Part 1: Introduction

Quantitative Research Methods Seminar

Part 1: Introduction

My goal:

Teach how to scientifically evaluate systems* using a quantitative user-centric approach

My approach:

- Intro to user-centric evaluation
- Why standard methods are insufficient
- Questionnaire construction and analysis with CFA
- Analyzing mediated regression paths with SEM
- Advanced topics (if we get to them)

Slides

Feel free to share these slides with anyone

This is version 1.3. 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

User Evaluation

An introduction

User Evaluation

A scientific method to investigate factors that influence how people interact with systems*

Systems can be anything:

- Software
- Hardware
- Other people
- Organizations
- Policies

Introduction

My goal:

Teach how to scientifically evaluate systems using a user-centric approach

How? User experiments! (and sometimes surveys)

My approach:

- I will provide a broad theoretical framework
- I will cover every step in conducting a user experiment
- I will teach the “statistics of the 21st century”

What to ask?

“Can you test if my system is good?”

Problem...

What does **good** mean?

- Learnability? (e.g. number of errors?)
- Efficiency? (e.g. time to task completion?)
- Usage satisfaction? (e.g. usability scale?)
- Outcome quality? (e.g. survey?)

We need to define **measures**

Better...

"Can you test if the user interface of my system scores high on this usability scale?"

However...

What does **high** mean?

Is 3.6 out of 5 on a 5-point scale “high”?

What are 1 and 5?

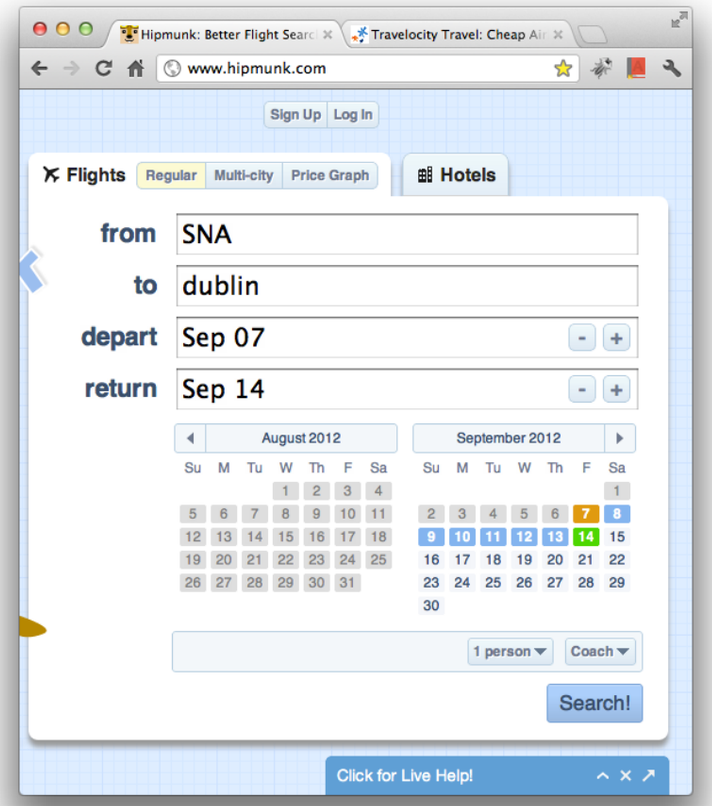
What is the difference between 3.6 and 3.7?

We need to **compare** the UI against something

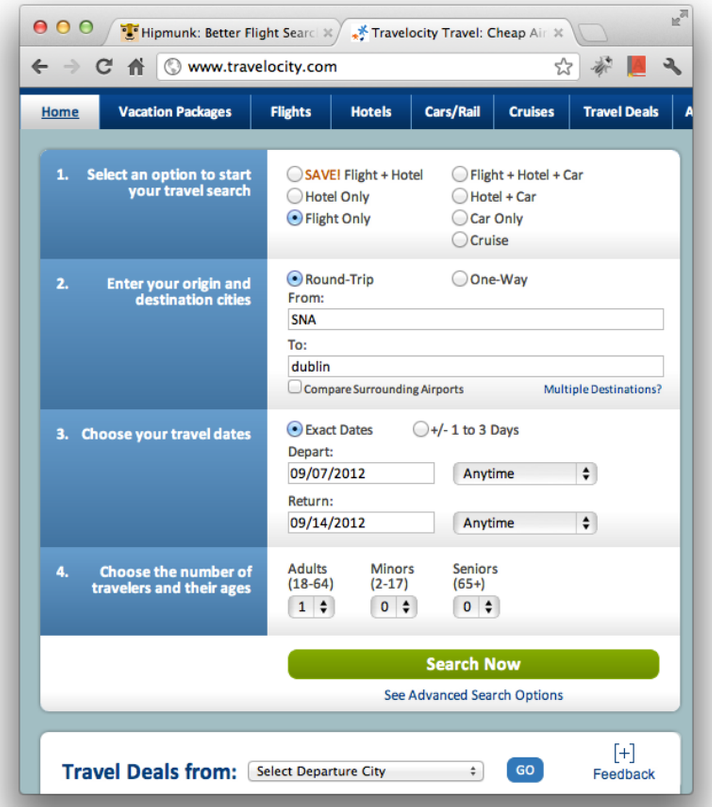
Even better...

"Can you test if the UI of my system scores high on this usability scale compared to this other system?"

Testing A vs. B



My new travel system



Travelocity

However...

Say we find that it scores higher on usability... **why** does it?

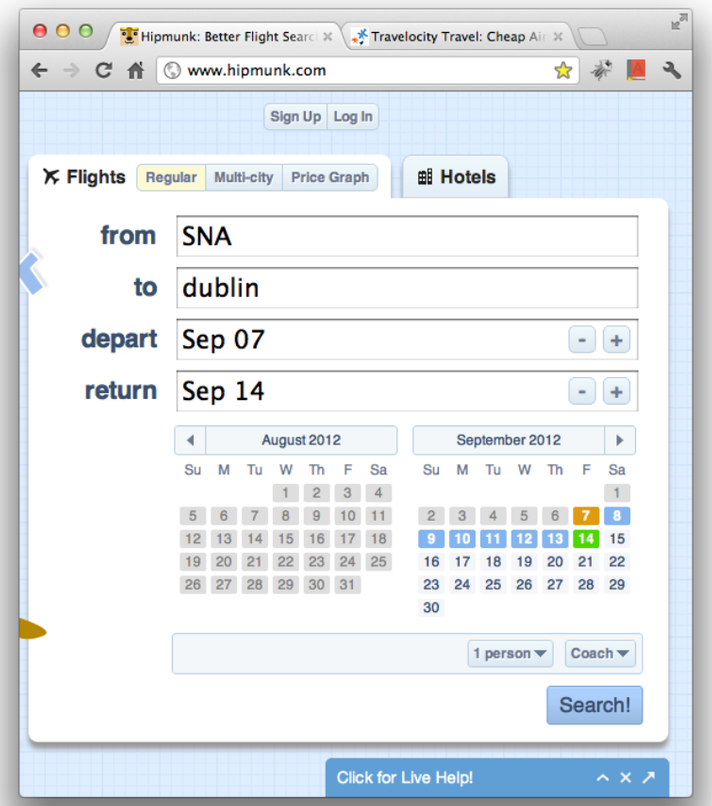
- different date-picker method
- different layout
- different number of options available

Apply the concept of **ceteris paribus** to get rid of confounding variables

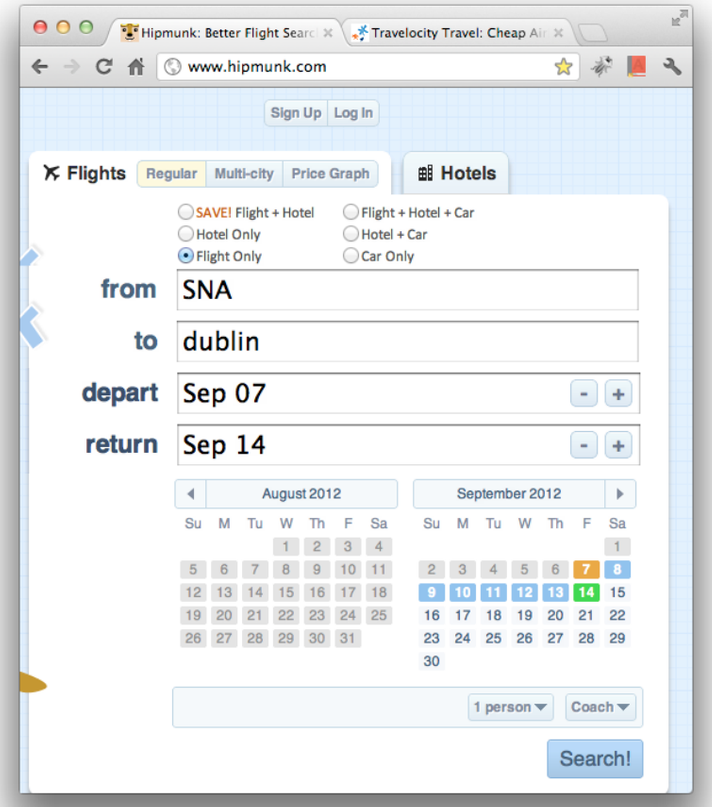
Keep everything the same, except for the thing you want to test (the manipulation)

Any difference can be attributed to the manipulation

Ceteris Paribus



My new travel system



Previous version
(too many options)

Survey/observation

What is the difference between men and women in Facebook usage satisfaction?

Downsides:

Purely observational

- No manipulations!
- What causes what?

No ceteris paribus

- Hard to get rid of confounding variables

Summary

"A user experiment systematically tests how different system aspects (manipulations) influence the users' experience and behavior (observations)."

"A survey systematically tests how certain aspects of the user (observations) influence the users' experience and behavior (observations)."

Participants

Population and sampling

Participants

"We are testing our system
on our colleagues/students."

-or-

"We posted the study link
on Facebook/Twitter."

Sampling

Are your connections, colleagues, or students **typical** users of your system?

- They may have more knowledge of the field of study
- They may feel more excited about the system
- They may know what the experiment is about
- They probably want to please you

You should sample from your **target population**

An unbiased sample of users of your system

Limiting scope

“We only use data from frequent users.”

Limiting scope

What are the consequences of **limiting** your scope?

You run the risk of catering to that subset of users only

You cannot make generalizable claims about users

For scientific experiments, the target population may be **unrestricted**

Especially when your study is more about human nature than about a specific system

Sample size

"We tested our system with 10 users."

Sample size

Is this a decent **sample size**?

Can you attain statistically significant results?

Does it provide a wide enough inductive base?

Make sure your sample is **large enough**

40 is typically the bare minimum

Anticipated effect size	Needed sample size
small	385
medium	54
large	25

Crowd-sourcing

Craigslist:

- Post in various cities under Jobs > Etcetera
- Create a geographically balanced sample

Amazon Mechanical Turk

- Often used for very small tasks, but Turk workers appreciate more elaborate studies
- Anonymous payment facilities.
- Set criteria for workers (e.g. U.S. workers with a high reputation)

Crowd-sourcing

Demographics reflect the general Internet population

- Craigslist users: a bit higher educated and more wealthy
- Turk workers: less likely to complain about tedious study procedures, but are also more likely to cheat

Make your study simple and usable

Use quality checks, add an open feedback item to catch unexpected problems

Manipulations

Testing A versus B

Manipulations

“Are our users more satisfied if our news recommender shows only recent items?”

Choosing a baseline

Proposed system or **treatment**:

Filter out any items > 1 month old

What should be my **baseline**?

- Filter out items < 1 month old?
- Unfiltered recommendations?
- Filter out items > 3 months old?

You should test against a **reasonable alternative**

"Absence of evidence is not evidence of absence"

Randomization

"The first 40 participants will get the baseline, the next 40 will get the treatment."

Randomization

These two groups cannot be expected to be **similar**!

Some news item may affect one group but not the other

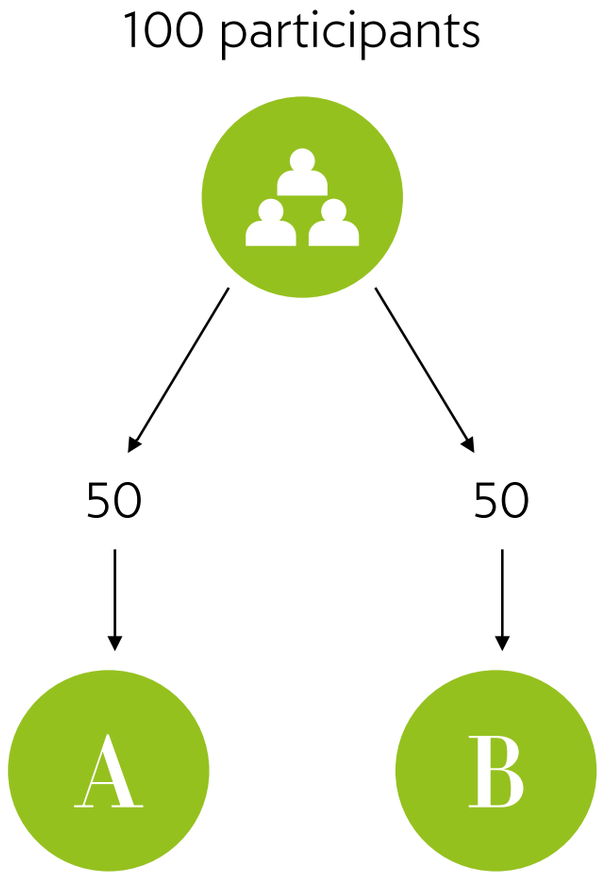
Randomize the assignment of conditions to participants

Randomization neutralizes (but doesn't eliminate) participant variation

Between-subjects

Randomly assign half the participants to A, half to B

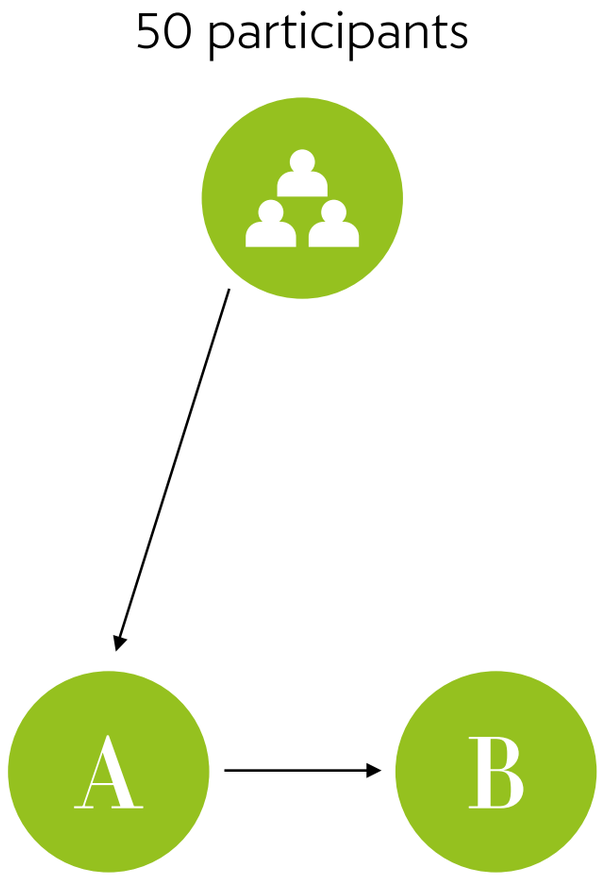
- Realistic interaction
- Manipulation hidden from user
- Many participants needed



Within-subjects

Give participants A first, then B

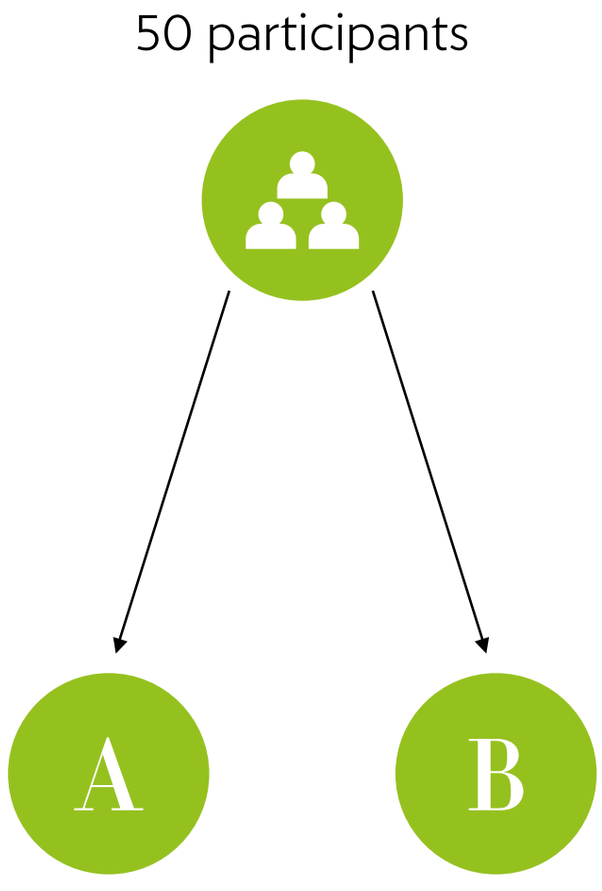
- Remove subject variability
- Participant may see the manipulation
- Spill-over effect



Within-subjects

Show participants A and B simultaneously

- Remove subject variability
- Participants can compare conditions
- Not a realistic interaction



Which one?

Should I do within-subjects or between-subjects?

Use **between-subjects** designs for user experience

- Closer to a real-world usage situation
- No unwanted spill-over effects

Use **within-subjects** designs for psychological research

- Effects are typically smaller
- Nice to control between-subjects variability

Factorial designs

You can test multiple manipulations in a **factorial design**

The more conditions, the **more participants** you will need!

	Low diversity	High diversity
5 items	5+low	5+high
10 items	10+low	10+high
20 items	20+low	20+high

Hawthorne effect

Beware of the **Hawthorne** effect

Participants may change their behavior just because they know they are being observed

When in doubt, triangulate!

Do standard AB-testing as well

Compare behavior between AB test and experiment

Placebo effect

Let's test an algorithm against random recommendations

- What should we tell the participant?

Beware of the **Placebo** effect!

