Intro to HCI evaluation

Measurement & Evaluation of HCC Systems

Intro

Today's goal:

Give an overview of the mechanics of how (and why) to evaluate HCC systems

Outline:

- Basics of user evaluation
- Selecting participants
- Selecting manipulations
- A look forward

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?

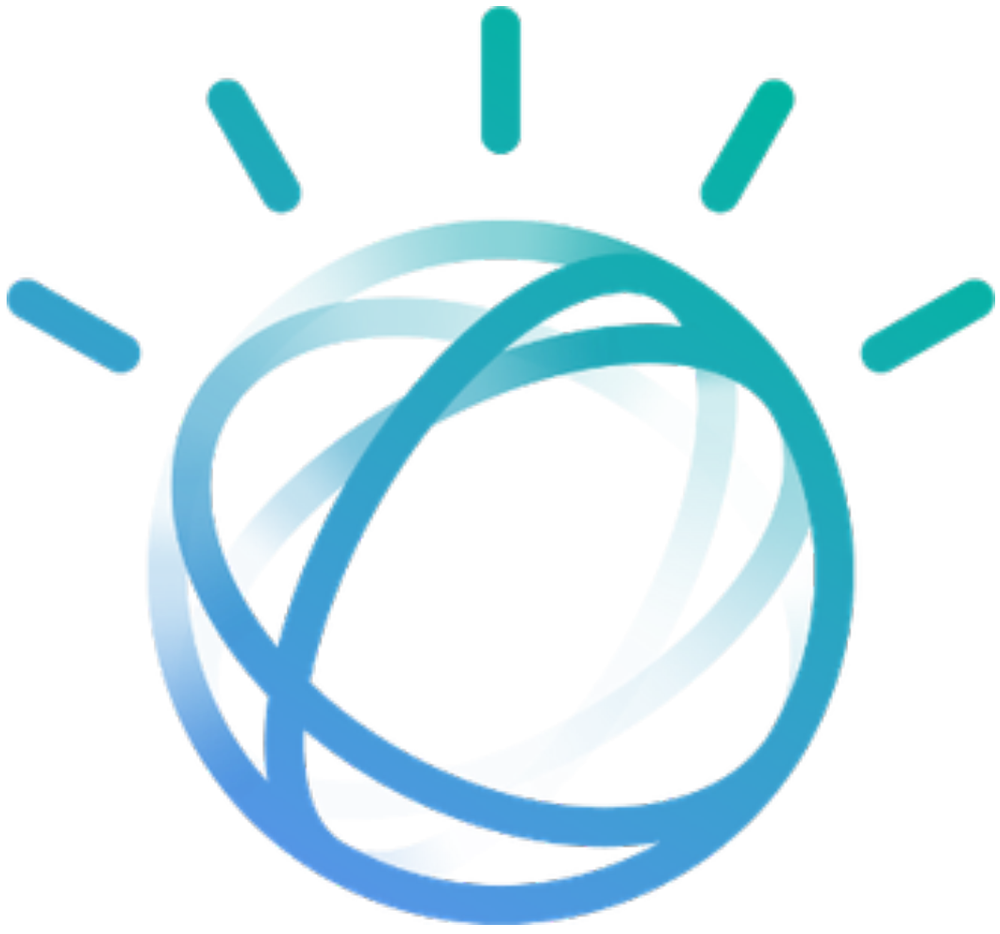
What is the difference between 3.6 and 3.7?

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



System A

System B

However...

Say we find that it scores higher... why does it?

- different skills
- different user models
- different voice

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

New version with **one** added/changed feature

Previous version

Survey/observation

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

Downsides:

Purely correlational

- No manipulations!
- What causes what?

No ceteris paribus

- Hard to get rid of confounding variables

The process

Ideal world:

theory (hypothesis) -> testing -> accepted theory (evidence)

Real world:

theory (hypothesis) -> testing -> completely unexpected results -> interpretation -> revision -> new theory -> ...

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 / Prolific:

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

Crowd-sourcing

Demographics roughly 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

See: Clemson > IRB > Resources > Presentations

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

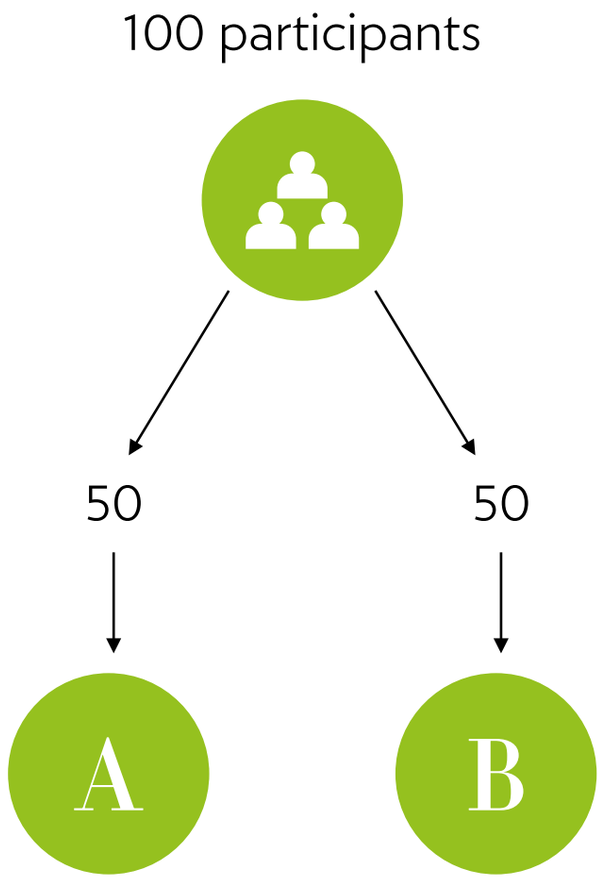
Randomize the assignment of conditions to participants

Randomization neutralizes (but doesn't eliminate) participant variation

Between-subjects

Randomly assign half the participants to A, half to B

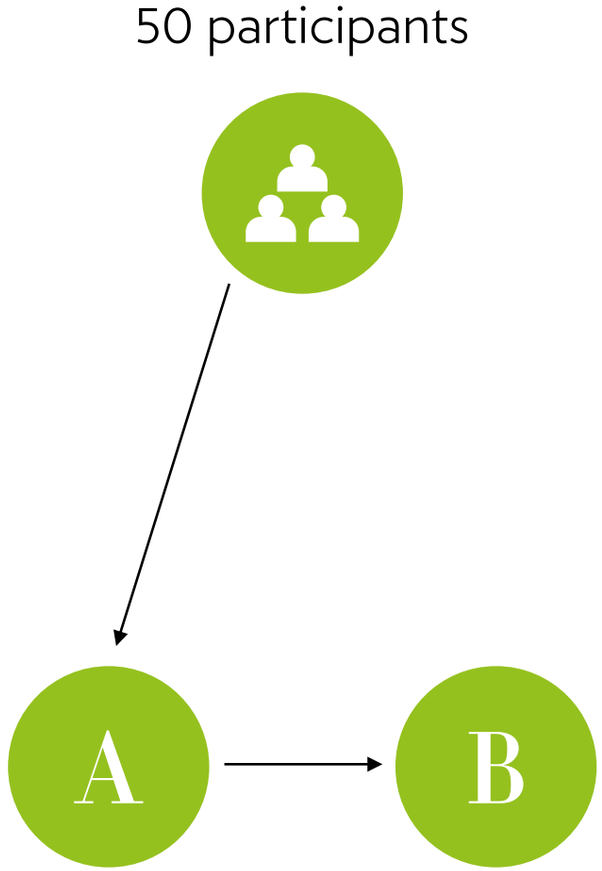
- Realistic interaction
- Manipulation hidden from user
- Many participants needed



