	-0.032
D	0.778	0.073
E	0.596	0.368
F	0.510	0.315

→

T	1	2
I	0.575	0.555
II	-1.071	1.081

Oblimin

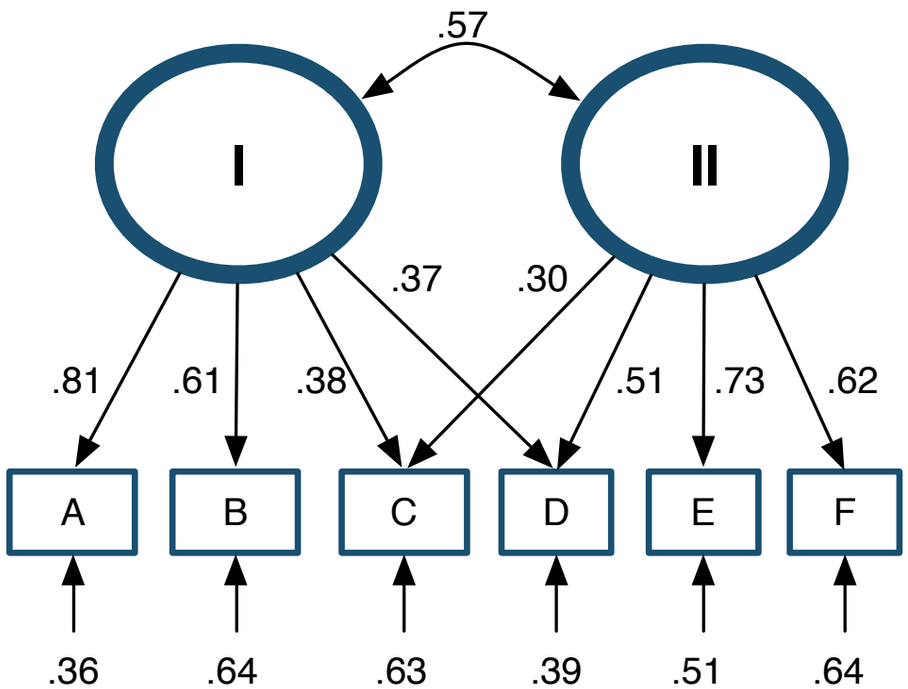
→

P0	1	2
A	0.81	0.02
B	0.61	-0.01
C	0.38	0.30
D	0.37	0.51
E	-0.05	0.73
F	-0.04	0.62

+

F	1	2
I	1.00	0.57
II	0.57	1.00

Final result



P0	1	2
A	0.81	0.02
B	0.61	-0.01
C	0.38	0.30
D	0.37	0.51
E	-0.05	0.73
F	-0.04	0.62

+

F	1	2
I	1.00	0.57
II	0.57	1.00

BRENICS
INFORMATION AND COMPUTER SCIENCES

Number of factors?

Method 1 (quick):

Obtain the communalities

In MPlus, add this to run a simple 1-factor model:

```
ANALYSIS: type = efa 1 1;
```

Then, look for "eigenvalues for sample correlation matrix"

Build a "scree plot" of communalities

Find the inflection point

Number of factors?

Method 2 (thorough):

Run with increasing number of factors

In MPlus, add this to run a simple 1-factor model:

```
ANALYSIS: type = efa 1 x;
```

Where x is higher than the number of factors you expect there to be

Find minimum of BIC, inflection in loglikelihood (LL) levels, and non-significant improvements (use a –2LL test)