- Remember: ceteris paribus!
- Other option: manipulate the message (factorial design)

Standard Methods

...and their use in experiments and surveys

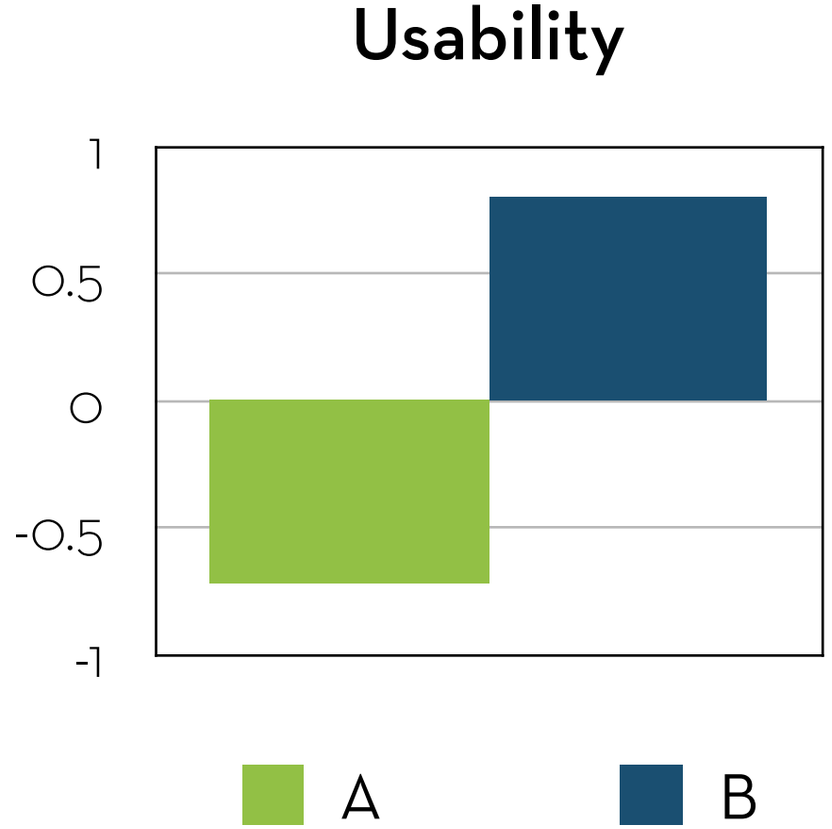
T-test

Difference between two systems:

Do these two UIs (A and B) lead to a different level of usability?

Differences between two groups of people:

Do men (A) and women (B) perceive different levels of usability?



T-test example

Usability for users of system A:

3, 2, 3, 4, 1

Usability for users of system B:

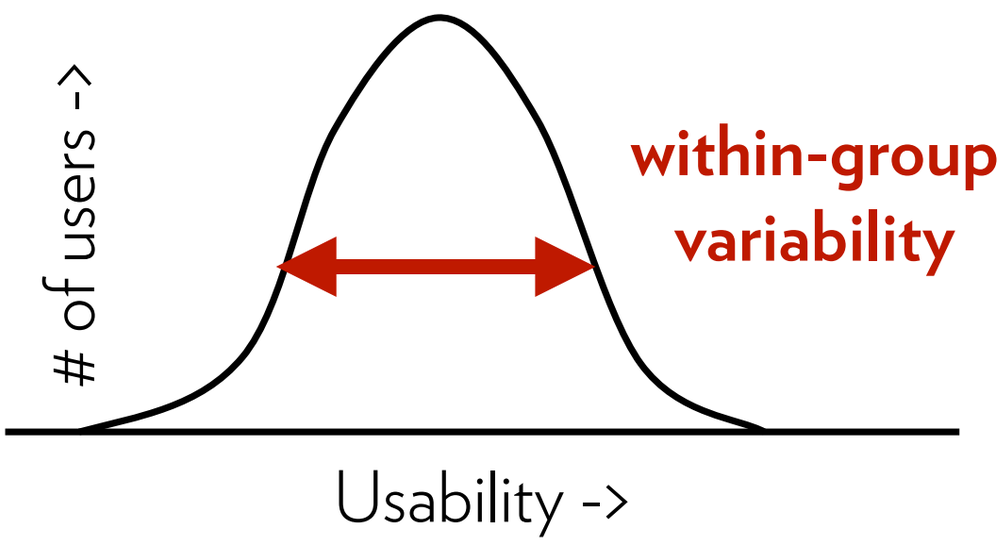
5, 4, 5, 4, 5

Which system is more usable?

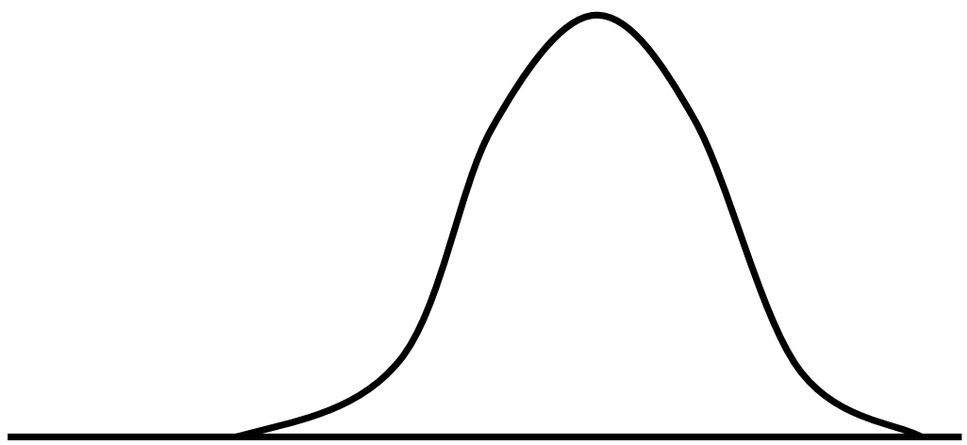
Is this difference significant?

T-test concept

Usability for users of **system A:**

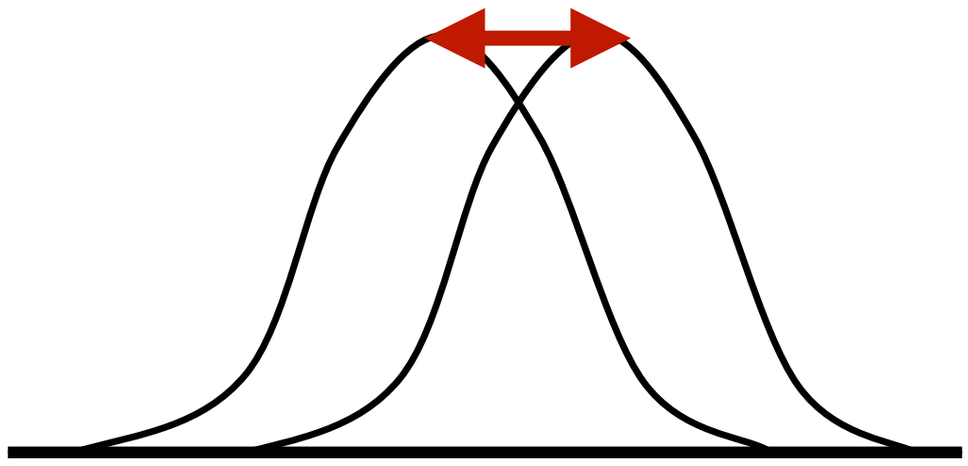


Usability for users of **system B:**



T-test concept

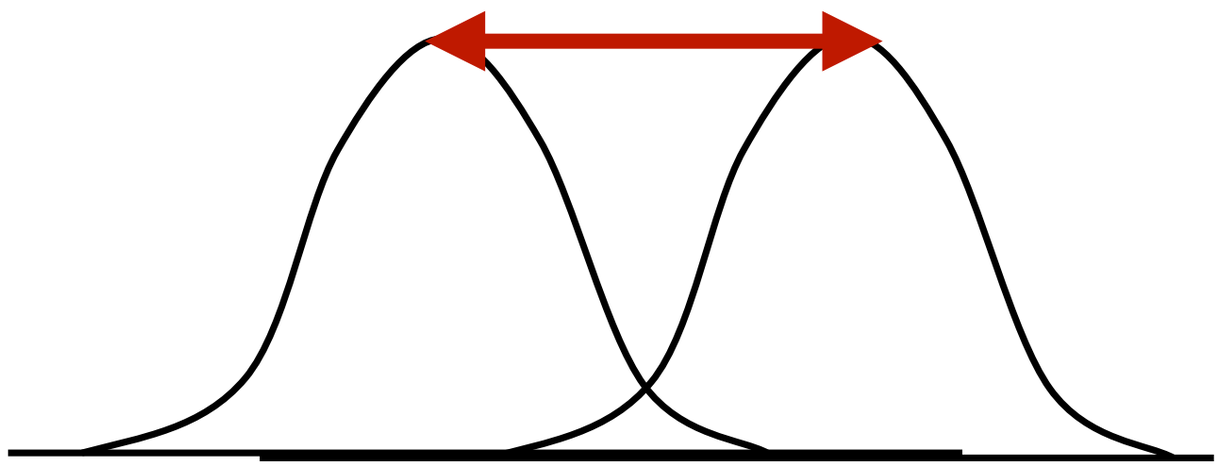
between-group
variability = small



effect is likely due to chance

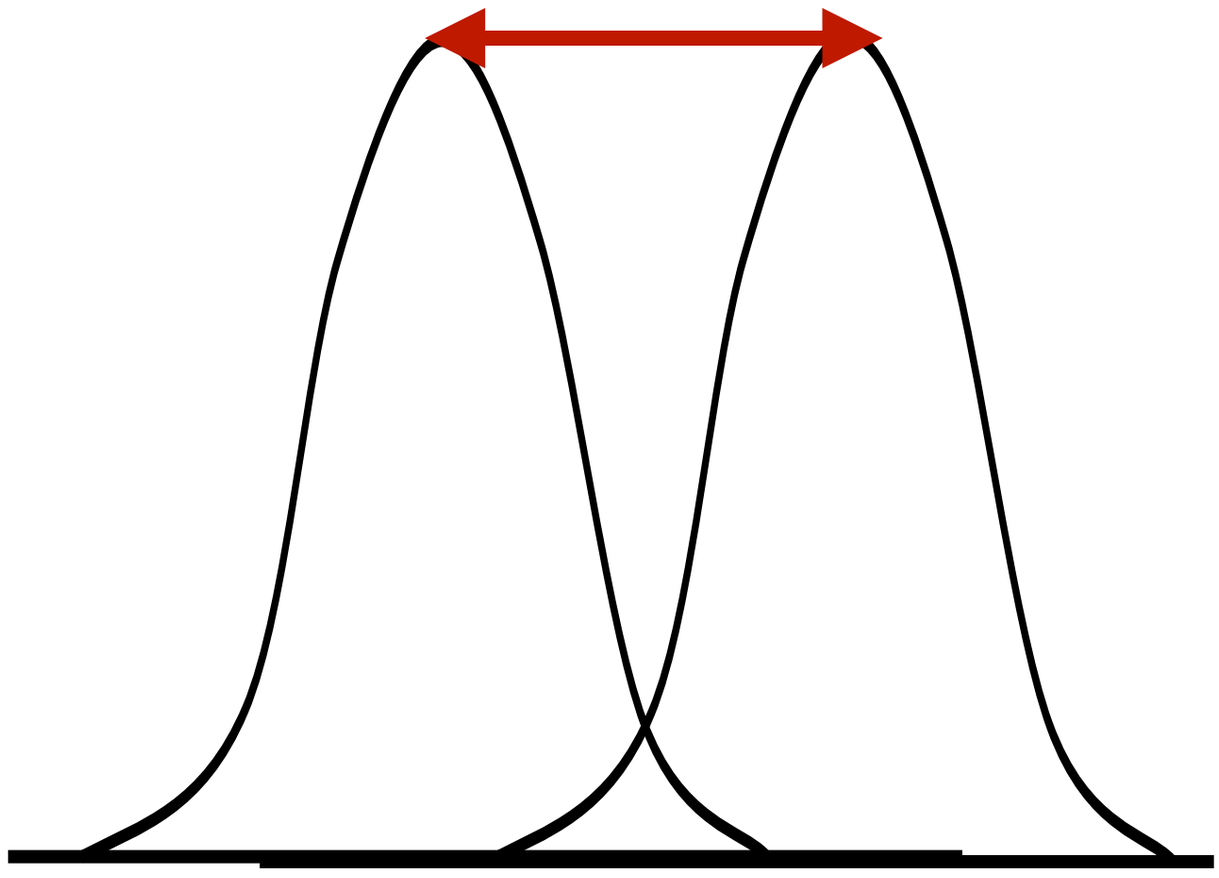
T-test concept

between-group
variability = large



effect is likely due to manipulation

T-test concept



more data = stronger test

T-test example

T-test: compare the difference in means (M) with the variability (V) and size (N) of the sample

$$t = (Ma - Mb)/\sqrt{(Va/Na + Vb/Nb)}$$

For our example:

Ma = 2.6, Va = 1.3, Na = 5

Mb = 4.6, Vb = 0.3, Nb = 5

t = 3.53, p = 0.01317

T-test example

In RStudio:

- Import the dataset
- Run the t-test:

```
t.test(usability~system, data=example)
```

- Inspect the output:

```
t = -3.5355, df = 5.753, p-value = 0.01317
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -3.3987283 -0.6012717
sample estimates:
mean in group X  mean in group Y
            2.6              4.6
```

T-test example

What does the p-value mean?

> The probability of observing this difference (or more extreme) if in reality there is no difference at all

What if the p-value is large?

> We cannot reject the null hypothesis (no difference)

What if the p-value is equal to or smaller than the significance level (we usually take 0.05)?

> We reject the null hypothesis

T-test example

In this specific case, the chance of observing a difference of 2.00 if there is no difference in reality, is very small (p = .013)

Hence we reject this null hypothesis...

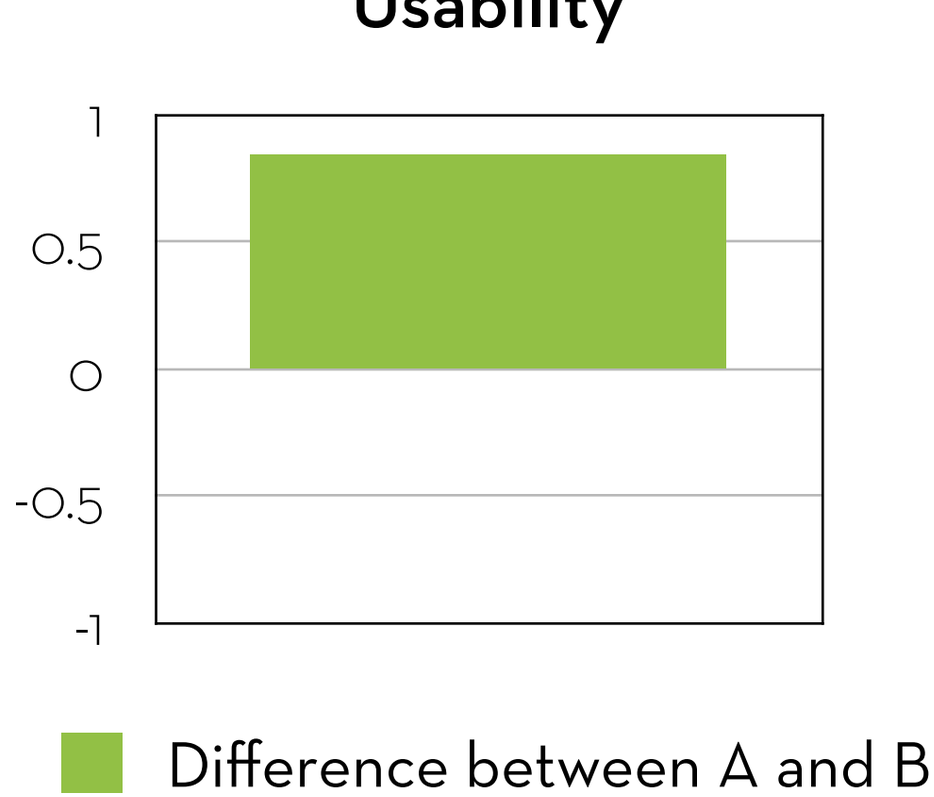
...and we take the result as evidence that system B may be more usable than system A

Note: this is evidence, not proof!

1-sample t-test

Difference between two systems, tested by the same user

Differences in user evaluation of Facebook vs. Google Plus



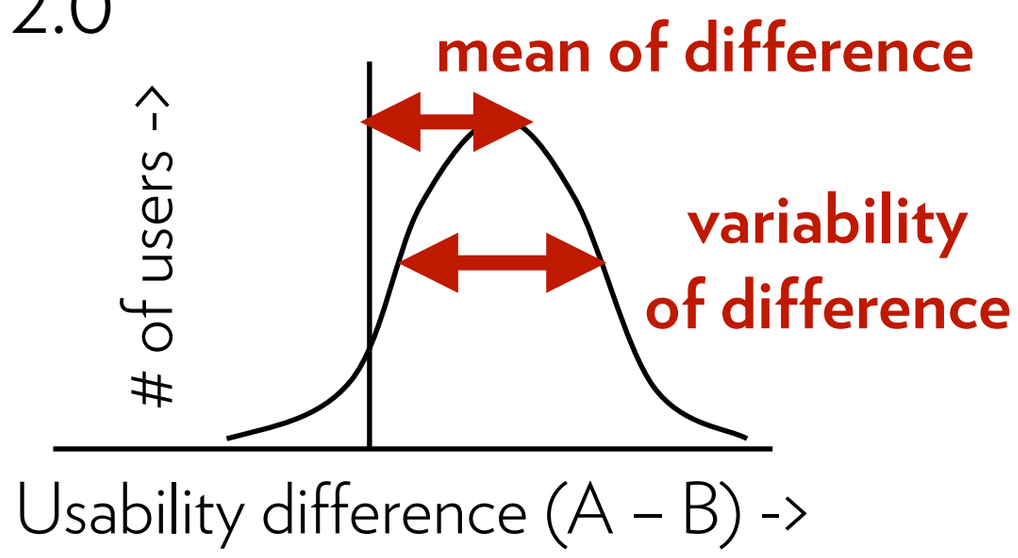
1-sample t-test

Participant uses system A —> usability evaluation: 4.0

Participant uses system B —> usability evaluation: 2.0

Calculate the difference: 2.0

Tabulate all differences:



T-test example

T-test: compare the difference (D) with the variability (Vd) and size (N) of the sample

$$t = (D)/\sqrt{(Vd/N)}$$

For our example:

D = 2.0, Vd = 2.0, N = 5

t = 3.16, p = 0.034

T-test example

In RStudio:

- Import the dataset
- Run the t-test:

```
t.test(example2$usability.X,example2$usability.Y,paired=TRUE)
```

- Inspect the output:

```
t = -3.1623, df = 4, p-value = 0.03411
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -3.7559781 -0.2440219
sample estimates:
mean of the differences
                     -2
```

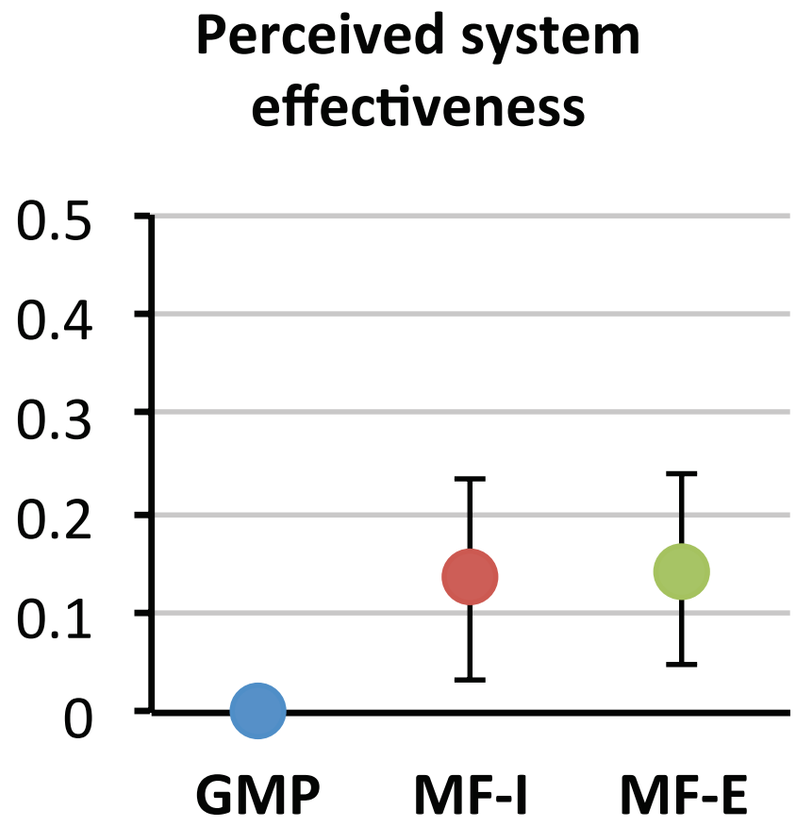
ANOVA

Differences between >2 systems / groups:

Are there differences in perceived system effectiveness between these 3 algorithms?