Within-subjects

Give participants A first, then B

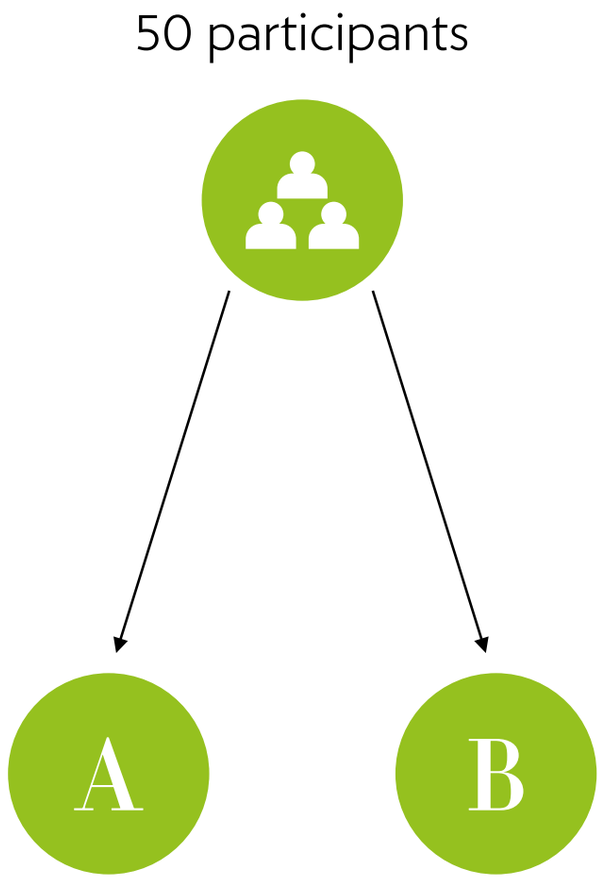
- Remove subject variability
- Participant may see the manipulation (induces demand characteristics)
- Spill-over effect



Within-subjects

Show participants A and B simultaneously

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



Signal + noise

Signal: true difference between A and B

Noise: random variation

- Environment
- Participants
- Measurements

In within-subjects experiments, you get rid of participant noise

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!

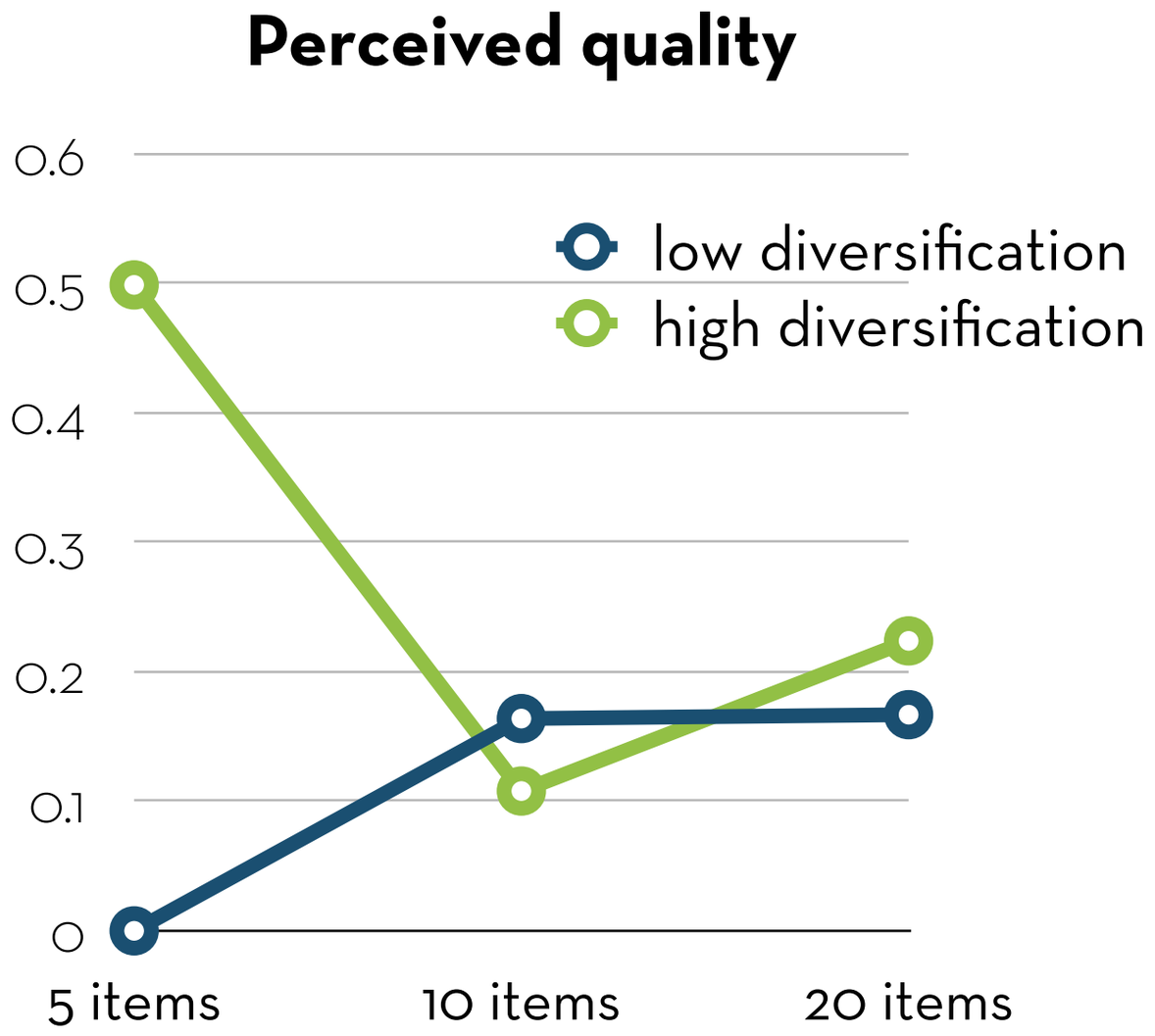
	Low diversity	High diversity
5 items	5+low	5+high
10 items	10+low	10+high
20 items	20+low	20+high

AB Factorial designs

Allows you to test **interaction effects**

Is the effect of diversification different per list length?

Is the effect of list length different for high and low diversification?



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

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)

A look forward

What are we going to learn?

Variables

Independent variables (X): things that are manipulated (experiment) or innate (survey)

- Low vs. high diversity
- Number of search results
- Gender
- Age

They are outside the participants' control (in the experiment)

Variables

Dependent variables (Y): things that are measured as an outcome of X

- Number of clicks
- Interaction time
- Facial expression
- Satisfaction*

Variables

Random variables (also X): variables that are not of interest, but they may influence Y, so we measure them just in case.