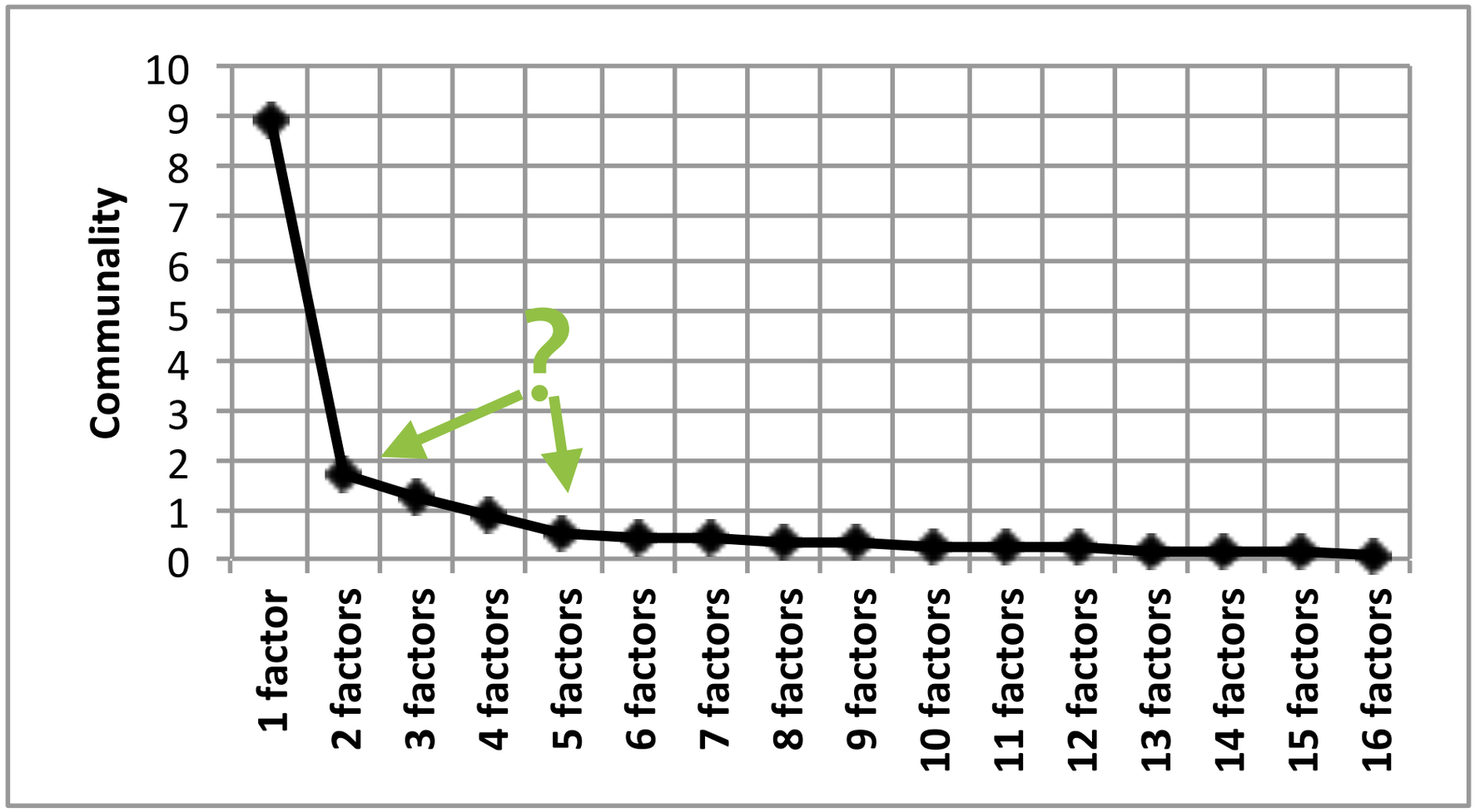
(see MPlus tutorials for details on the –2LL test)

BRENICS
INFORMATION AND COMPUTER SCIENCES

Example

ID	Items
1	Wall
2	Status updates
3	Shared links
4	Notes
5	Photos
6	Hometown
7	Location (city)
8	Location (state/province)
9	Residence (street address)
10	Employer
11	Phone number
12	Email address
13	Religious views
14	Interests (favorite movies, etc.)
15	Facebook groups
16	Friend list

Example



Example

Table 7
A comparison of the fit of different factor solutions.