First do an omnibus test, then post-hoc tests or planned contrasts

Family-wise error!



Family-wise error

One statistical test: is the observed effect is “real” or due to chance variation?

We cannot be 100% certain, so we take p(chance) < .05

1 out of every 20 significant results could be a mistake!

Test all possible pairs of 5 conditions: 10 tests!

Family-wise error rate (chance of at least one mistake) is 40%!

Preventing this:

Always perform an omnibus test first

Not significant? Stop here!

Then, 3 options:

Pick a baseline condition and compare all conditions against that condition

Conduct “planned contrast” tests

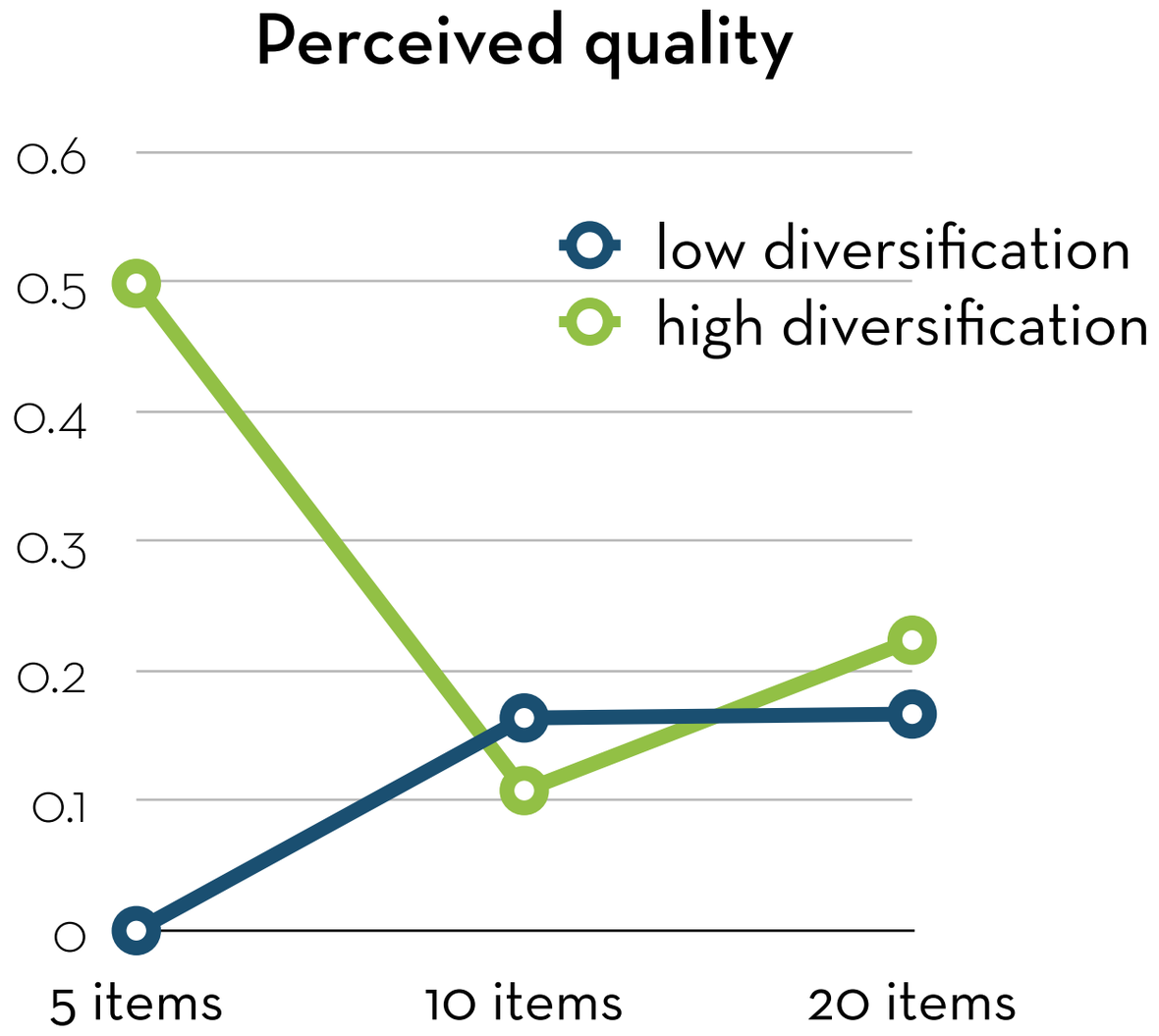
Perform all tests but use post-hoc test methods (e.g. Bonferroni correction)

Factorial ANOVA

Two manipulations at the same time:

What is the combined effect of list diversity and list length on perceived recommendation quality?

Test for the interaction effect!



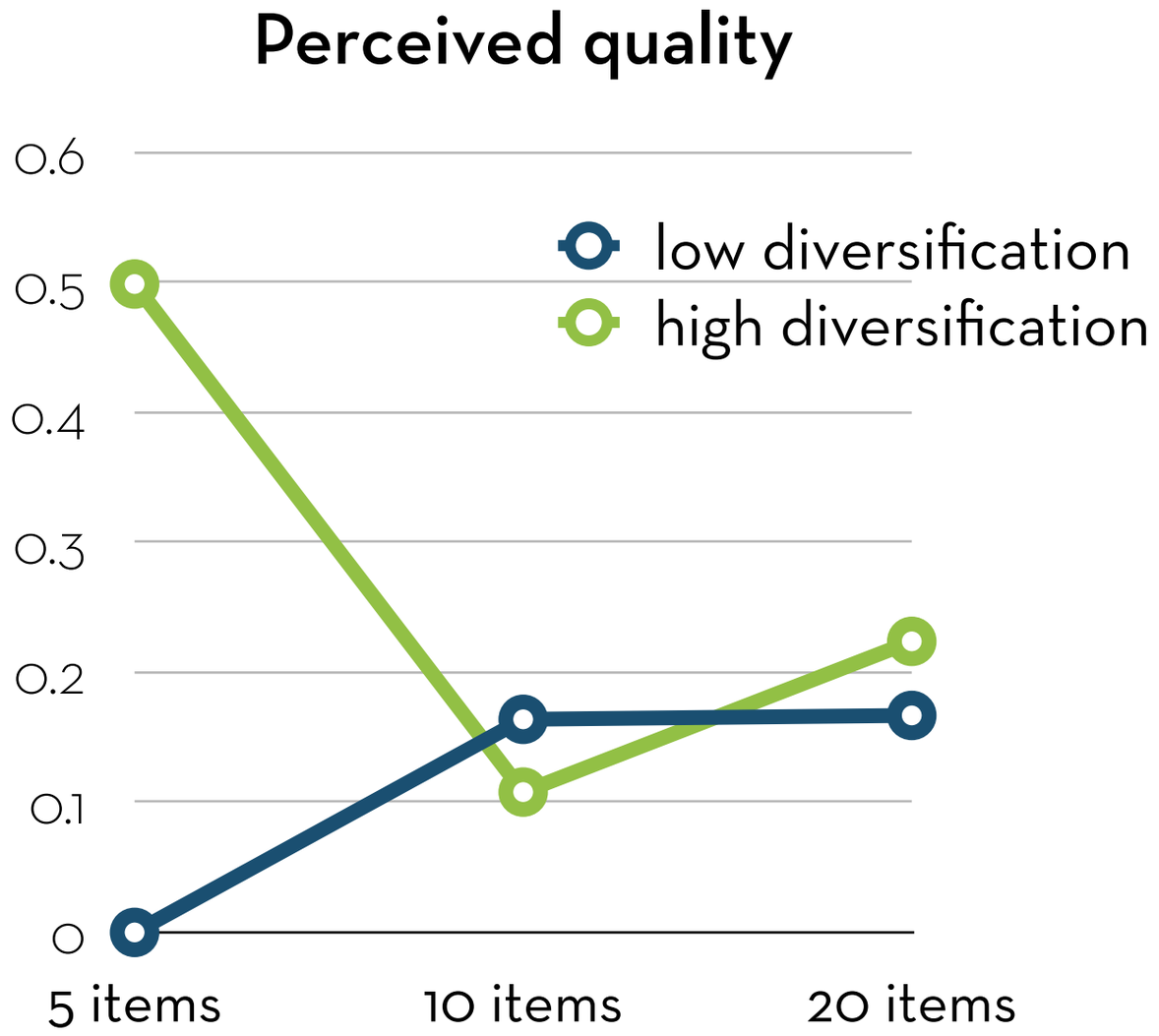
Willemsen et al.: "Understanding the Role of Latent Feature Diversification on Choice Difficulty and Satisfaction", submitted to UMUAI

Factorial ANOVA

Interaction effect:

"5-item lists have a higher perceived quality than 10- or 20-item lists, but only when diversification is high"

"High diversification lists have a higher perceived quality, but only for 5-item lists"



Willemsen et al.: "Understanding the Role of Latent Feature Diversification on Choice Difficulty and Satisfaction", submitted to UMUAI

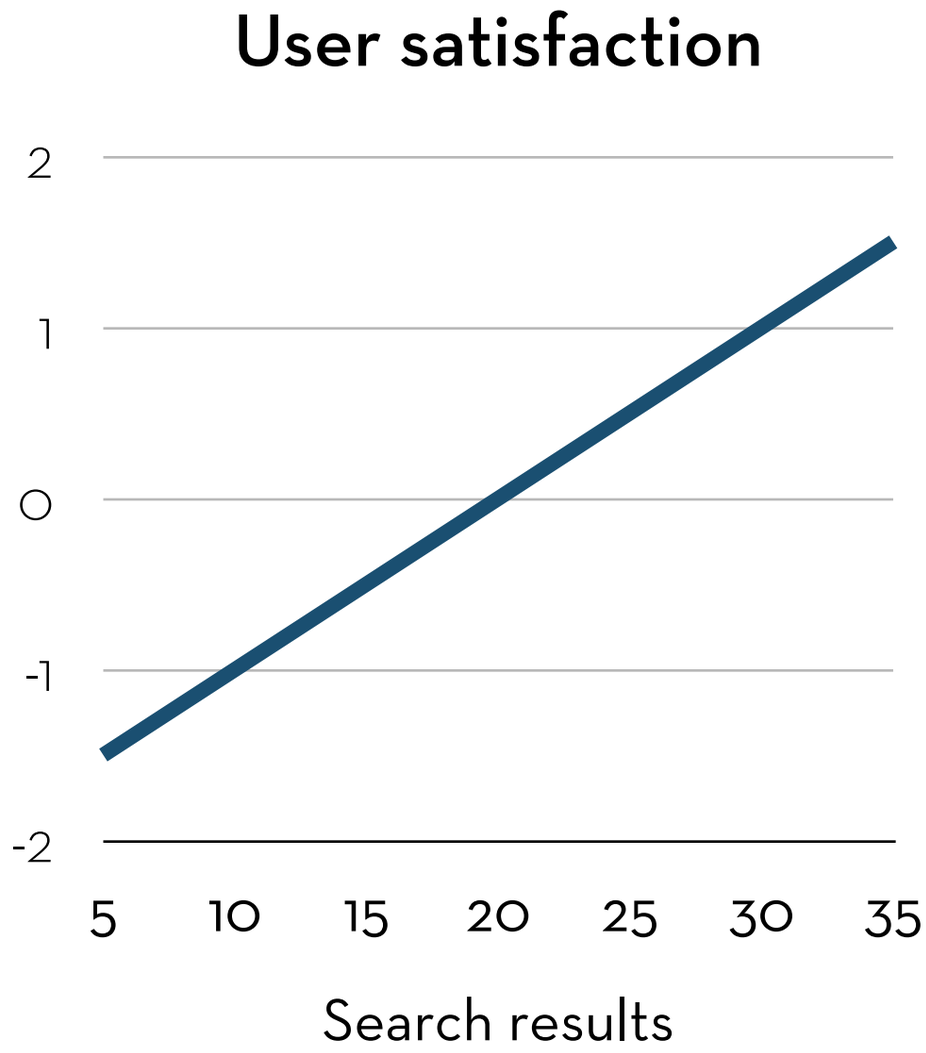
Regression

More of X -> more of Y:

Does user satisfaction increase with the number of search results?

More of X -> less of Y:

Does Facebook usage satisfaction decrease with age?



Regression concept

Compare slope (b) against variability of errors (S.E):

t = b/S.E.

Regression example

In RStudio:

- Run the t-test:

```
reg <- lm(attitude~usability, data=example)
```

- Run a summary of the results:

```
summary(reg)
```

- Inspect the output:

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.08537    0.79096  -0.108  0.91671
usability    0.82927    0.20700   4.006  0.00392 **
```

Regression example

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.08537    0.79096  -0.108  0.91671
usability    0.82927    0.20700   4.006  0.00392 **

Residual standard error: 0.8383 on 8 degrees of freedom
Multiple R-squared:  0.6673, Adjusted R-squared:  0.6258
F-statistic: 16.05 on 1 and 8 DF,  p-value: 0.003916
```

Attitude when usability is zero: –0.085 (intercept)

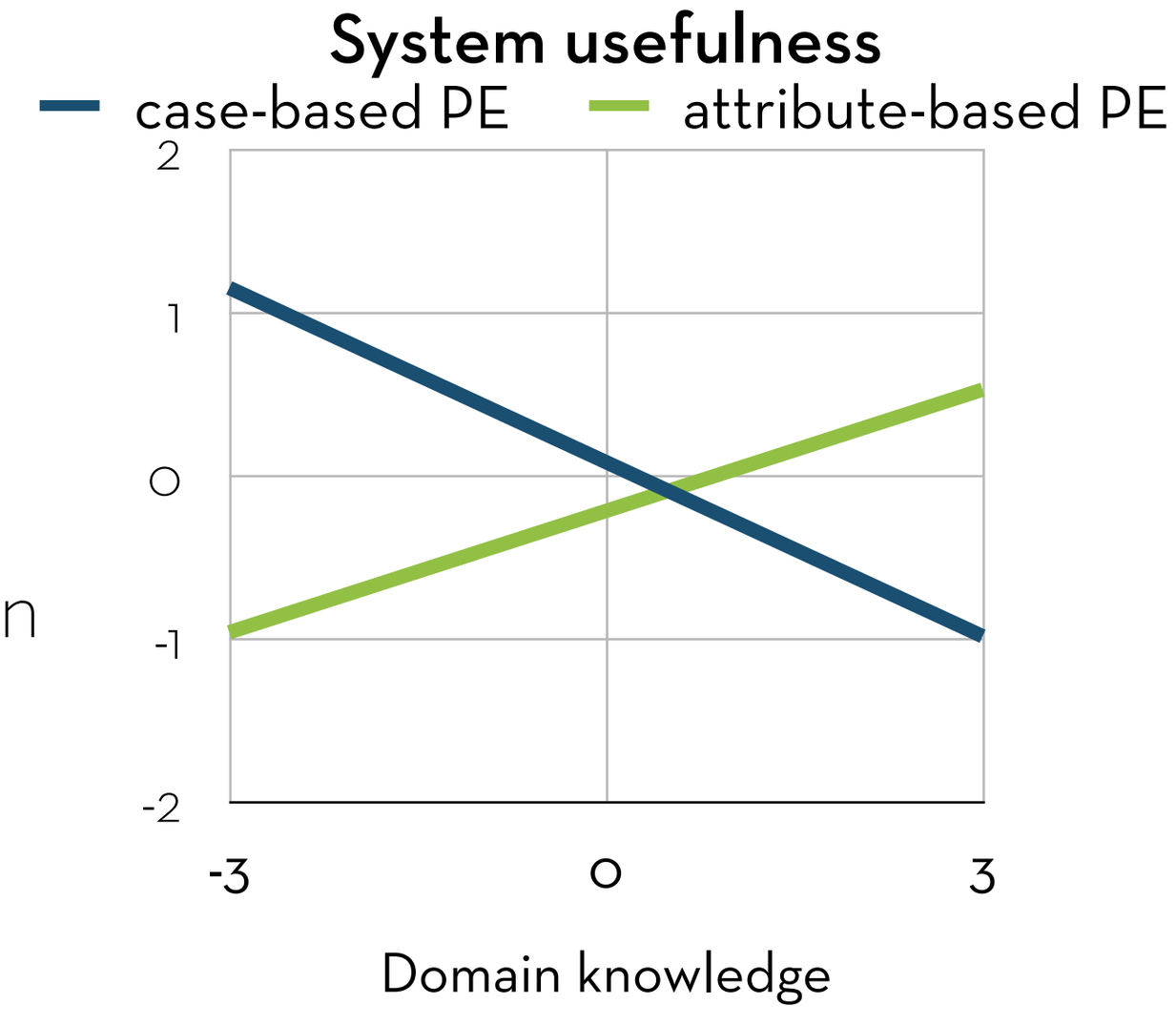
Increase in attitude for 1pt increase in usability: 0.829

Effect is highly significant (p = 0.004); explains 67% of variance (R-squared = 0.6673)

More complex...

Manipulation x personal characteristic -> outcome

Do experts and novices rate these two interaction methods differently in terms of usefulness?