Control variables (not X): variables that are not of interest, but they may influence Y, so we try to keep them stable

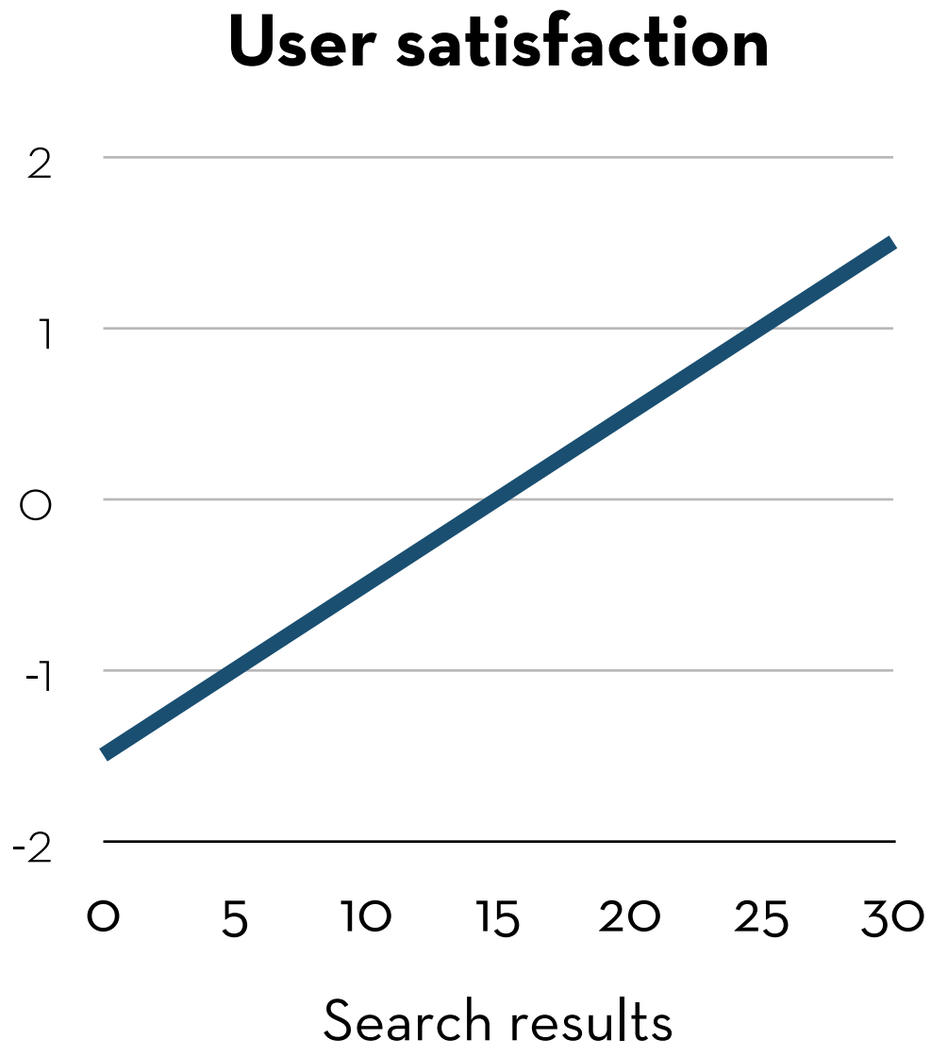
Regression

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?



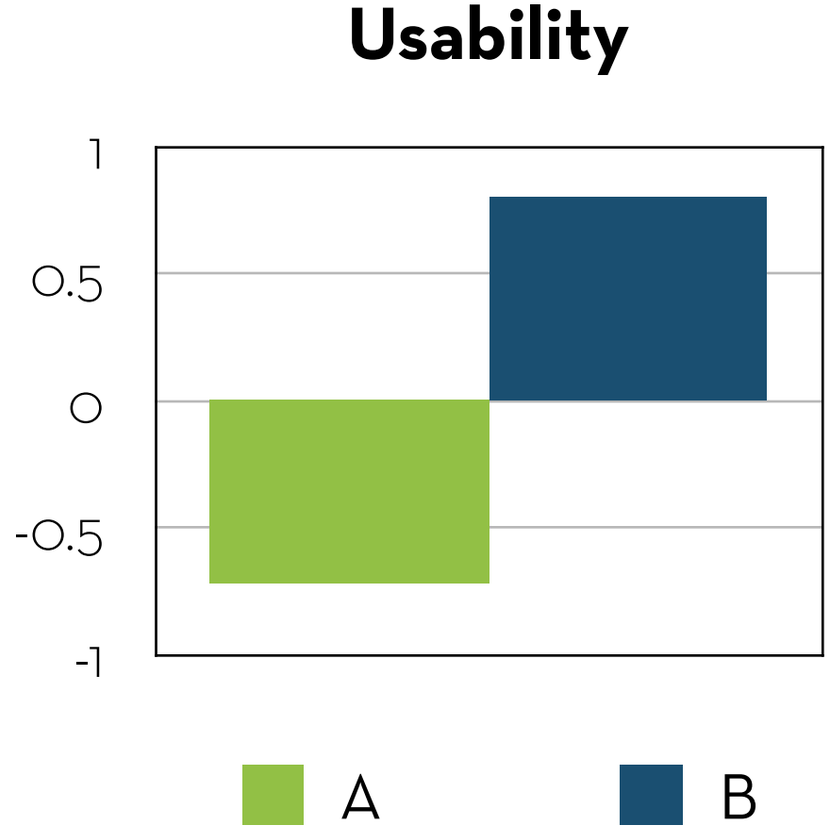
T-test

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?



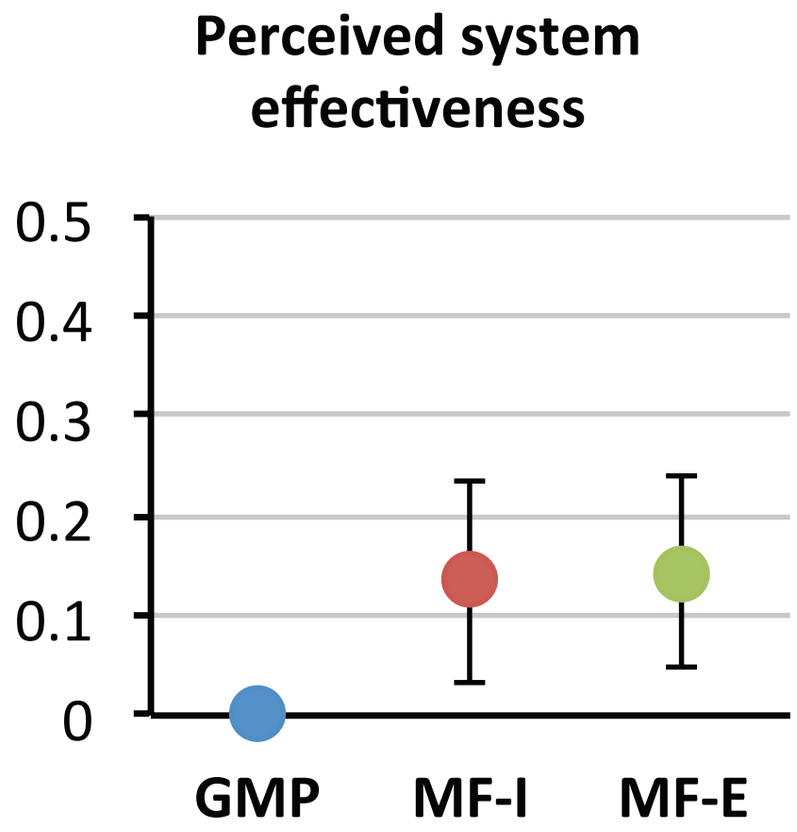
ANOVA

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!