	BIC	LL	# of par.	*p*-Value
1 factor	20,611	−10164.489	48	
2 factors	20,207	−9918.105	63	< 0.001
3 factors	19,574	−9560.411	77	< 0.001
4 factors	**19,320**	−9395.040	90	< 0.001
5 factors	19,360	−9379.961	102	**0.237**
6 factors	19,402	−9368.779	113	0.428

The bold values are mentioned in the text as indicators of the optimal number of dimensions.

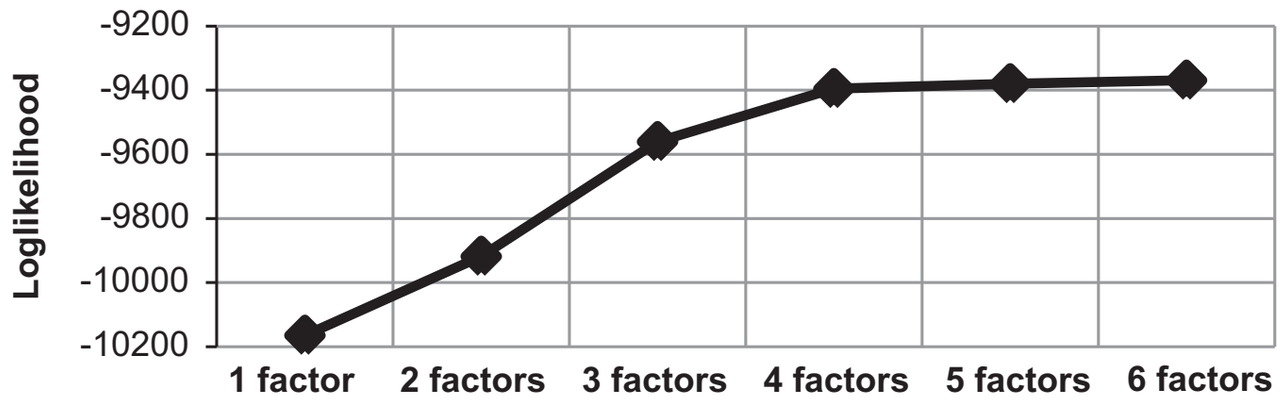


Fig. 7. Change in loglikelihood between subsequent factor solutions.

Example

GEOMIN ROTATED LOADINGS (* significant at 5% level)

	1	2	3	4
CWALL	0.801*	0.023	-0.011	-0.027
CSTATUS	0.934*	0.001	0.005	0.012
CLINKS	0.777*	-0.024	-0.022	0.150*
CNOTES	0.783*	0.010	0.129*	0.028
CPHOTO	0.568*	0.206*	0.144*	0.009
CTOWN	0.168*	0.683*	0.007	0.117*
CLOCCITY	-0.006	0.960*	0.043	-0.016
CLOCSTAT	0.041	0.943*	-0.042	0.004
CLOCADRE	0.081	0.118*	0.742*	-0.081*
CEMPLOYE	-0.134	0.302*	0.398*	0.301*
CPHONE	0.001	-0.033	0.928*	0.003
CEMAIL	0.068	-0.029	0.642*	0.226*
CRELIGIO	-0.026	-0.060	0.040	0.795*
CINTERES	0.095	0.019	-0.036	0.841*
CGROUPS	0.181	0.050	-0.014	0.741*
CFRIENDS	0.332*	0.098	0.038	0.457*

BREN:ICS
INFORMATION AND COMPUTER SCIENCES

Example

Type of data	ID	Items
Facebook activity	1	Wall
	2	Status updates
	3	Shared links
	4	Notes
	5	Photos
Location	6	Hometown
	7	Location (city)
	8	Location (state/province)
Contact info	9	Residence (street address)
	11	Phone number
	12	Email address
Life/interests	13	Religious views
	14	Interests (favorite movies, etc.)
	15	Facebook groups