Knijnenburg & WIllemsen: "Understanding the Effect of Adaptive Preference Elicitation Methods", RecSys2009

It is all the same!

Regression: Y = a + bX + e

Parameters: set the intercept (a) and slope (b), in a way that minimizes error (e)

Statistical test: P(b = 0) < 0.05

Is this slope significant (i.e. is the chance that it is actually zero smaller than 5%)?

If so: X has an effect on Y

If not: X has no effect on Y

It is all the same!

T-test: let's say you test system A versus B

Create a new variable (a "dummy"):

X = 0 for system A, and 1 for system B

Formula: Y = a + bX + e

For system A: Y = a + b*0 = a

For system B: Y = a + b*1 = a + b

Parameter b tests the **difference** between system A and B!

It is all the same!

One sample t-test: let's say you test system A versus B

Y = difference between system A and B for each user

Formula: Y = a + e

Parameter a tests the **difference** between system A and B!

It is all the same!

ANOVA: Let's say you have three systems: A, B, and C

Create two dummies:

X_B = 1 for users of system B, otherwise it is 0

X_C = 1 for users of system C, otherwise it is 0

Formula: $Y = a + b_B X_B + b_C X_C + e$

For system A: $Y = a + b_B{*}0 + b_C{*}0 = a$

For system B: $Y = a + b_B{*}1 + b_C{*}0 = a + b_B$

For system C: $Y = a + b_B{*}0 + b_C{*}1 = a + b_C$

It is all the same!

Formula: $Y = a + b_B X_B + b_C X_C + e$

Differences between systems:

A vs B: test $P(b_B = 0)$

A vs C: test $P(b_C = 0)$

B vs C: test $P(b_B - b_C = 0)$

Omnibus test: $P(b_B = 0 \text{ and } b_C = 0)$

It is all the same!

Factorial ANOVA: Let's say you have 2 binary variables X_1 and X_2

Create two dummies: X_1 and X_2

Formula: $Y = a + b_1X_1 + b_2X_2 + b_3X_1X_2 + e$

b_1 and b_2 are main effects, b_3 is the interaction effect

It is all the same!

Conclusion: every standard test (t-test, ANOVA, Factorial ANOVA, ANCOVA) can be expressed as a regression!

Learn more?

Take a class (Clemson):

STAT 8010 Statistical Methods I

STAT 8050 Design and Analysis of Experiments

PSYC 8100 Research Design and Quantitative Methods I

HCC 8810 Measurement and Evaluation of HCC systems

Take a class (UC Irvine):

STATS 201

SocEcol 264A and B

Learn more?

Learn it yourself:

Jessica Utts, “Seeing Through Statistics”

Andy Field, “Discovering Statistics” series

BRENICS
INFORMATION AND COMPUTER SCIENCES

Pitfalls

Why these methods often don't work

Overview

Y is not normal

Why? Measuring time, counts, yes/no, etc.

Correlated errors

Why? Y is repeated / X is grouped

Y is unobserved

Why? You want to measure subjective evaluations

You want to test X -> M -> Y

Why? To test a theory

Y is not normal

Standard tests assume that the dependent variable (Y) is an continuous, unbounded, normally distributed interval variable

Continuous: variable can take on any value, e.g. 4.5 or 3.23 (not just whole numbers)

Unbounded: range of values is unlimited (or at least does not stop abruptly)

Interval: differences between values are comparable; is the difference between 1 and 2 the same as the difference between 3 and 4?

Y is not normal

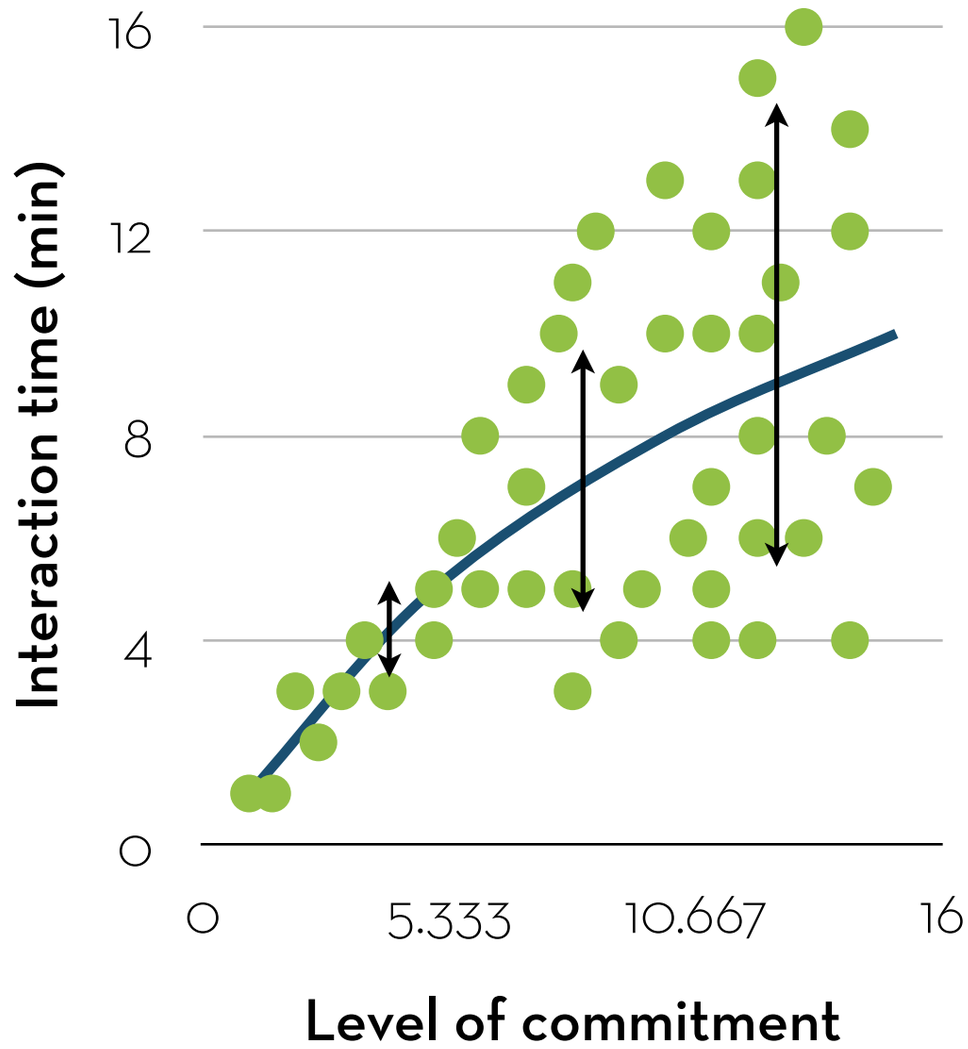
Not true for most behaviors!

Number of clicks (discrete, bounded by zero, not normal)

Time, money (bounded by zero, not normal)

1-5 ratings (bounded, discrete, not interval)

Decisions (yes/no)



Bad solution...

Use “distribution-free” or “non-parametric” tests

- Mann–Whitney U test (t-test)
- Wilcoxon signed-rank test (within-subjects t-test)
- Kruskal-Wallis test (ANOVA)
- Friedman’s test (within-subjects ANOVA)

These are old-fashioned solutions

- They do not work for non-continuous data types
- Other methods are typically much more powerful

Good solution

Transform the dependent variable to make it more normal

E.g. log transformation for zero-bounded variables:
$x_t = \ln(x + a)$

Use the “generalized linear models” (GLMs)

- Binary data: logit/probit regression
- 5- or 7-point scales: ordered logit/probit regression
- Count data: Poisson regression

If no correct method exists, use a robust estimator

Learn more?

Take a class (Clemson):

STAT 8020 Statistical Methods II

HCC 8810 Measurement and Evaluation of HCC systems

Take a class (UC Irvine):

STATS 202

Learn it yourself:

Alan Agresti, “Categorical Data Analysis”, 2nd ed.

Correlated errors

Standard regression requires **uncorrelated errors**

This is not the case when...

...you have repeated measurements of the same participant (e.g. you measured 5 task performance times per participant, for 60 participants)

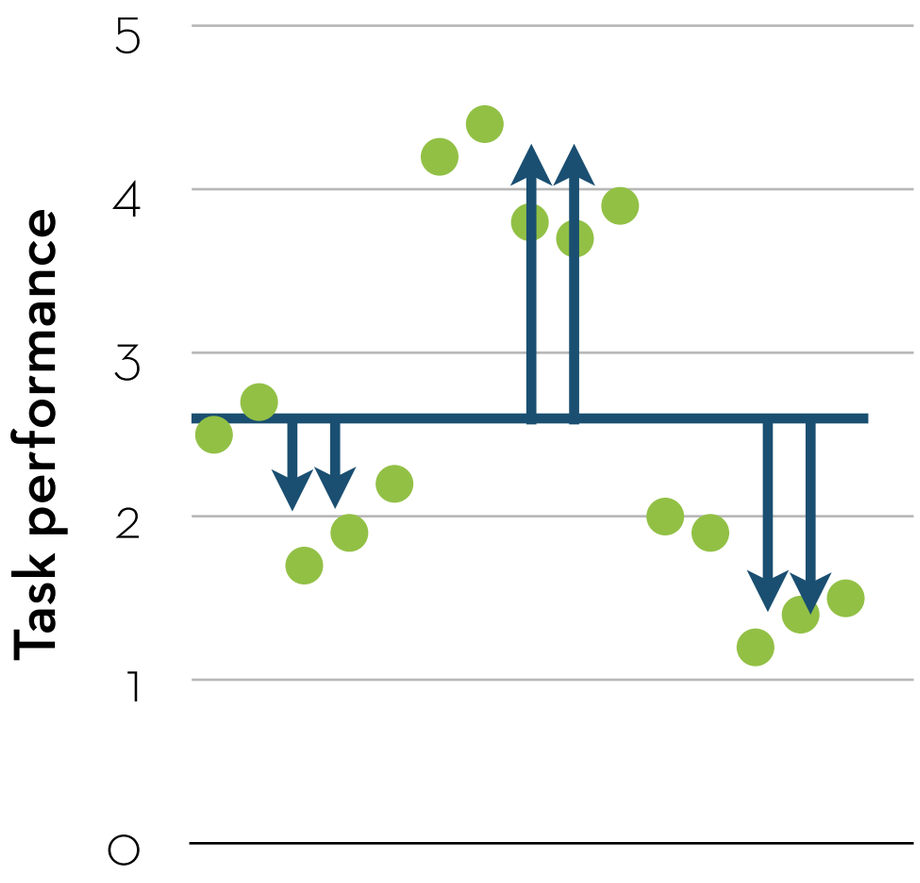
...participants are somehow related (e.g. you measured the performance of 5 group members, for 60 groups)

Correlated errors

Consequence: errors are correlated

There will be a user-bias (and maybe an task-bias)

Golden rule: data-points should be **independent**

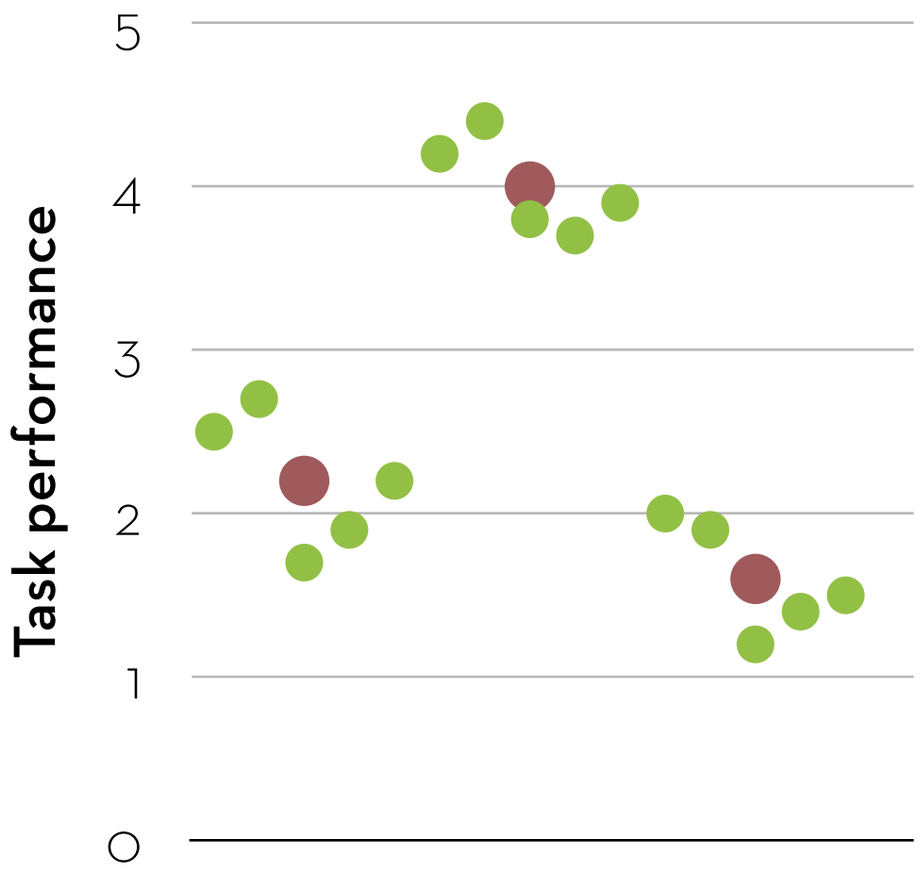


OK solution...

Take the average of the repeated measurements

Reduces the number of observations

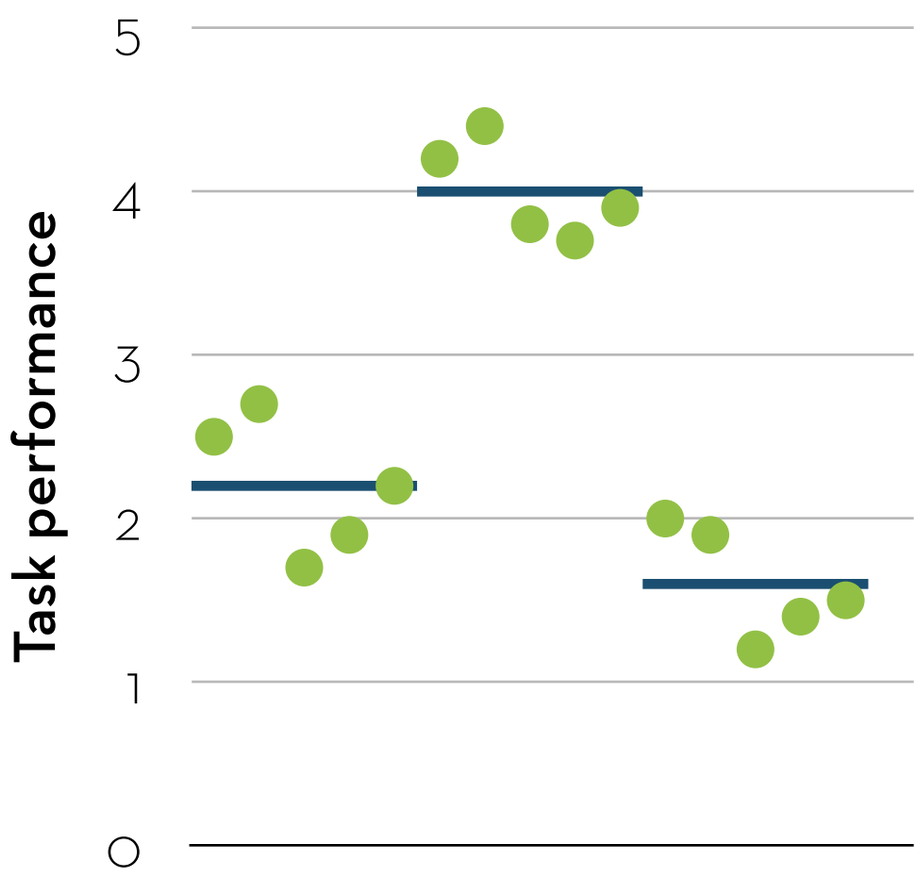
It becomes impossible to make inferences about individual tasks/users/etc.



Good solution

Use a multi-level regression method that allows one to estimate the error correlations:

- …by defining a random intercept for each user (GLMM)
- …by imposing an error covariance structure (GEE)



Learn more?

Take a class (Clemson):

STAT 8020 Statistical Methods II

HCC 8810 Measurement and Evaluation of HCC systems

Take a class (UC Irvine):

STATS 203

Learn it yourself:

Fitzmaurice, Laird and Ware, "Applied Longitudinal Analysis"

Y is unobserved

Behavior is an “observed” variable

- Relatively easy to quantify
- E.g. time, money spent, click count, yes/no decision

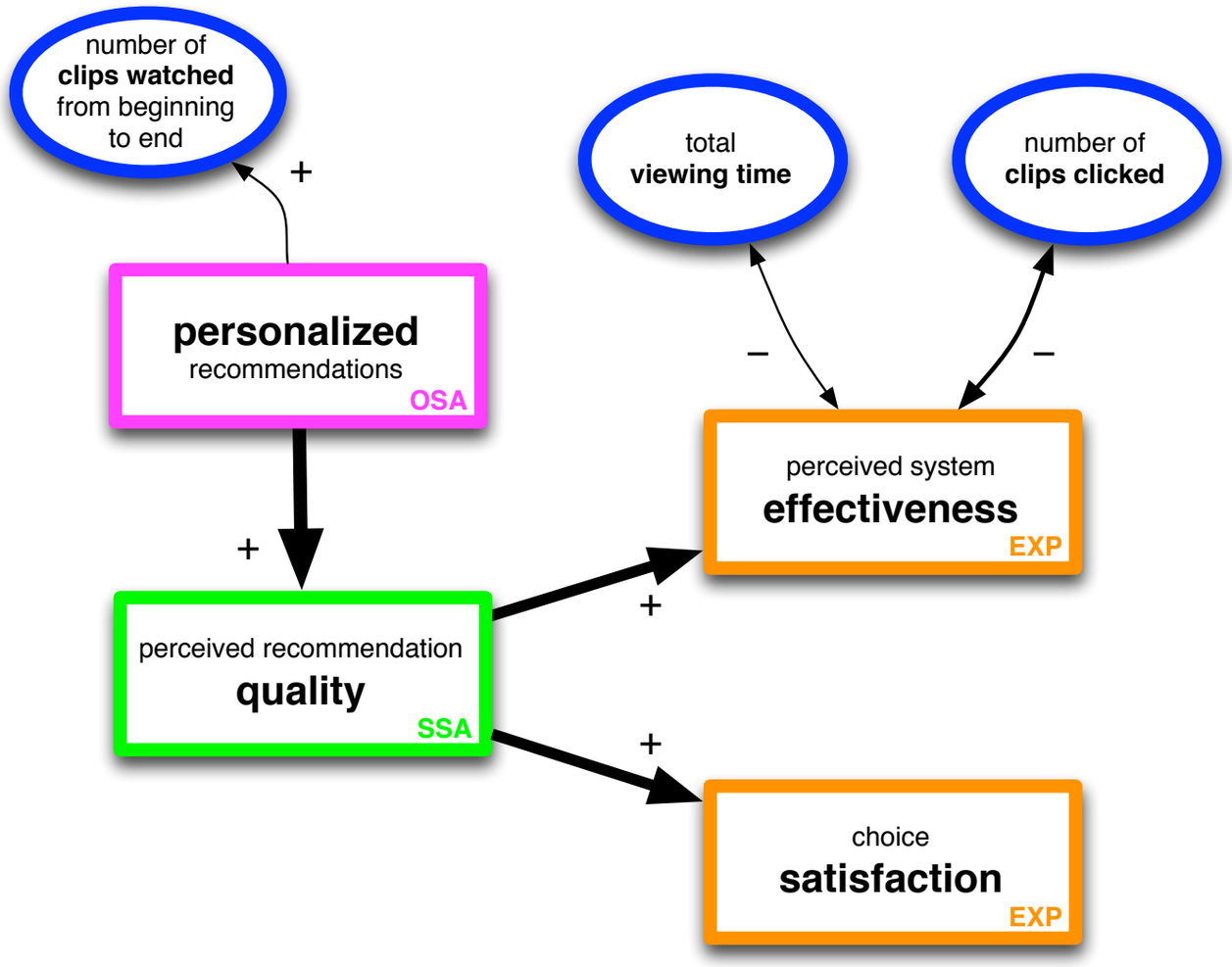
Perceptions, attitudes, and intentions (subjective valuations) are “unobserved” variables

- They happen in the user’s mind
- How can we quantify them?
- But first: why should we measure them at all?