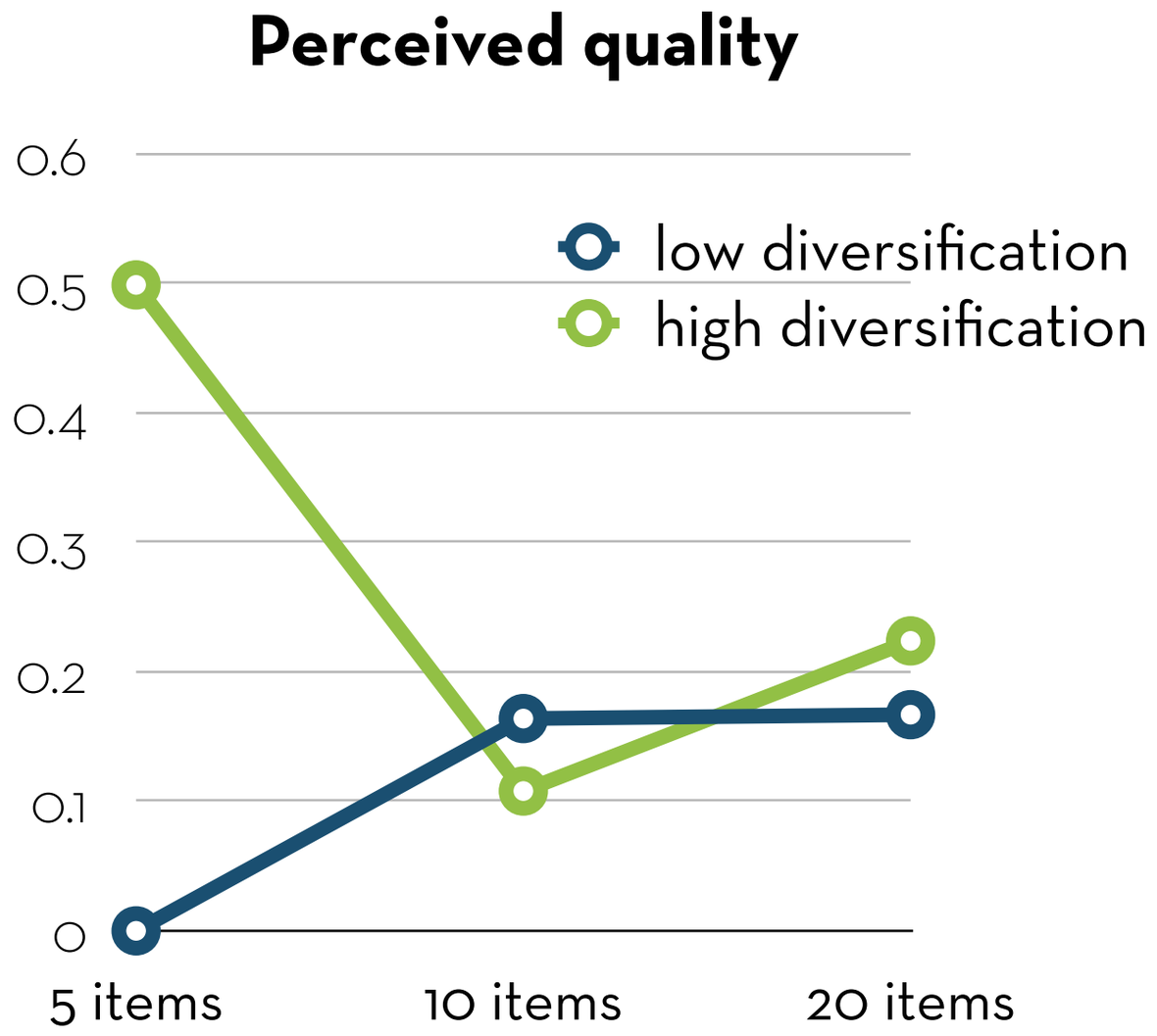


Factorial ANOVA

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", UMUAI

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

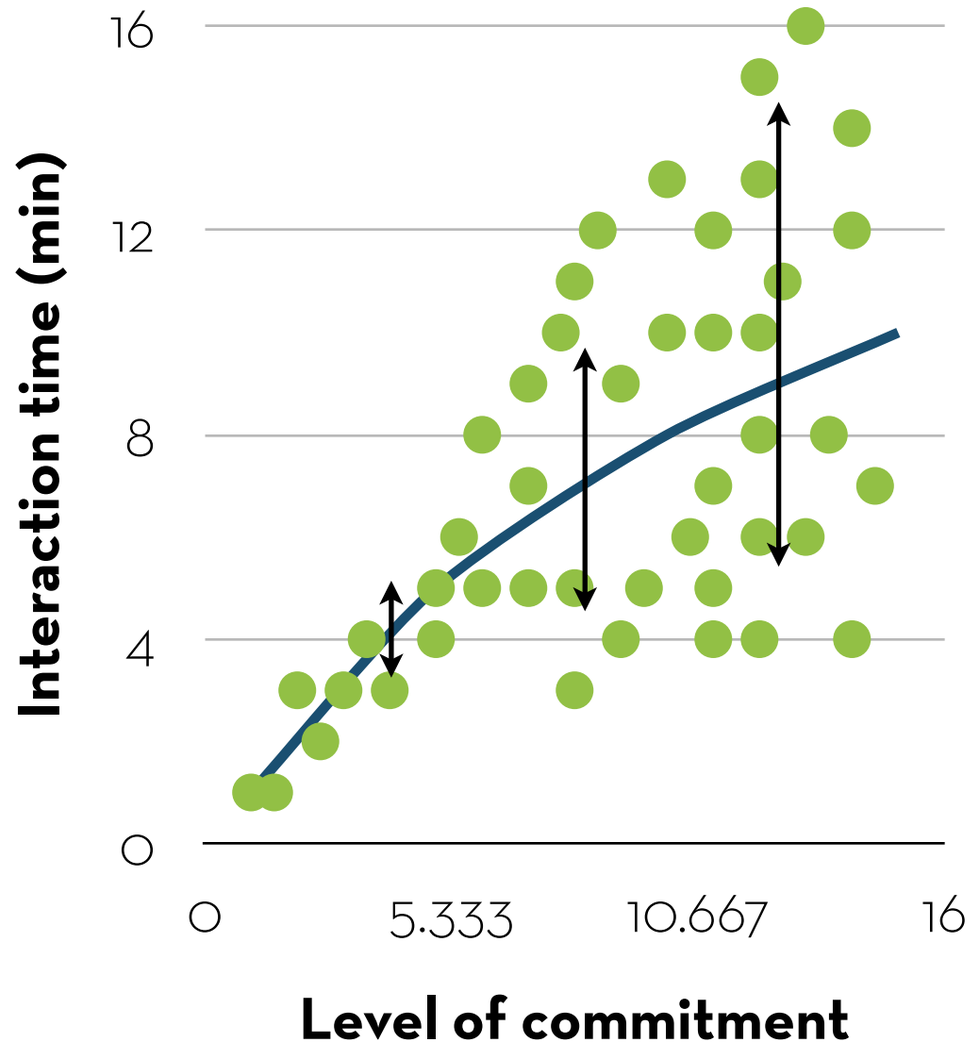
Not true for most behaviors!