Why go subjective?

"Testing a recommender against a random videoclip system, the number of clicked clips and total viewing time went down!"

Why go subjective?



Knijnenburg et al.: “Receiving Recommendations and Providing Feedback”, EC-Web 2010

Why go subjective?

Behavior is hard to interpret

- Relationship between behavior and satisfaction is not always trivial

User experience is a better predictor of long-term retention

- With behavior only, you will need to run for a long time

Questionnaire data is more robust

- Fewer participants needed

Why go subjective?

Measure **subjective valuations** with questionnaires

- Perception, experience, intention

Triangulate these data with behavior

- Ground subjective valuations in observable actions
- Explain observable actions with subjective valuations

Y is unobserved

Behavior is an “observed” variable

- Relatively easy to quantify
- E.g. time, money spent, click count, yes/no decision

Perceptions, attitudes, and intentions (subjective valuations) are “unobserved” variables

- They happen in the user’s mind
- **How can we quantify them?**
- But first: why should we measure them at all?

Bad solution...

“To measure satisfaction, we asked users whether they liked the system (on a 5-point rating scale).”

Why is this bad?

Does the question mean the **same** to everyone?

- John likes the system because it is convenient
- Mary likes the system because it is easy to use
- Dave likes it because the outcomes are useful

A single question is not enough to establish **content validity**

We need a multi-item measurement scale

Example scale

Perceived system effectiveness:

- "Using the system is annoying"
- "The system is useful"
- "Using the system makes me happy"
- "Overall, I am satisfied with the system"
- "I would recommend the system to others"
- "I would quickly abandon using this system"

5- or 7-point scale: from "completely disagree" to "completely agree"

OK solution...

"We asked users ten 5-point scale questions and summed the answers."

What is missing?

Is the scale really measuring a **single** thing?

- 5 items measure satisfaction, the other 5 convenience
- The items are not related enough to make a reliable scale

Are two scales really measuring **different** things?

- They are so closely related that they actually measure the same thing

We need to establish **convergent** and **discriminant validity**

This makes sure the scales are unidimensional

Good solution

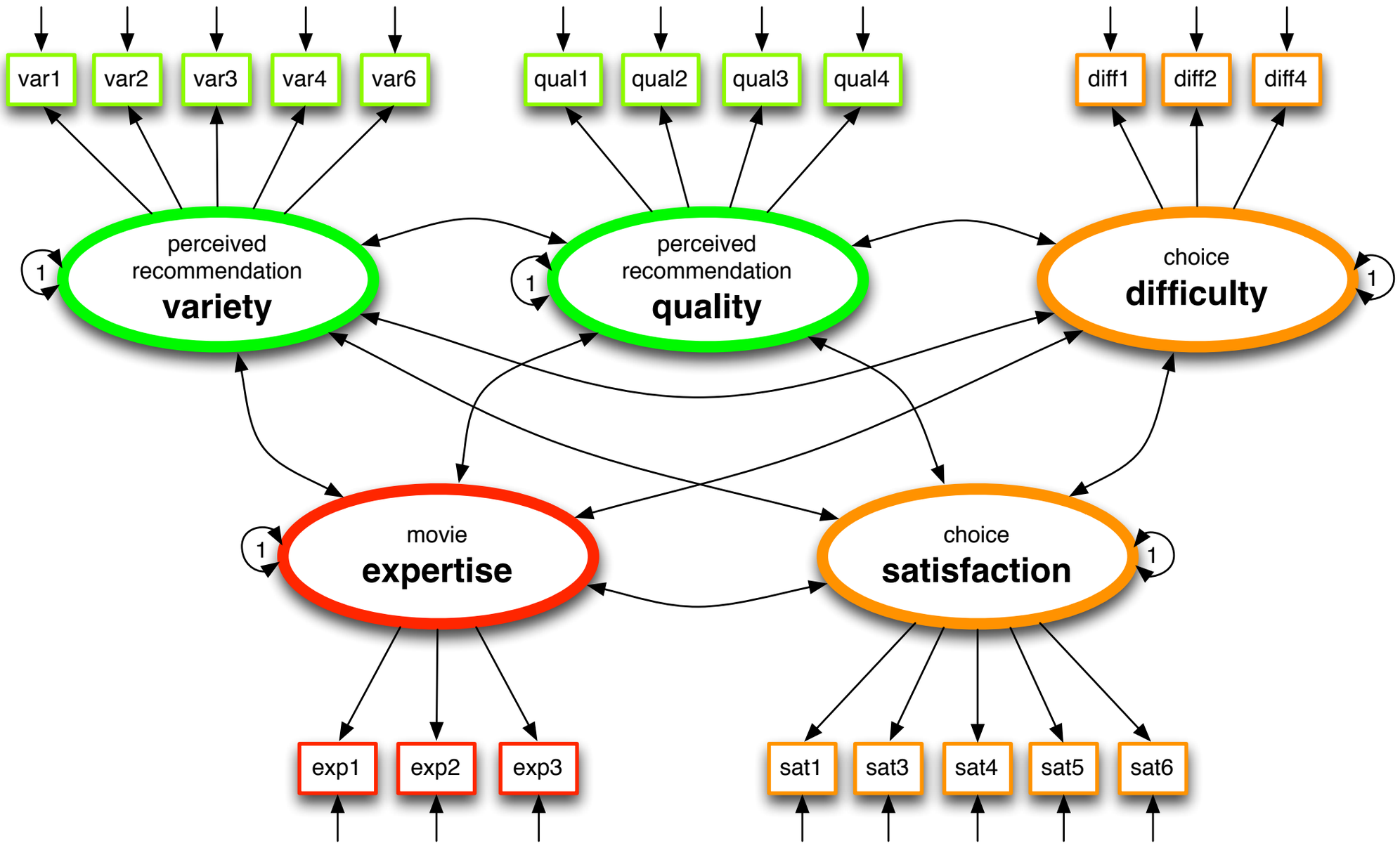
Use **factor analysis**

- Define latent factors, specify how items "load" on them
- Factor analysis will determine how well the items "fit"
- It will give you suggestions for improvement

Benefits of factor analysis:

- Establishes convergent and discriminant validity
- Outcome is a normally distributed measurement scale
- The scale captures the "shared essence" of the items

Factor Analysis



Learn more?

Take a class (Clemson):

This one! (measurement will be covered next time)

PSYC 8710 Psychological Tests and Measurement

MGT 9050 Research Methods

HCC 8810 Measurement and Evaluation of HCC systems

Take a class (UC Irvine):

Prof. Jone Pearce's Measurement Practicum (part of "Mgmt 291: Doctoral Seminar in Organizational Behavior")

Learn more?

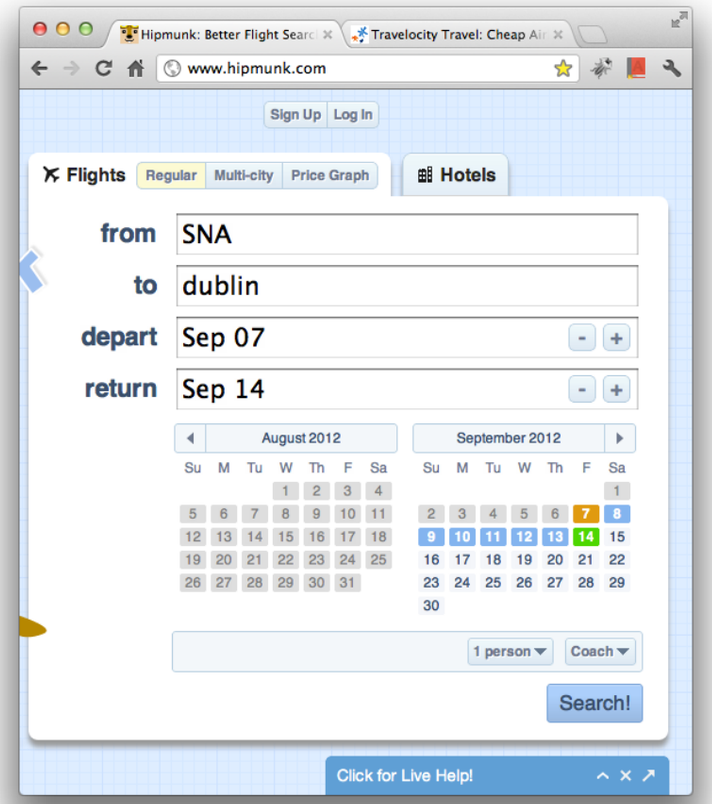
Learn it yourself:

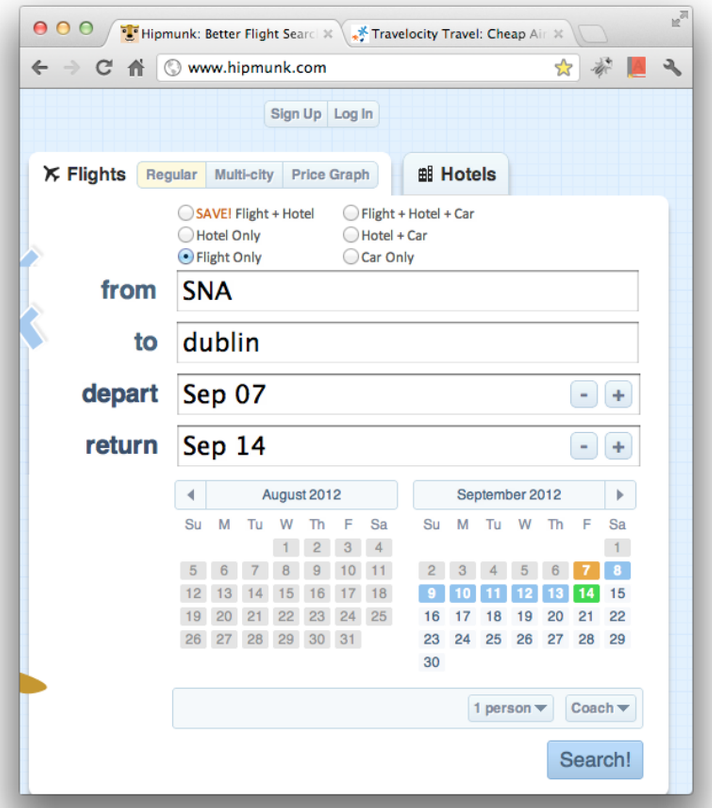
Robert DeVellis, “Scale Development”, 2nd ed.

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

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

Theory behind x->y





Why would the new system (X) have a higher usability (Y)?

Mediation: x->m->y

To learn something from a study, we need a **theory** behind the effect

- This makes the work generalizable
- This may suggest future work

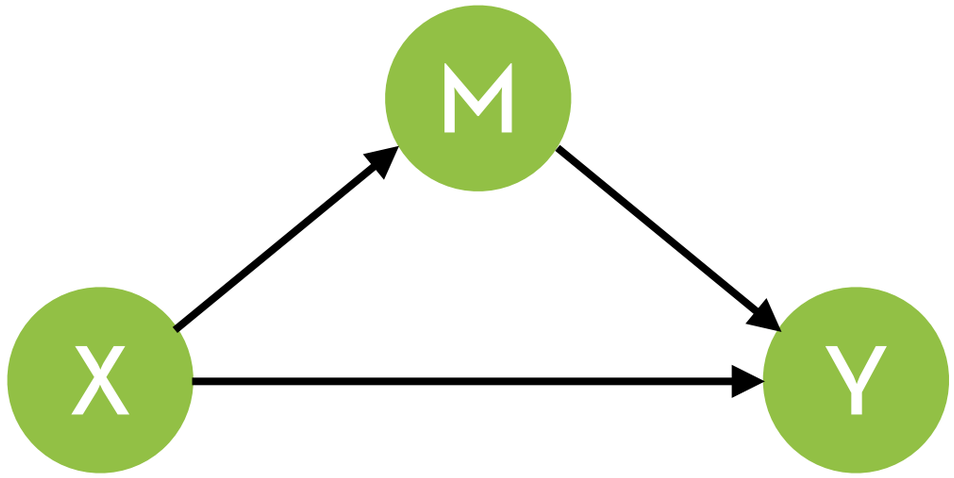
Measure **mediating variables**

- Measure understandability (and a number of other concepts) as well
- Find out how they mediate the effect on usability

Mediation Analysis

Manipulation -> perception
-> experience

Does the system
influence usability
via understandability?



Types of mediation

Partial mediation

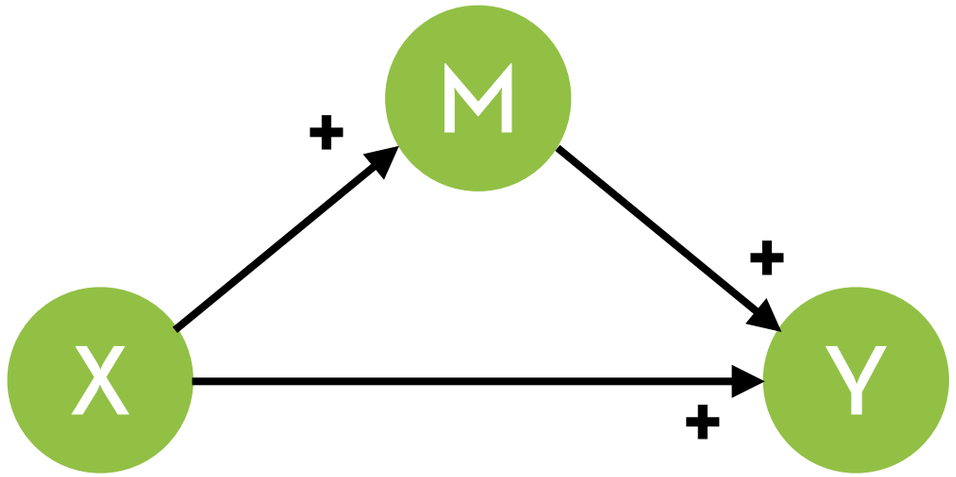
Full mediation

Negative mediation

Mediation Analysis

Manipulation -> perception
-> experience

Does the system
influence usability
via understandability?



Types of mediation

Partial mediation

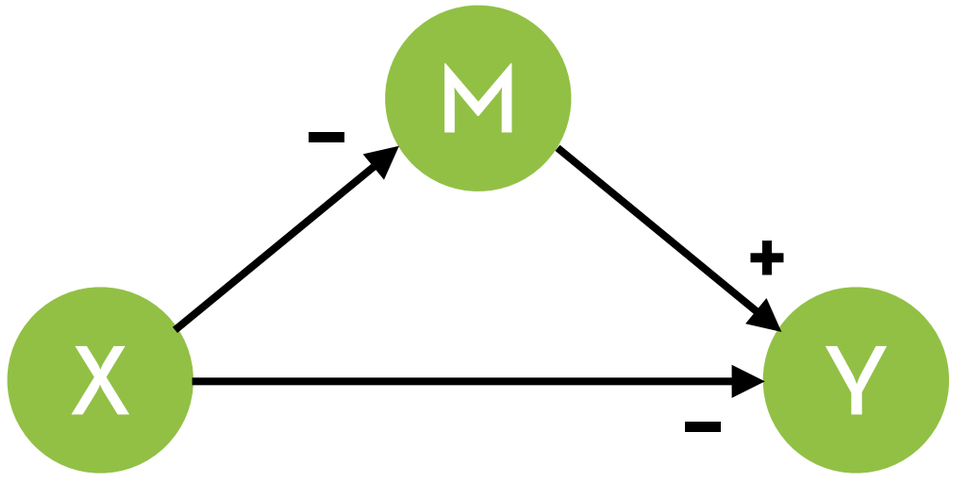
Full mediation

Negative mediation

Mediation Analysis

Manipulation -> perception
-> experience

Does the system
influence usability
via understandability?



Types of mediation

Partial mediation

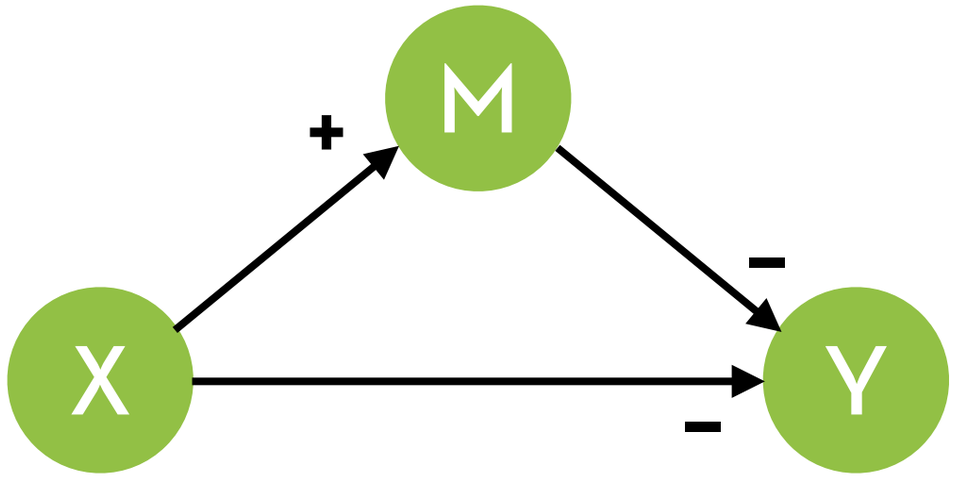
Full mediation

Negative mediation

Mediation Analysis

Manipulation -> perception
-> experience

Does the system
influence usability
via understandability?