Number of clicks

Time, money

1-5 ratings

Decisions



Correlated errors

Standard regression requires uncorrelated errors

This is not the case when...

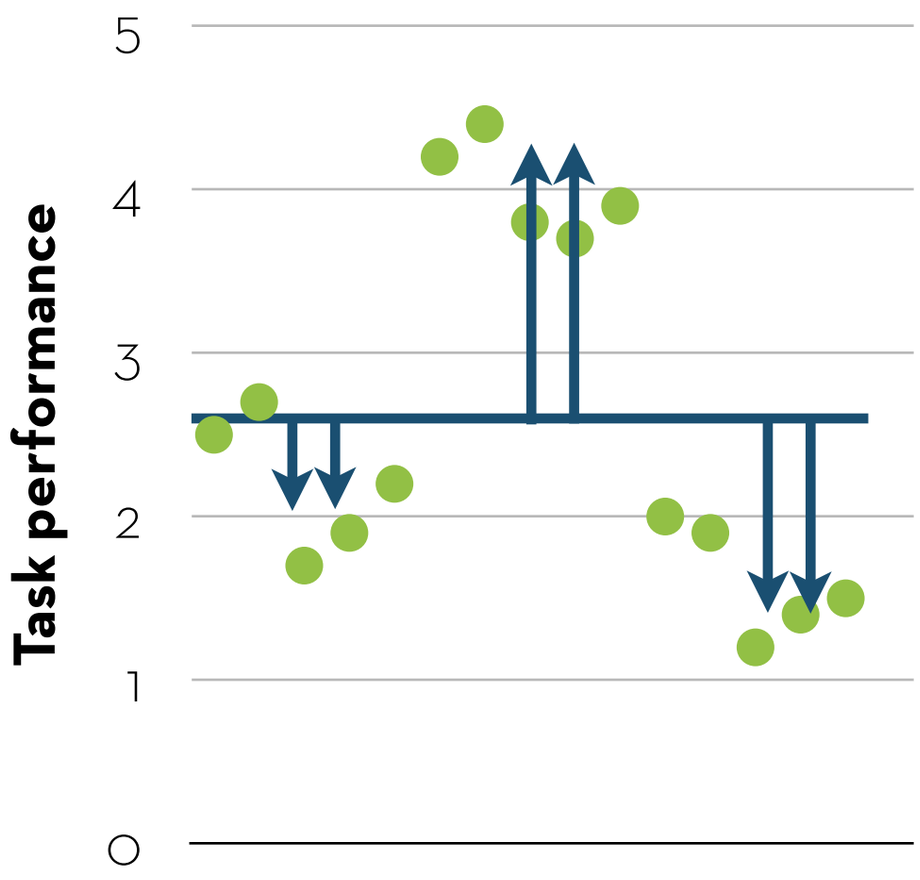
...you have repeated measurements of the same participant (e.g. you measured 5 task performance times per participant, for 60 participants)

...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)



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

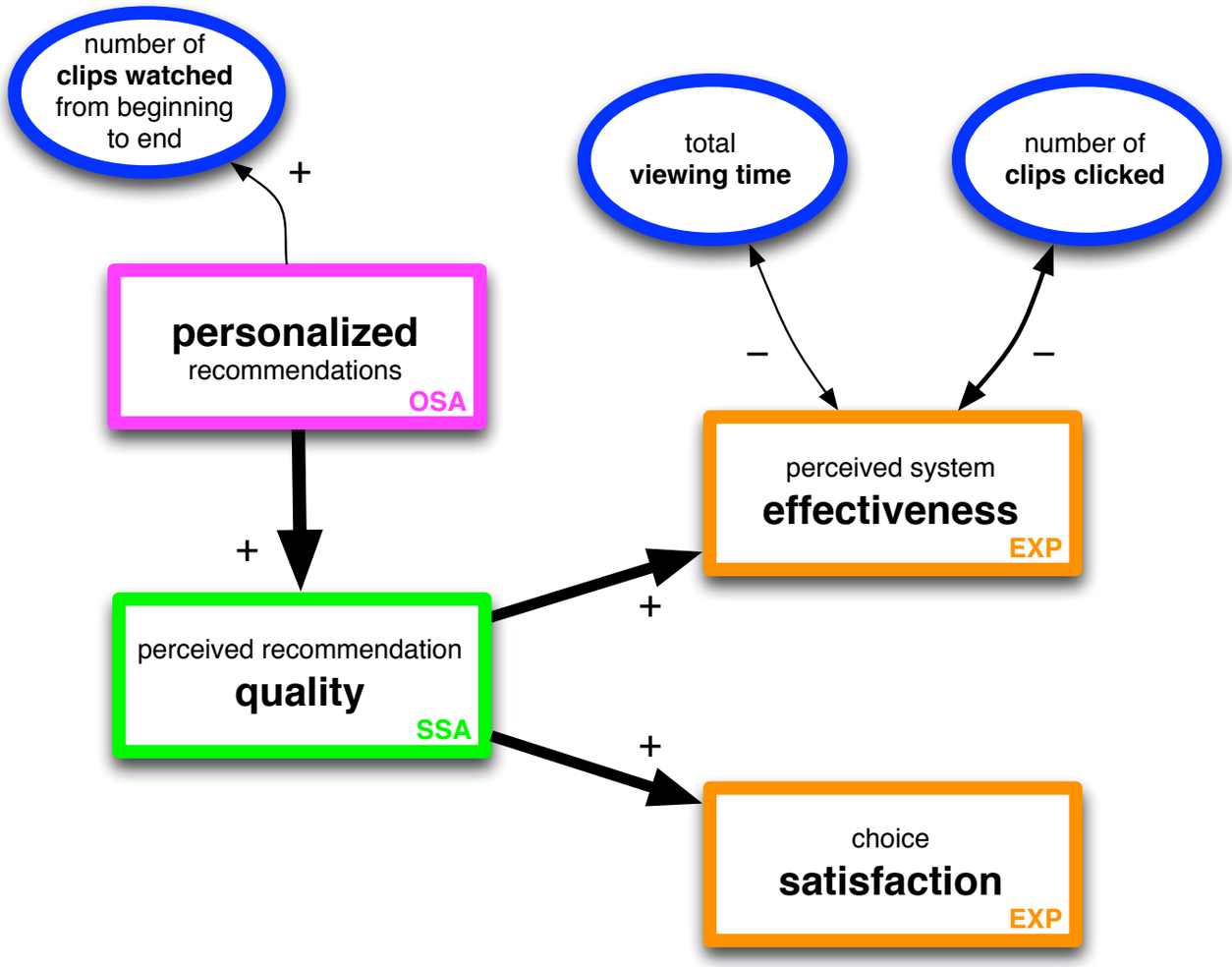
Why should we measure these things?

And how can we quantify them?

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

How to quantify?