Types of mediation

Partial mediation

Full mediation

Negative mediation

Mediation Analysis

Manipulation -> perception
-> experience

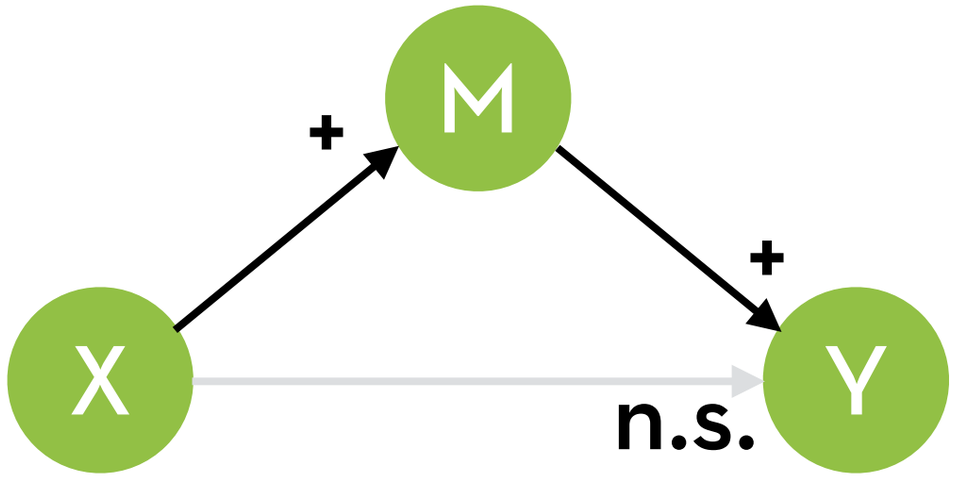
Does the system
influence usability
via understandability?

Types of mediation

Partial mediation

Full mediation

Negative mediation



Mediation Analysis

Manipulation -> perception
-> experience

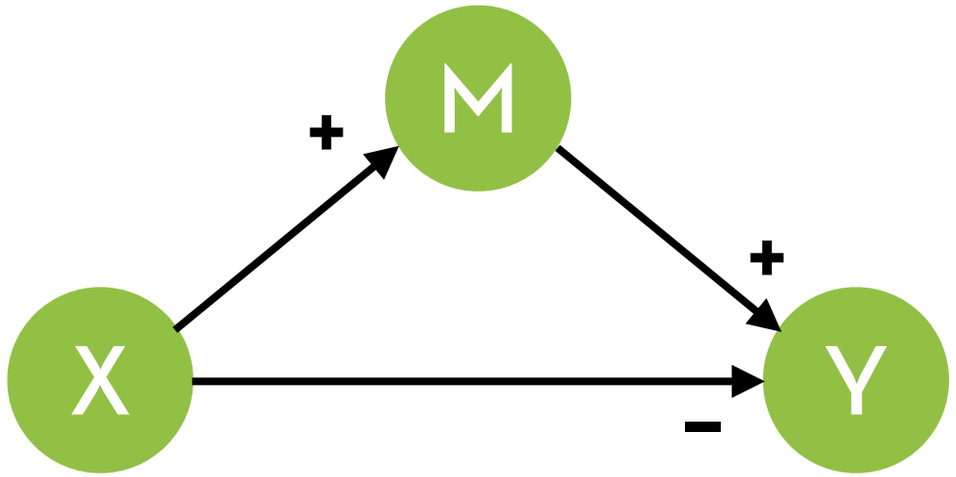
Does the system
influence usability
via understandability?

Types of mediation

Partial mediation

Full mediation

Negative mediation



Old way of testing

The four steps of Baron & Kenny, 1986 (see www.davidkenny.net)

1. X -> Y should be significant (note: this step has been contested!)
2. X -> M should be significant
3. M -> Y should be significant in a regression that controls for X
4. For complete mediation, X -> Y should be "zero" in a regression that controls for M (same regression as step 3)

Old way of testing

Finally, test the significance of the indirect effect (X->M->Y)

Methods:

- Sobel test (simple but conservative)
- Bootstrapping (a bit too liberal)
- Monte-Carlo simulation (complicated)

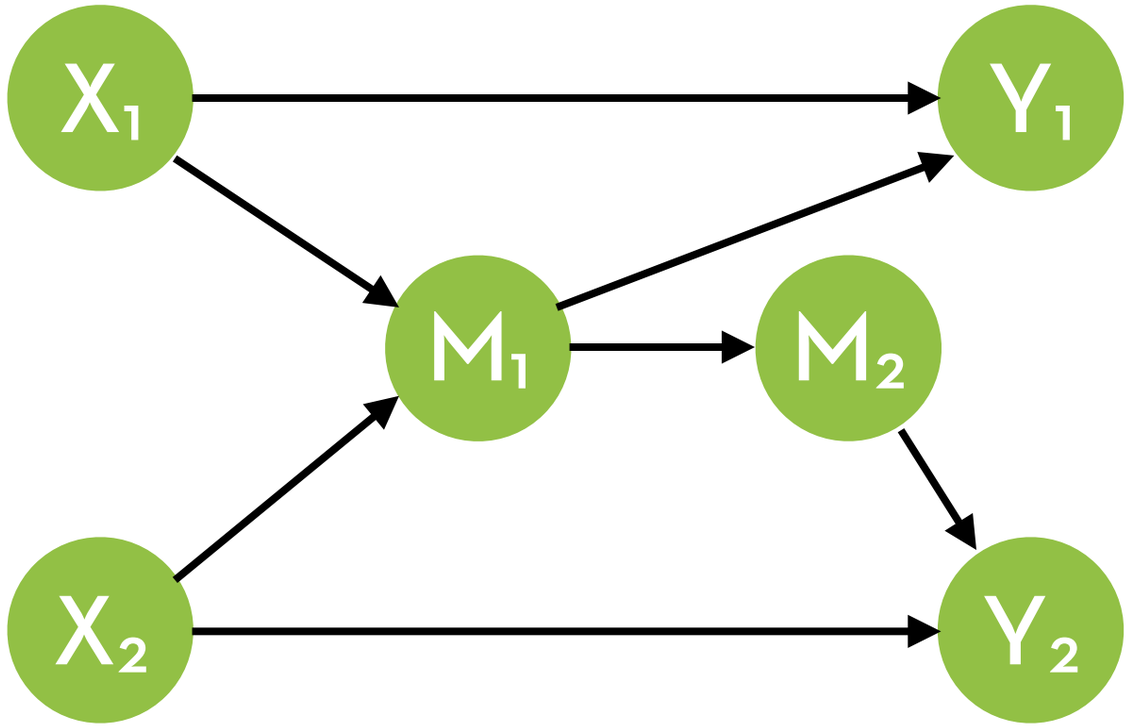
Problems...

Mediation Analysis is a lot of work

- Many tests to conduct
- Many findings to report

Gets even more complicated with more "interesting" models

No "overall" test of the model



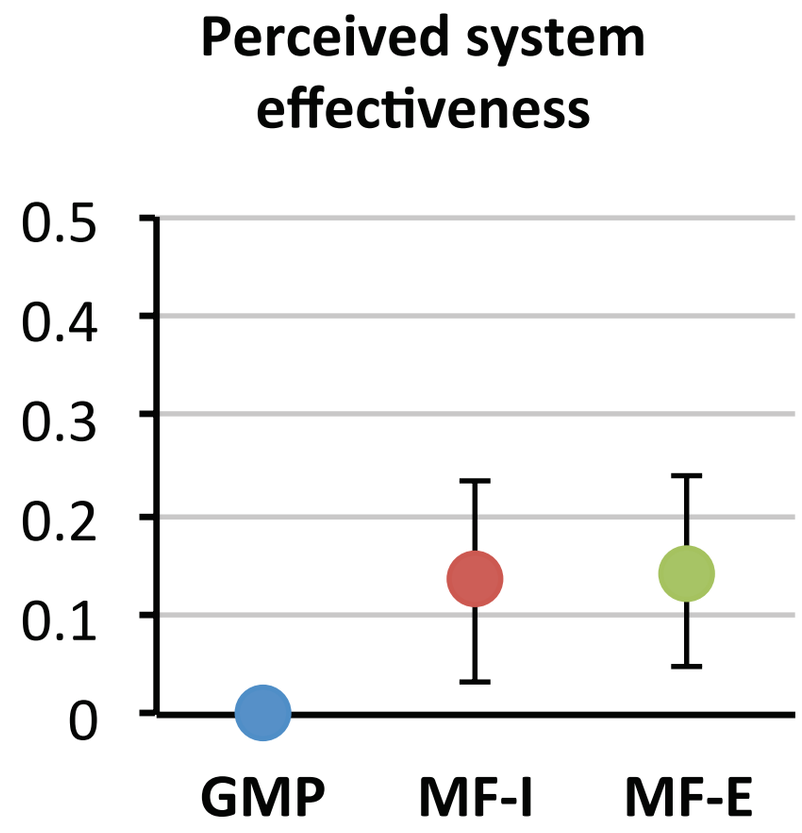
Example

We compared three recommender systems

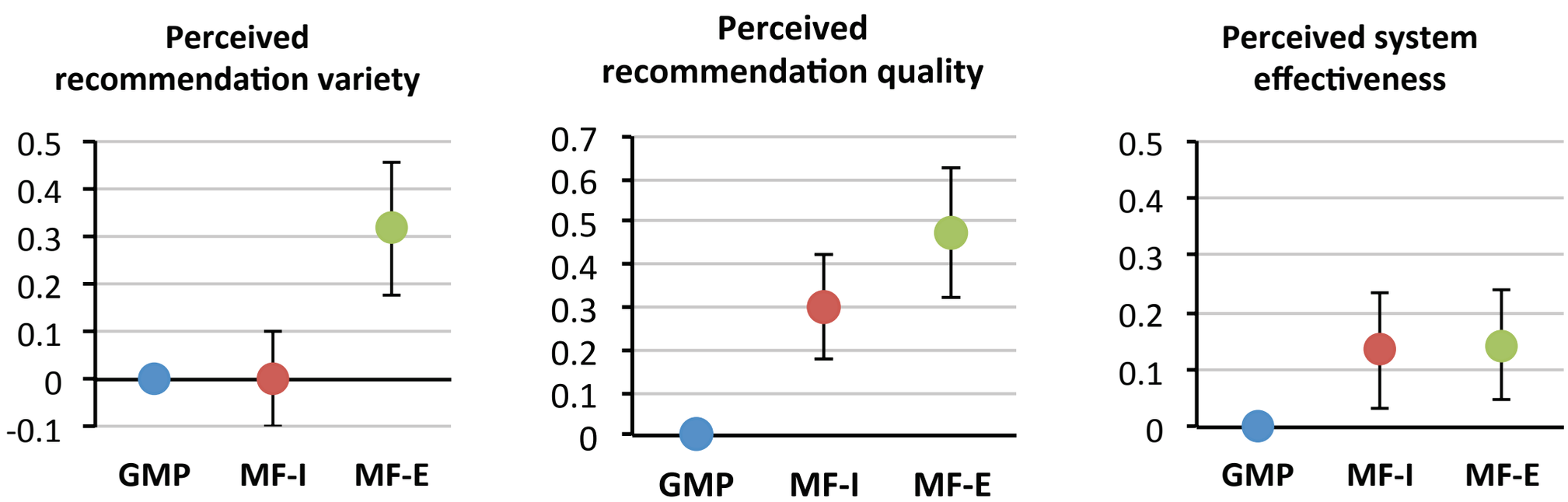
- Most popular items
- MF w/ implicit feedback
- MF w/ explicit feedback

MF-I and MF-E make the system more effective

- Why?



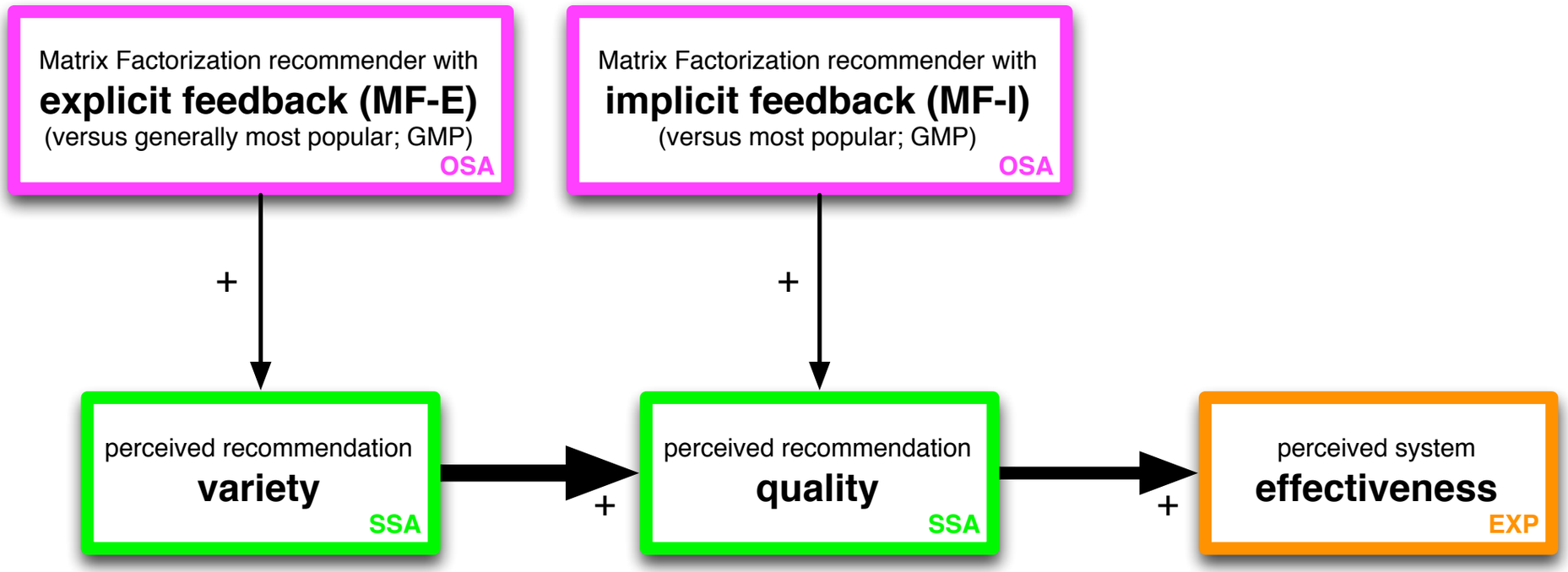
Example



Knijnenburg et al.: "Explaining the user experience of recommender systems", UMUAI 2012

The mediating variables show the entire story

Example



Knijnenburg et al.: "Explaining the user experience of recommender systems", UMUAI 2012

Advantages

Overall model fit statistics:

E.g. chi-square model fit, CFI, TLI, and RMSEA

Model coefficients:

E.g. the regression of perceived quality on effectiveness has b = 0.846, s.e. = 0.127, p < 0.001

Other useful tests:

E.g. modification indices, indirect and total effects, omnibus tests

Causality

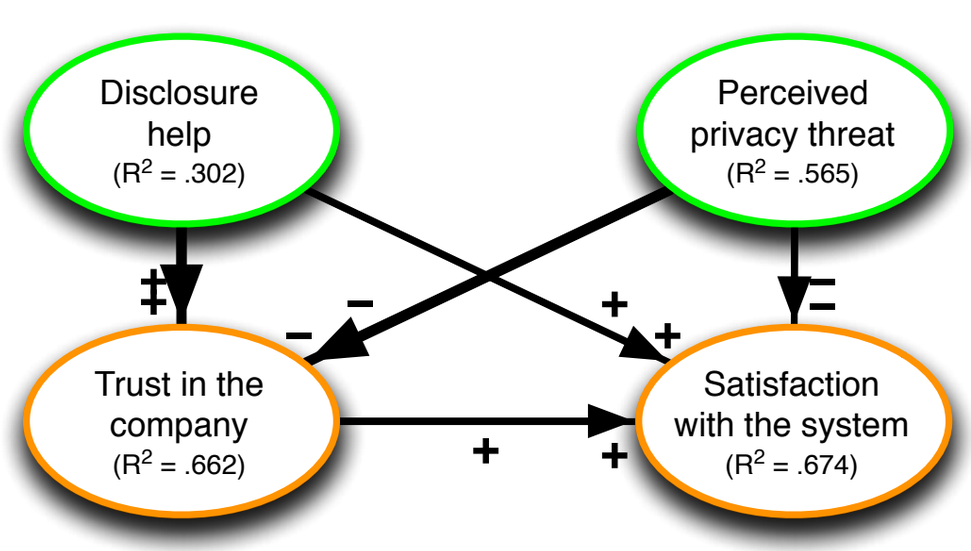
What causes what?

A **manipulation** only causes things

For all other variables:

- Common sense
- Existing work
- Existing theory/models

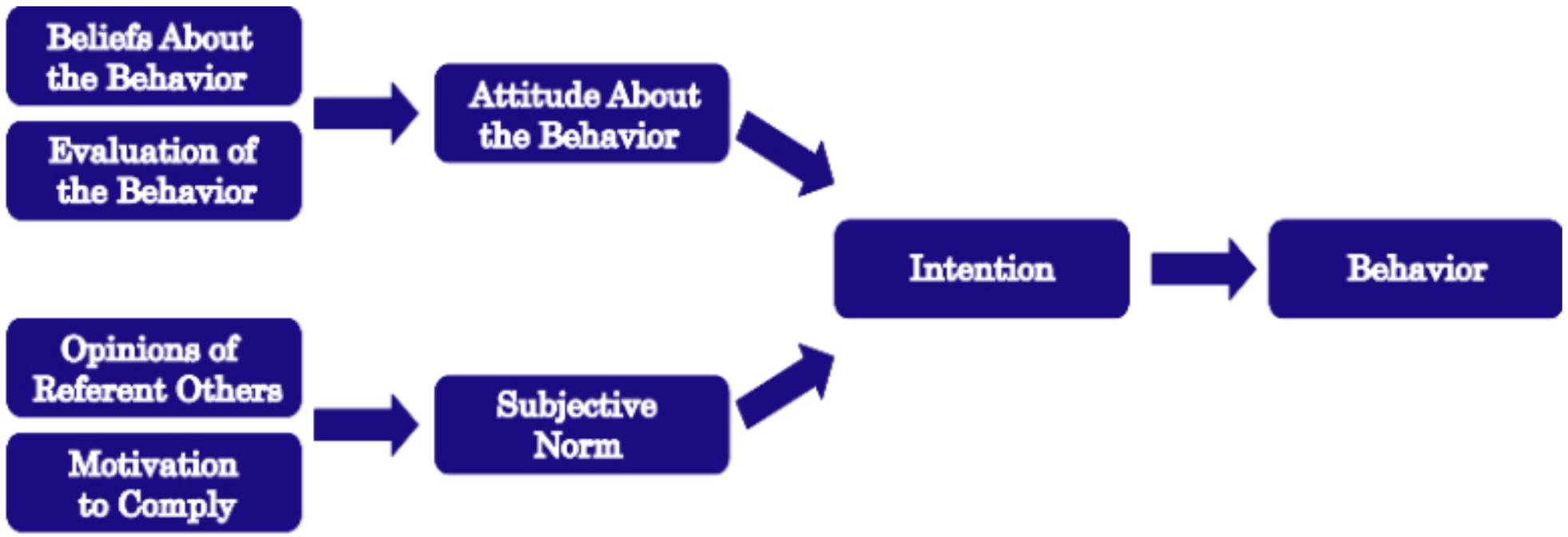
Example: privacy study



Knijnenburg & Kobsa.: "Making Decisions about Privacy"

Existing models

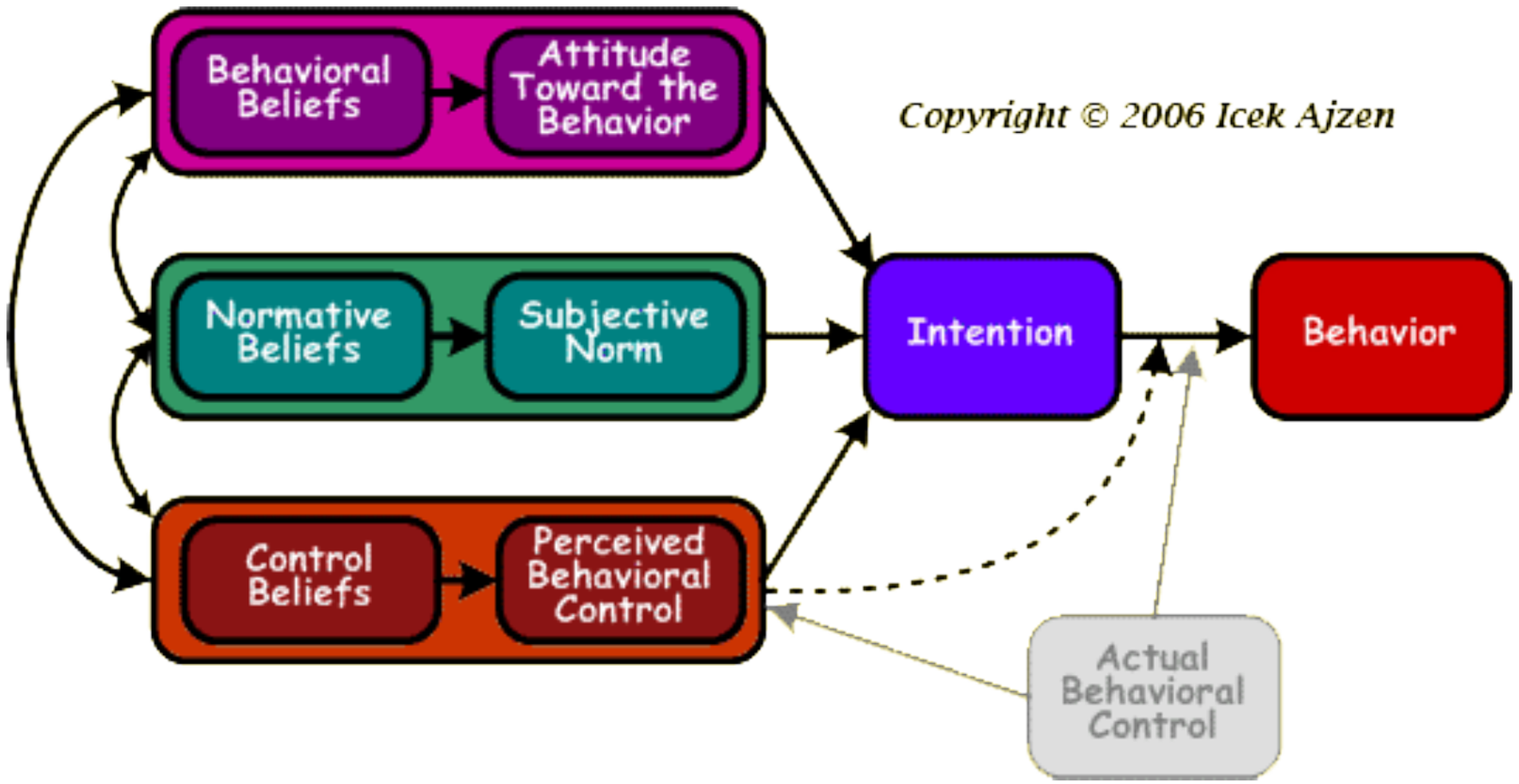
Theory of Reasoned Action (TRA)



Fishbein-Aizen Theory of Reasoned Action

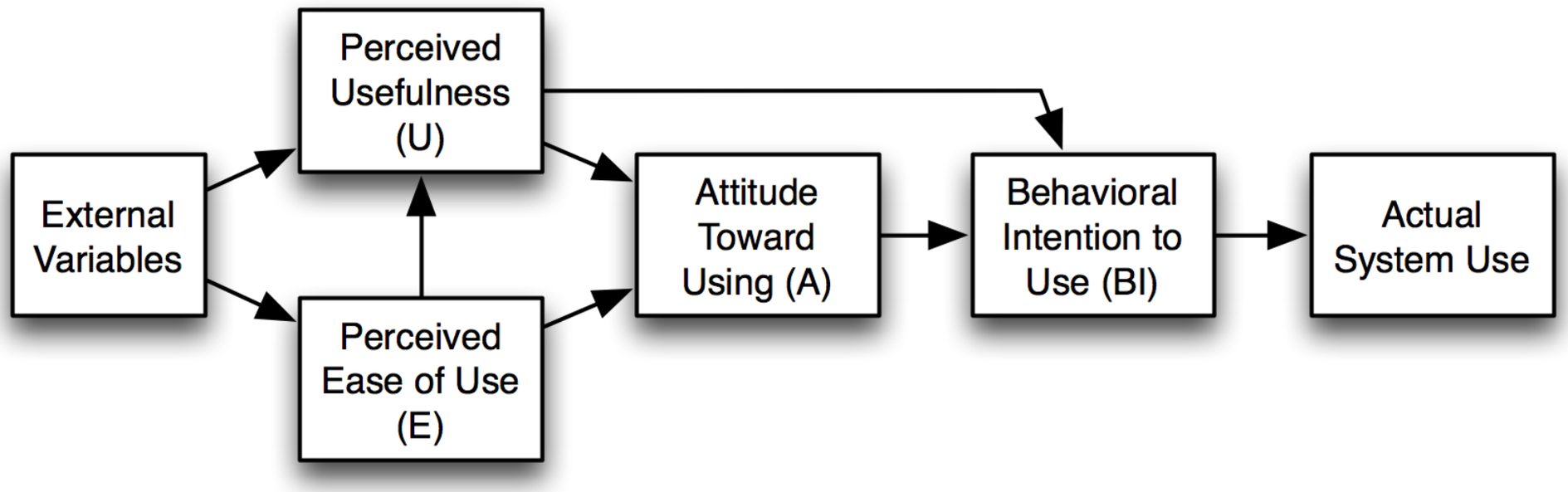
Existing models

Theory of Planned Behavior (TPB)



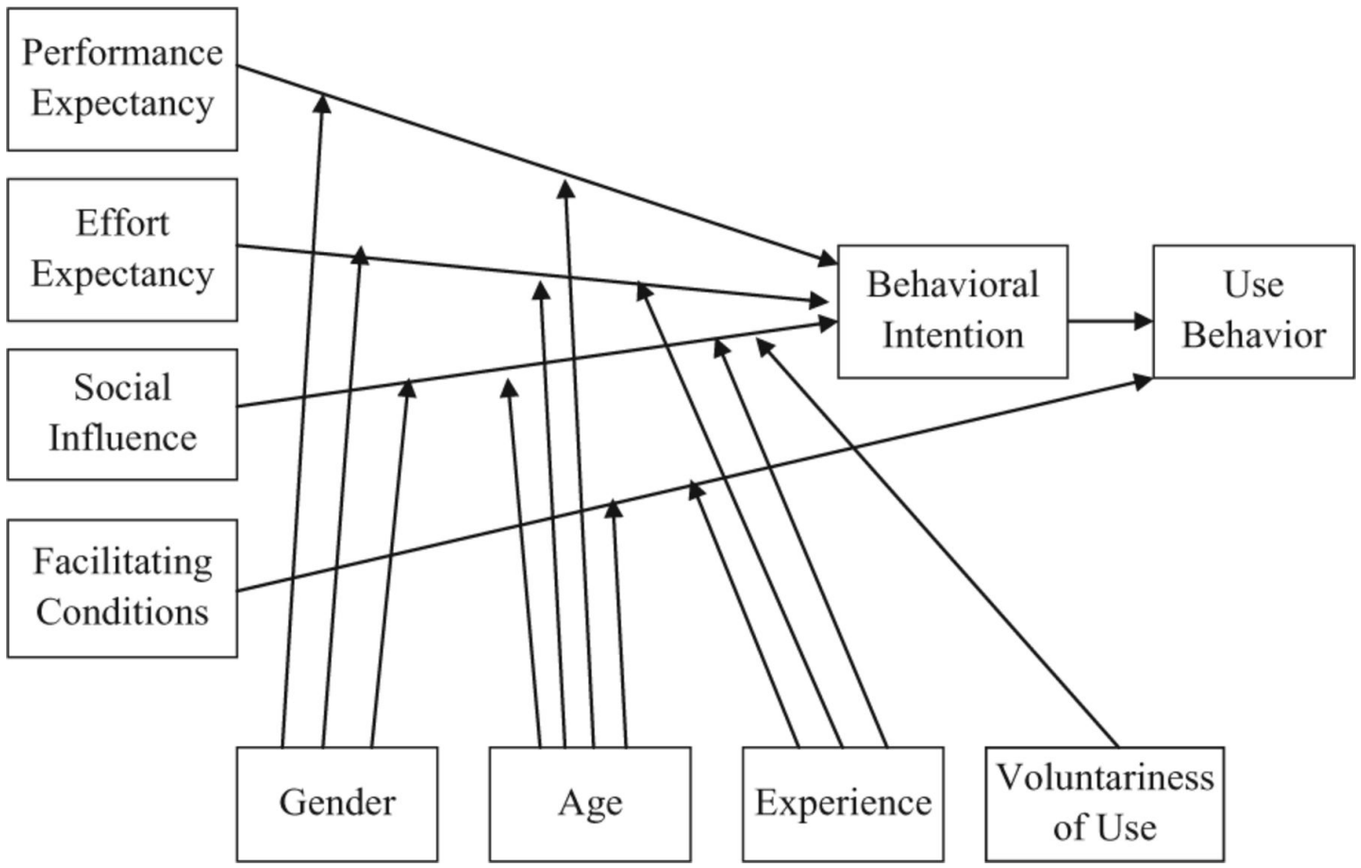
Existing models

Technology Acceptance Model (TAM)



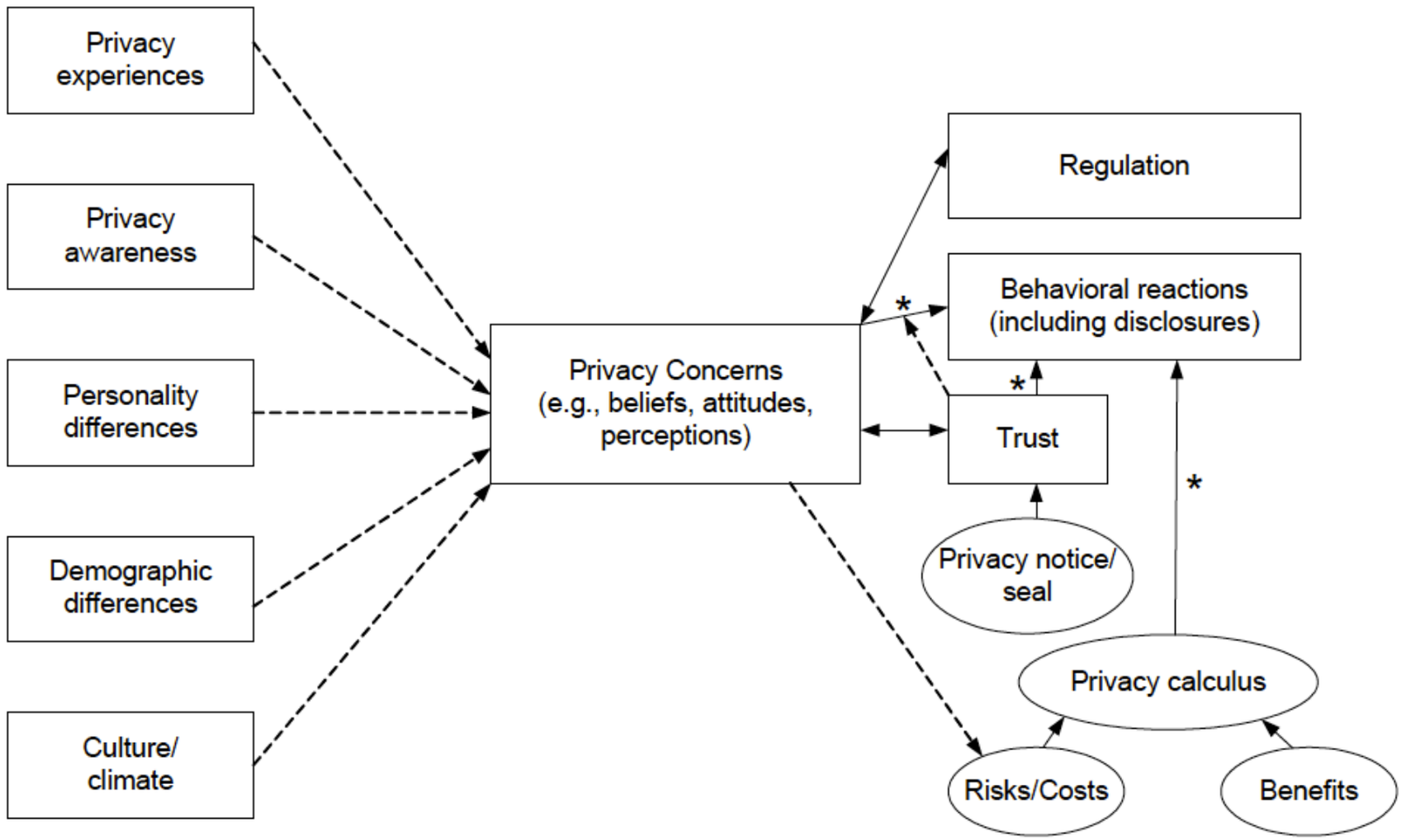
Existing models

Unified Theory of Acceptance and Use of Technology (UTAUT)



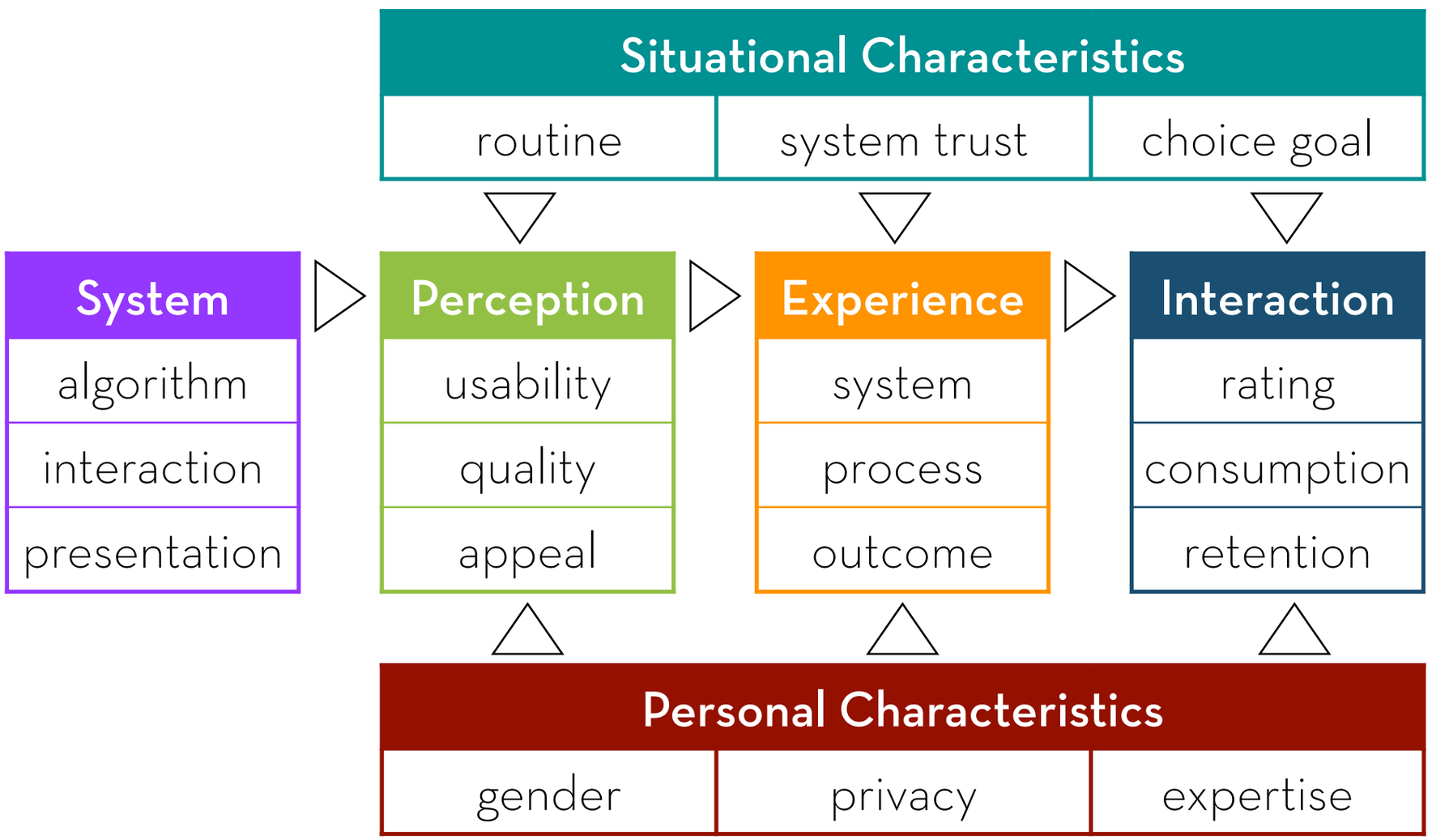
Existing models

Field-specific, e.g. privacy: Smith et al., MISQ



Existing models

Field-specific, e.g. recommender systems: Knijnenburg et al., UMUAI



Analysis

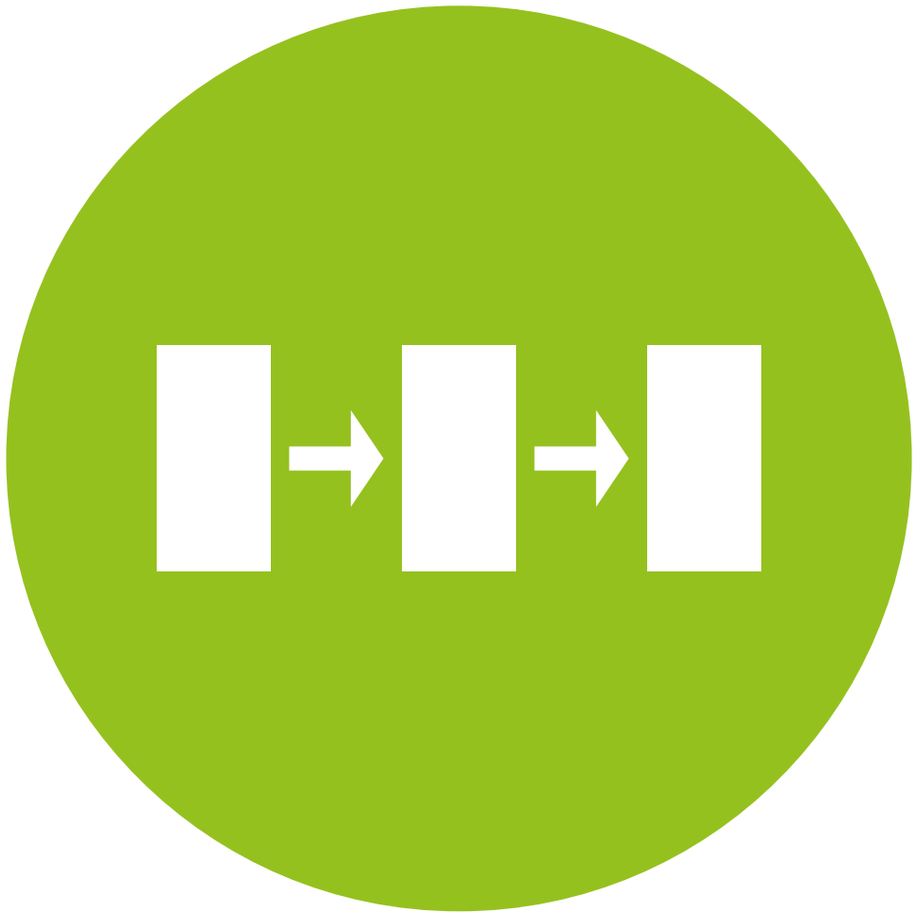
“All models are wrong, but some are useful.”