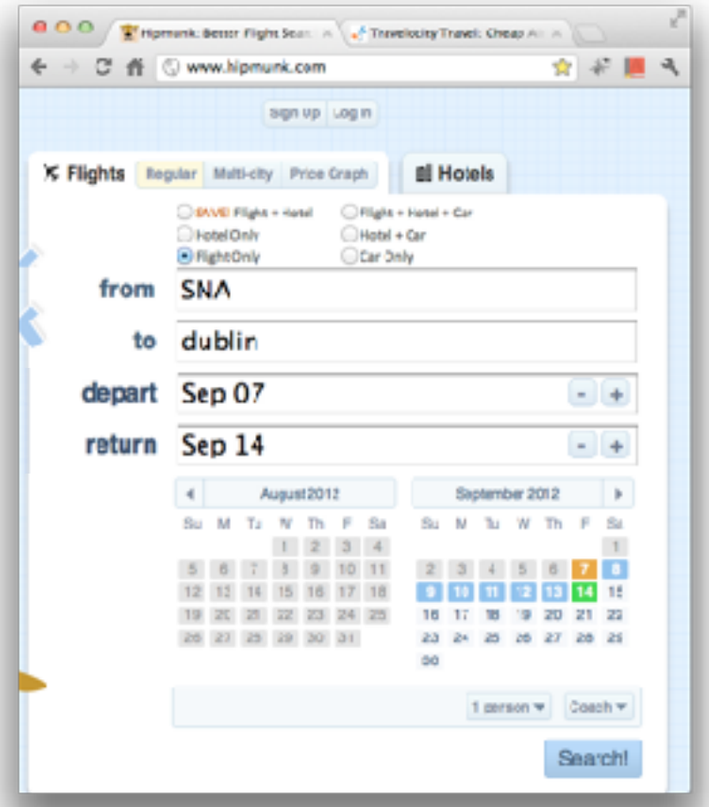
Psychometrics:

Ask multiple questions on a 5- or 7-point scale

E.g. 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"

Use factor analysis to validate the scales



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

Theory behind x->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

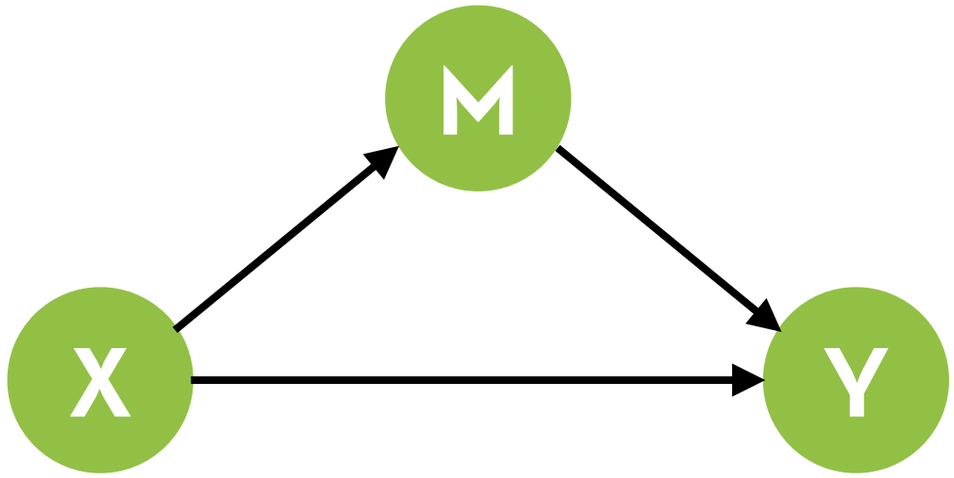
- Find out how they mediate the effect on usability

Evaluate the data using structural equation modeling

Mediation analysis

Manipulation -> perception
-> experience

Does the system
influence usability
via understandability?



Types of mediation

Partial mediation

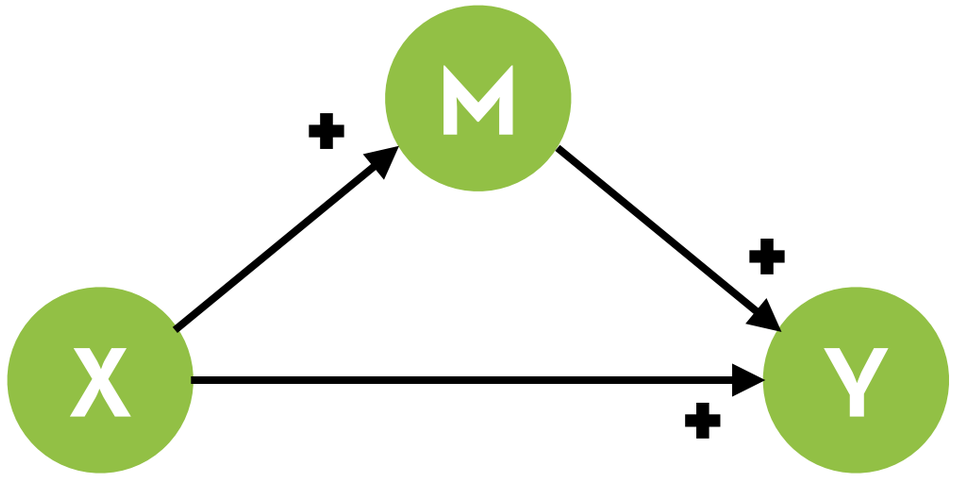
Full mediation

Negative mediation

Mediation Analysis

Manipulation -> perception
-> experience

Does the system
influence usability
via understandability?



Types of mediation

Partial mediation

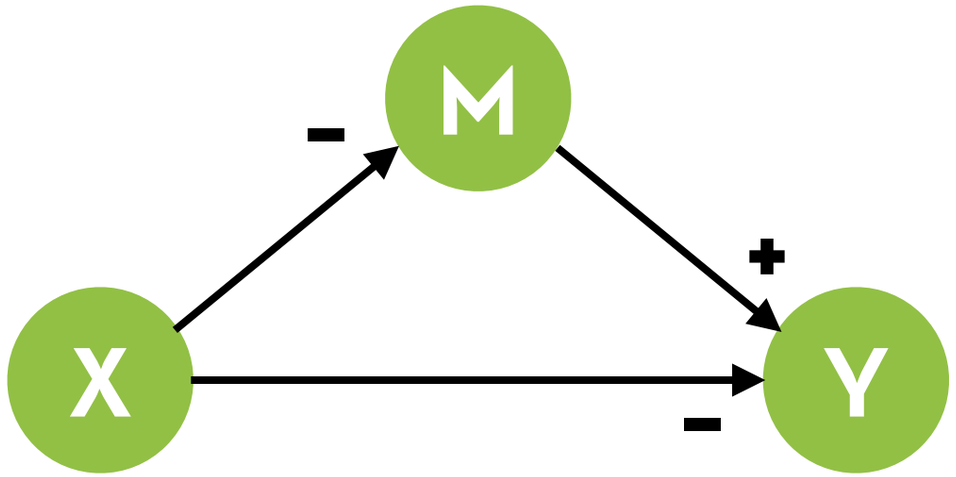
Full mediation

Negative mediation

Mediation Analysis

Manipulation -> perception
-> experience

Does the system
influence usability
via understandability?



Types of mediation

Partial mediation

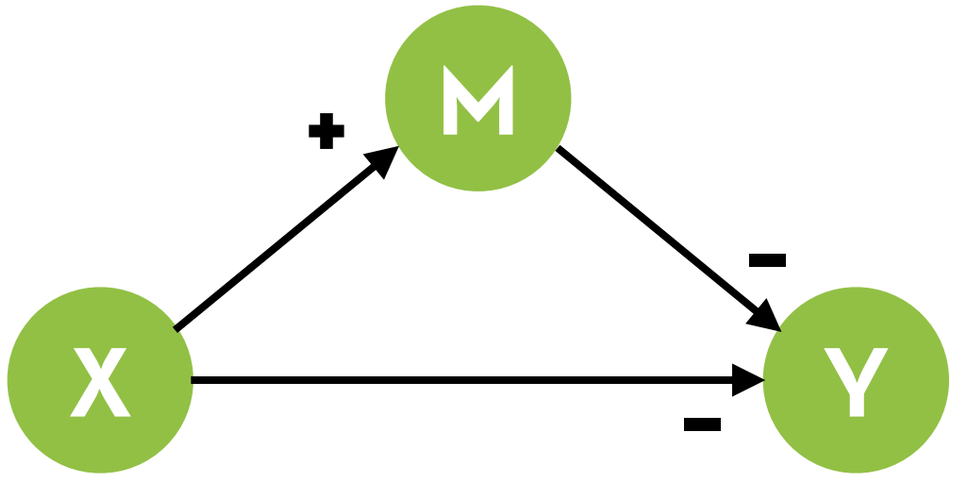
Full mediation

Negative mediation

Mediation Analysis

Manipulation -> perception
-> experience

Does the system
influence usability
via understandability?



Types of mediation

Partial mediation

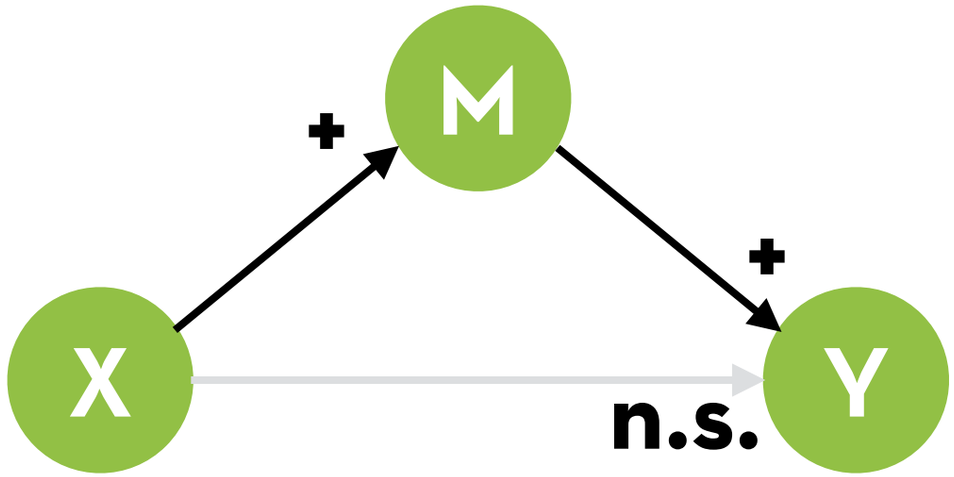
Full mediation

Negative mediation

Mediation Analysis

Manipulation -> perception
-> experience

Does the system
influence usability
via understandability?



Types of mediation

Partial mediation

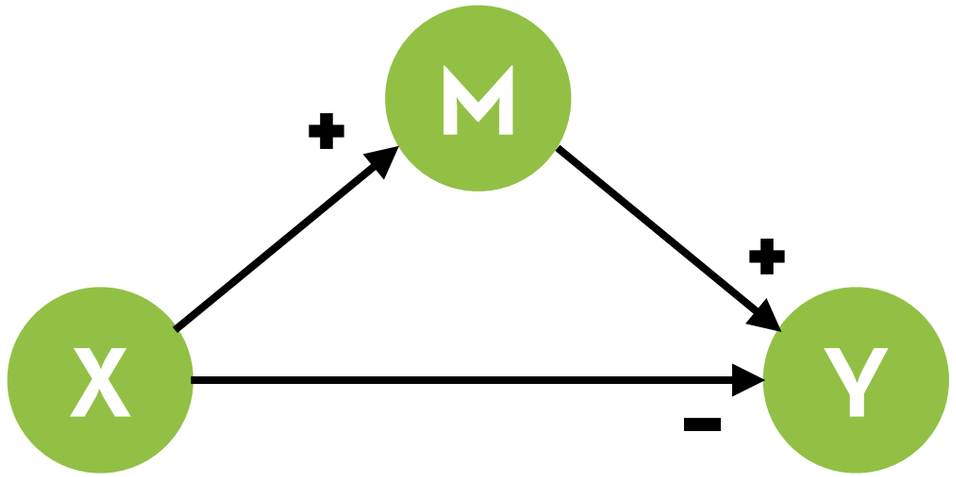
Full mediation

Negative mediation

Mediation Analysis

Manipulation -> perception
-> experience

Does the system
influence usability
via understandability?



Types of mediation

Partial mediation

Full mediation

Negative mediation

Example

6 jams

Less attractive
30% sales
Higher choice satisfaction

ıst

24 jams

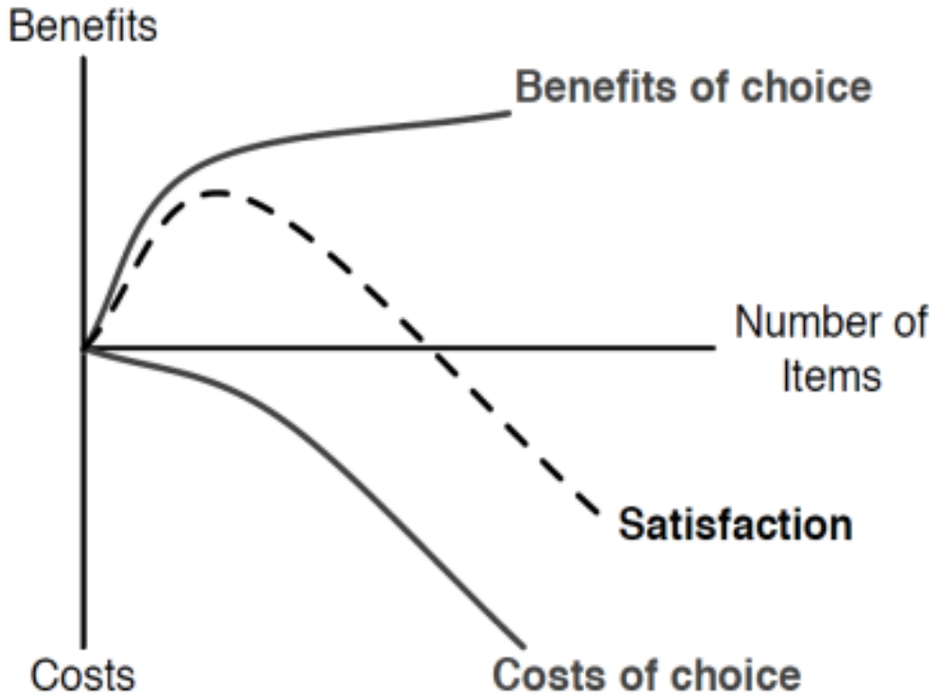
More attractive
3% sales
Lower choice satisfaction

Example

Satisfaction = benefit – cost

Benefit of more options: easier to find the right option

Cost of more options: more comparisons, higher potential regret



Is this also true for recommendations?