George Box

Learn more?

Path models are a special case of Structural Equation Models

See the SEM “Learn more?” slide for classes / books / tutorials



Intro to SEM

"The statistical method of the 21st century"

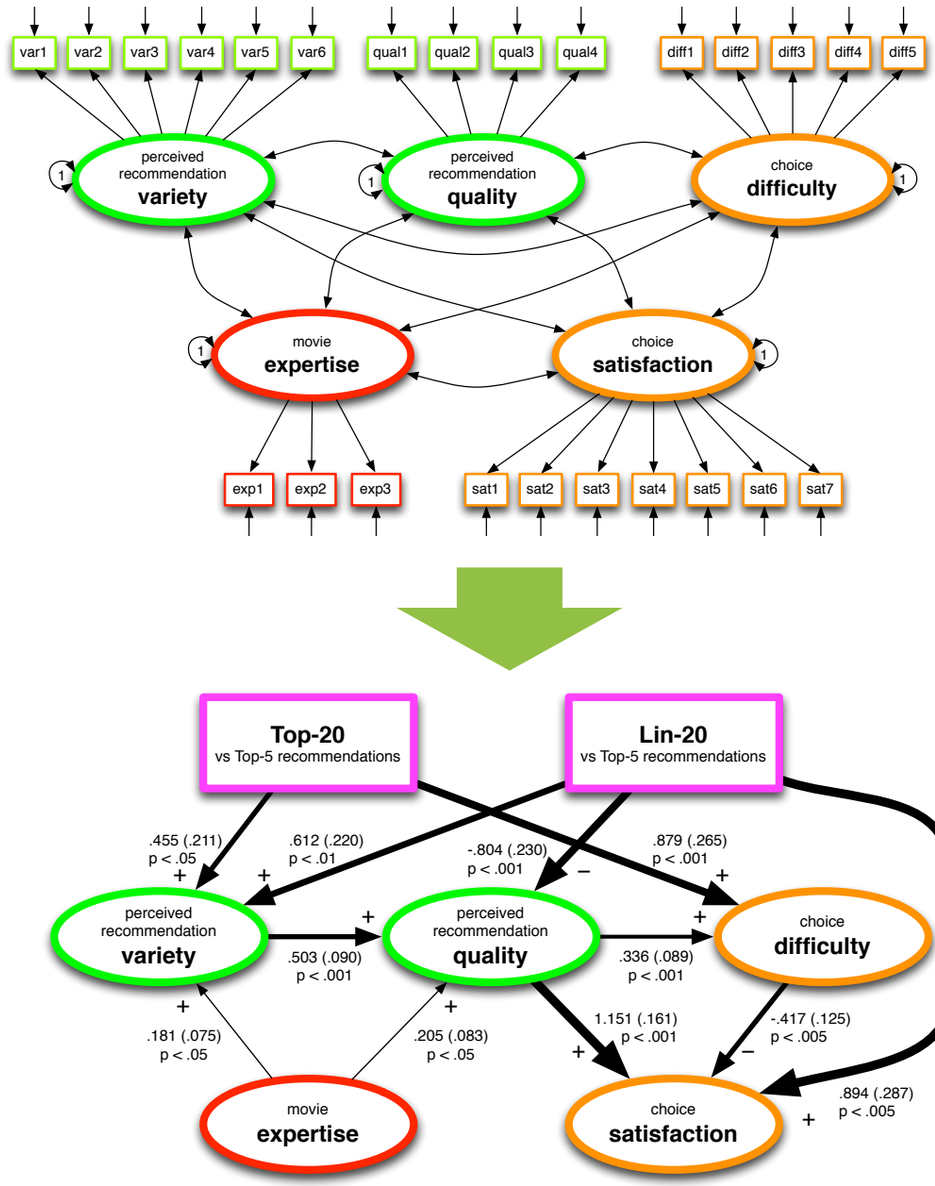
Structural Models

Combine **factor analysis** and **path models**

- Turn items into factors
- Test causal relations

Very **simple reporting**

- Report overall fit + effect of each causal relation
- A path that explains the effects



Example

Example from Bollen et al.: “Choice Overload”

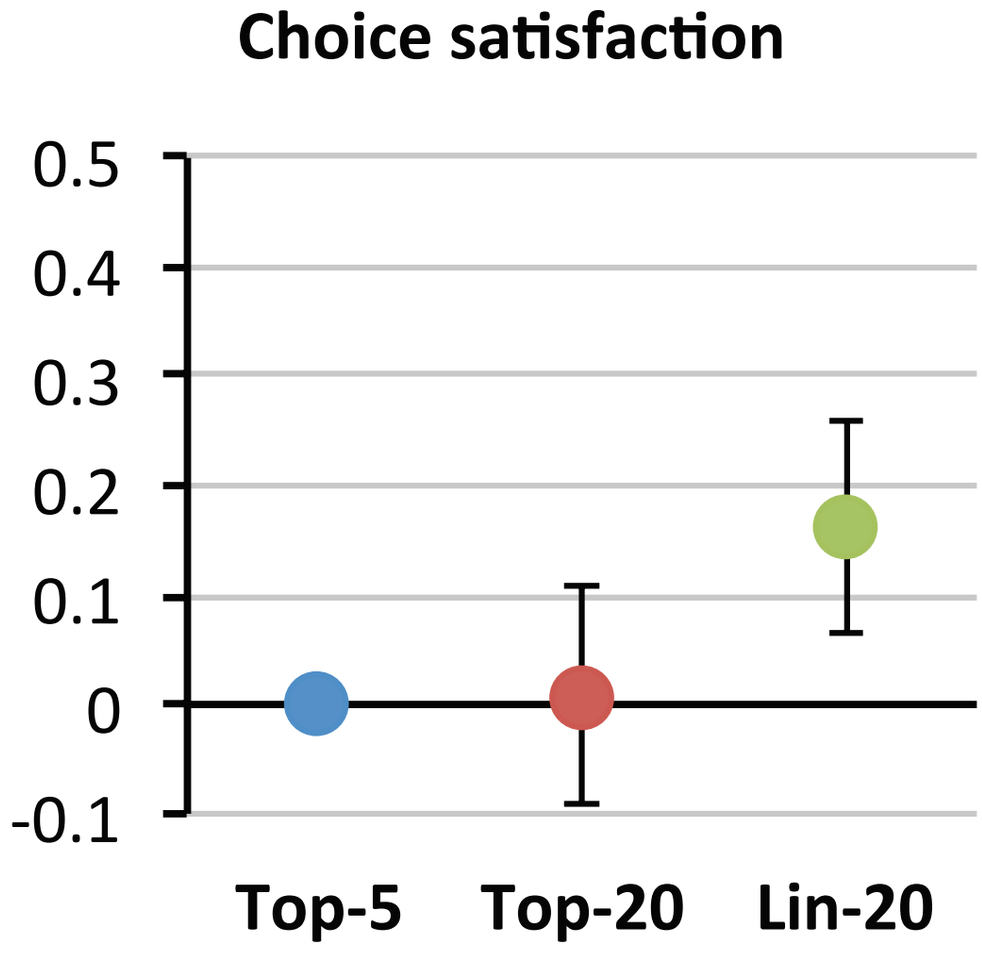
What is the effect of the number of recommendations?

What about the composition of the recommendation list?

Tested with **3 conditions**:

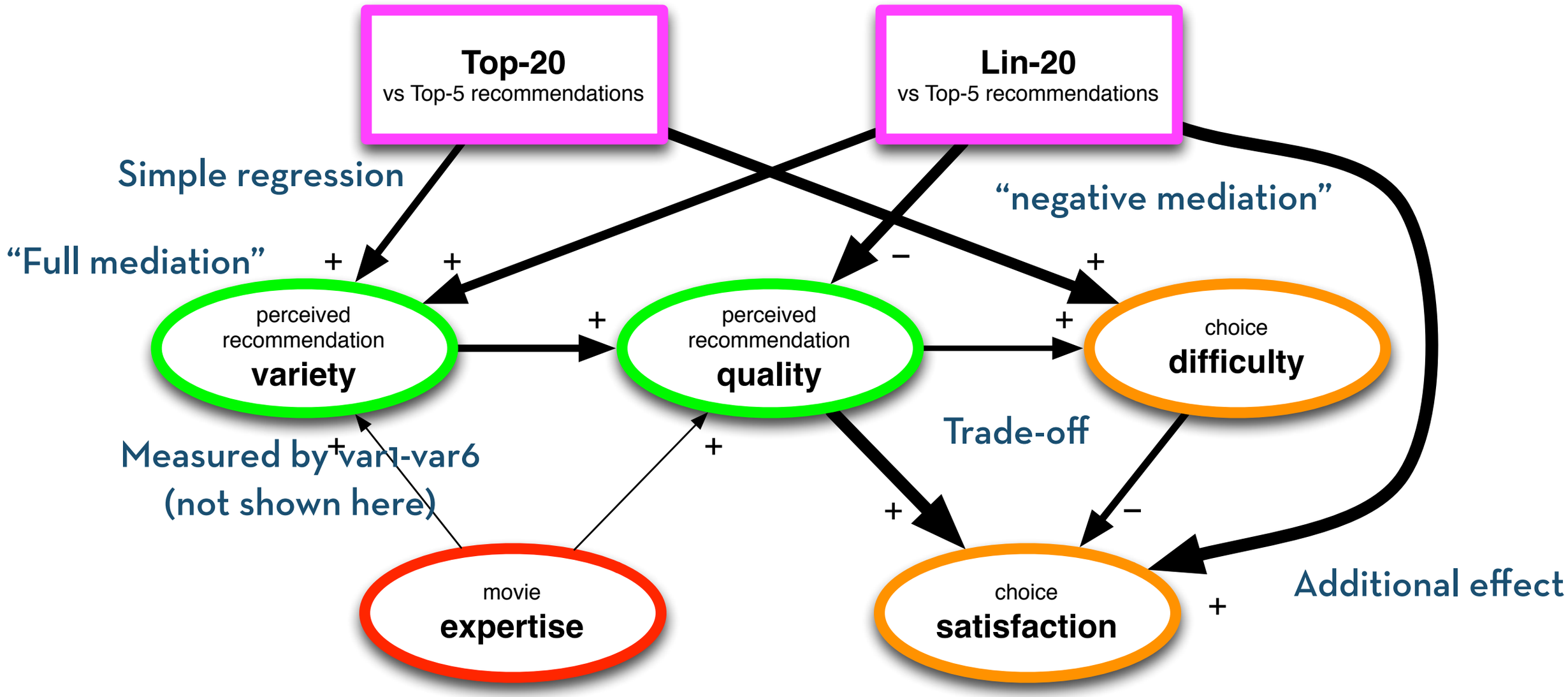
- Top 5:
  - recs: 1 2 3 4 5
- Top 20:
  - recs: 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
- Lin 20:
  - recs: 1 2 3 4 5 99 199 299 399 499 599 699 799 899 999 1099 1199 1299 1399 1499

Example



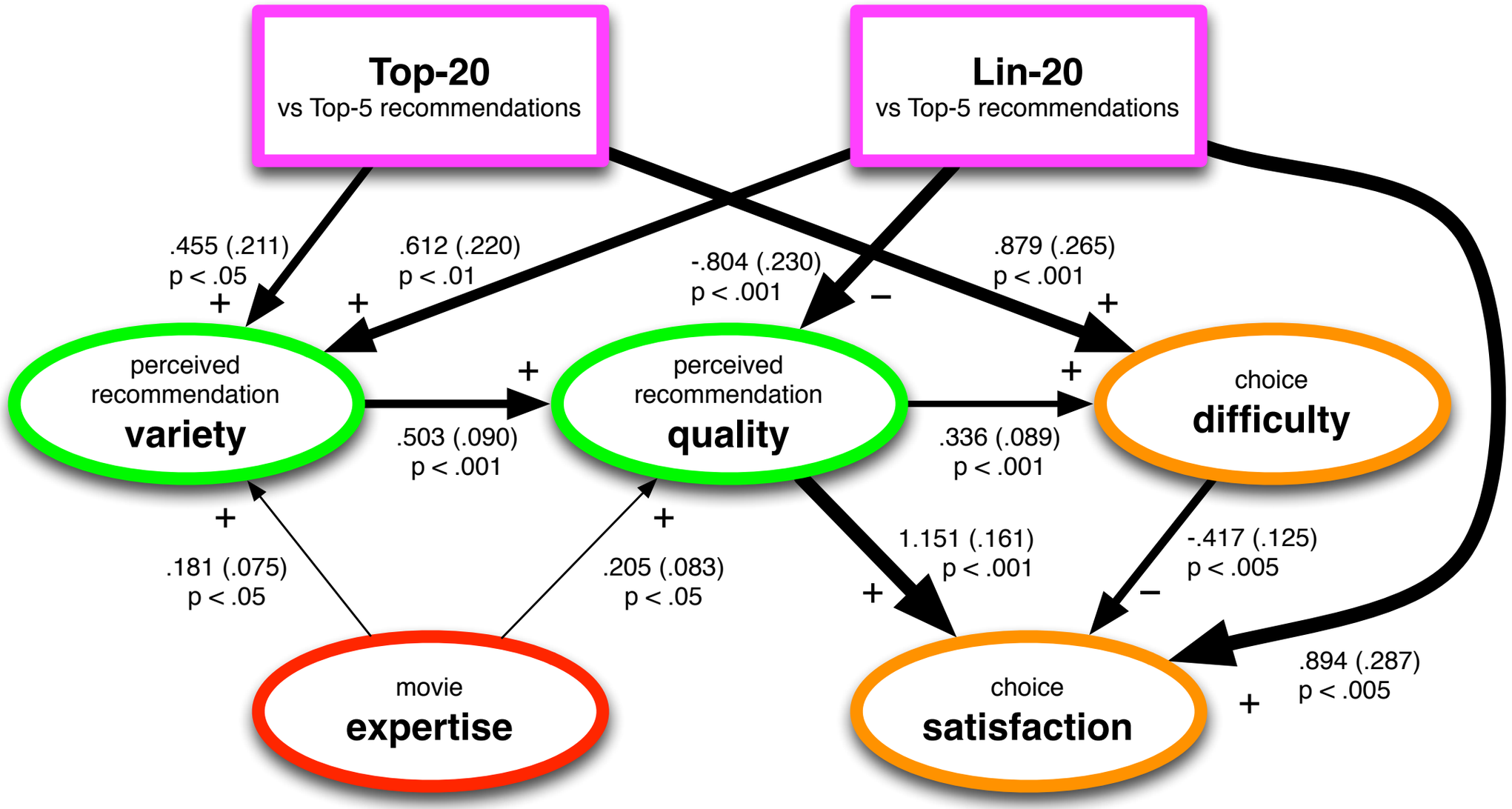
Bollen et al.: “Understanding Choice Overload in Recommender Systems”, RecSys 2010

Example



Bollen et al.: "Understanding Choice Overload in Recommender Systems", RecSys 2010

Example



Bollen et al.: "Understanding Choice Overload in Recommender Systems", RecSys 2010

Learn more?

Take a class (Clemson):

This one! (SEM will be covered after measurement)

PSYC 8730 Structural Equation Modeling in Applied Psychology

HCC 8810 Measurement and Evaluation of HCC systems

Learn more?

Take a class (UC Irvine):

John Hipp: “SocEcol 266A: Structural Equation Modeling” and “SocEcol 275: Structural Equation Modeling II”

George Farkas: “Educ 288B: Structural Equation Modeling”

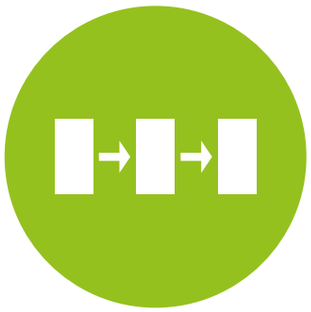
Alex Liu: “Mgmt 291: Structural Equation Modeling”

Learn more?

Learn it yourself:

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

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



Software

What statistical software are we going to use?

Preferred software: MPlus. Free: R with package “lavaan”

Capabilities:

Able to handle non-normal variables

Able to handle repeated measures (lavaan: either or)

Able to handle interactions (some with a trick)

Find total effects, look at mod-indices, etc.

MPlus has great support and course videos

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

T H A N K S !

George Bernard Shaw