Example

Example from Bollen et al.: “Choice Overload”

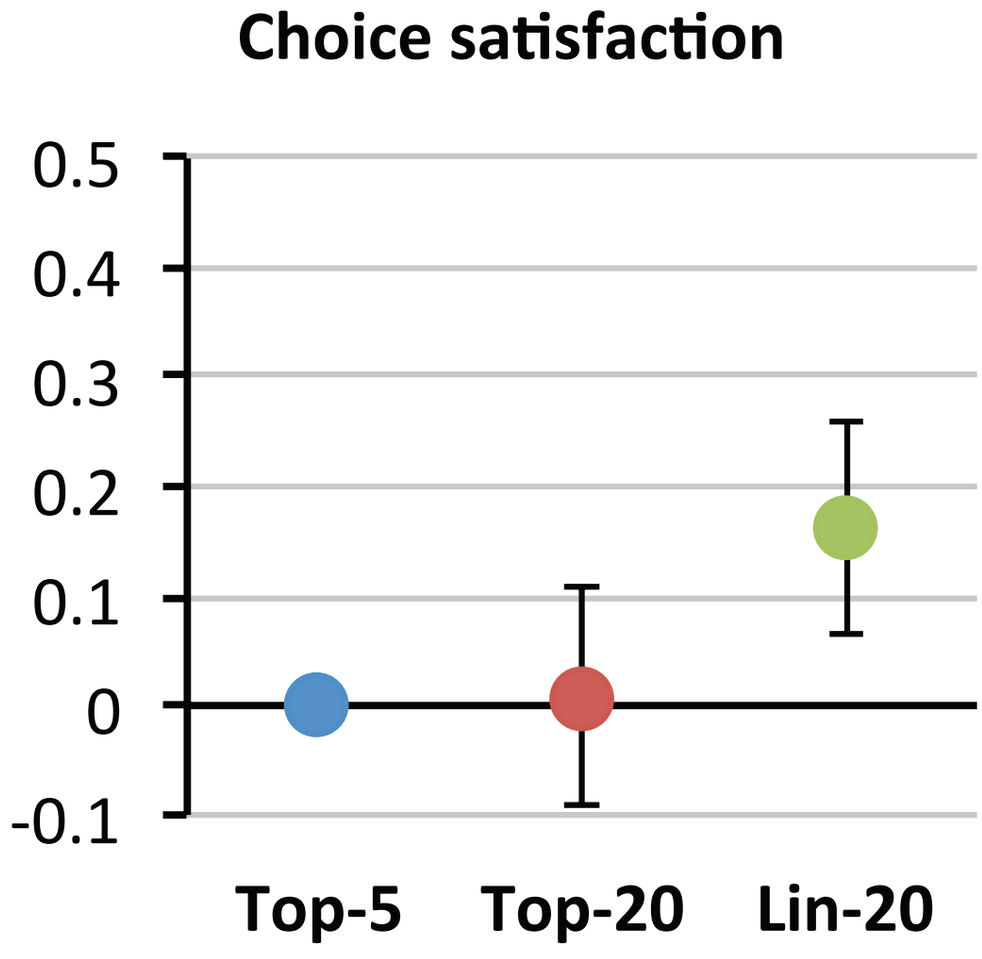
What is the effect of the number of recommendations?

What about the composition of the recommendation list?

Tested with 3 conditions:

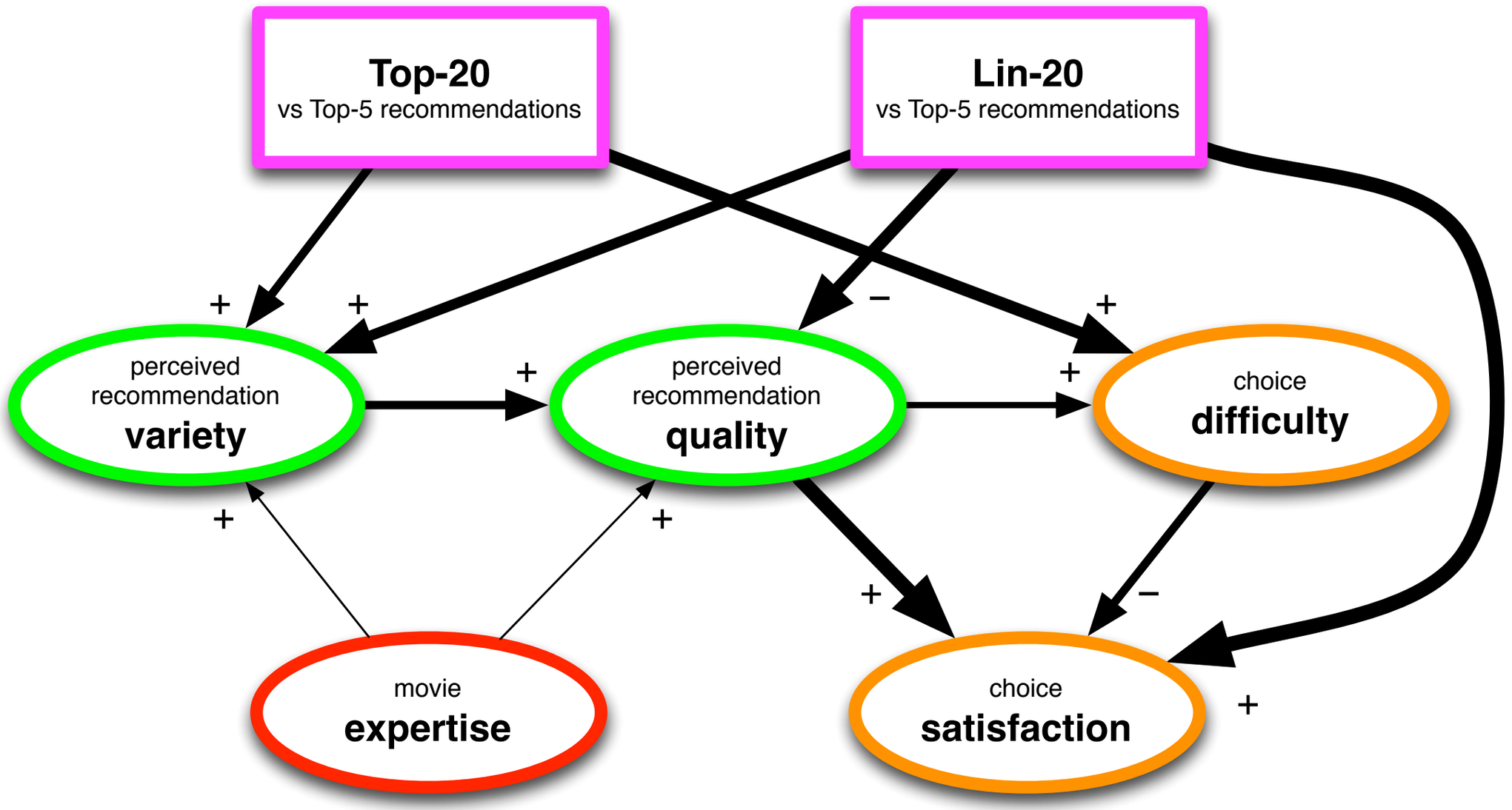
- Top 5:
  - recs: 1 2 3 4 5
- Top 20:
  - recs: 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
- Lin 20:
  - recs: 1 2 3 4 5 99 199 299 399 499 599 699 799 899 999 1099 1199 1299 1399 1499

Example



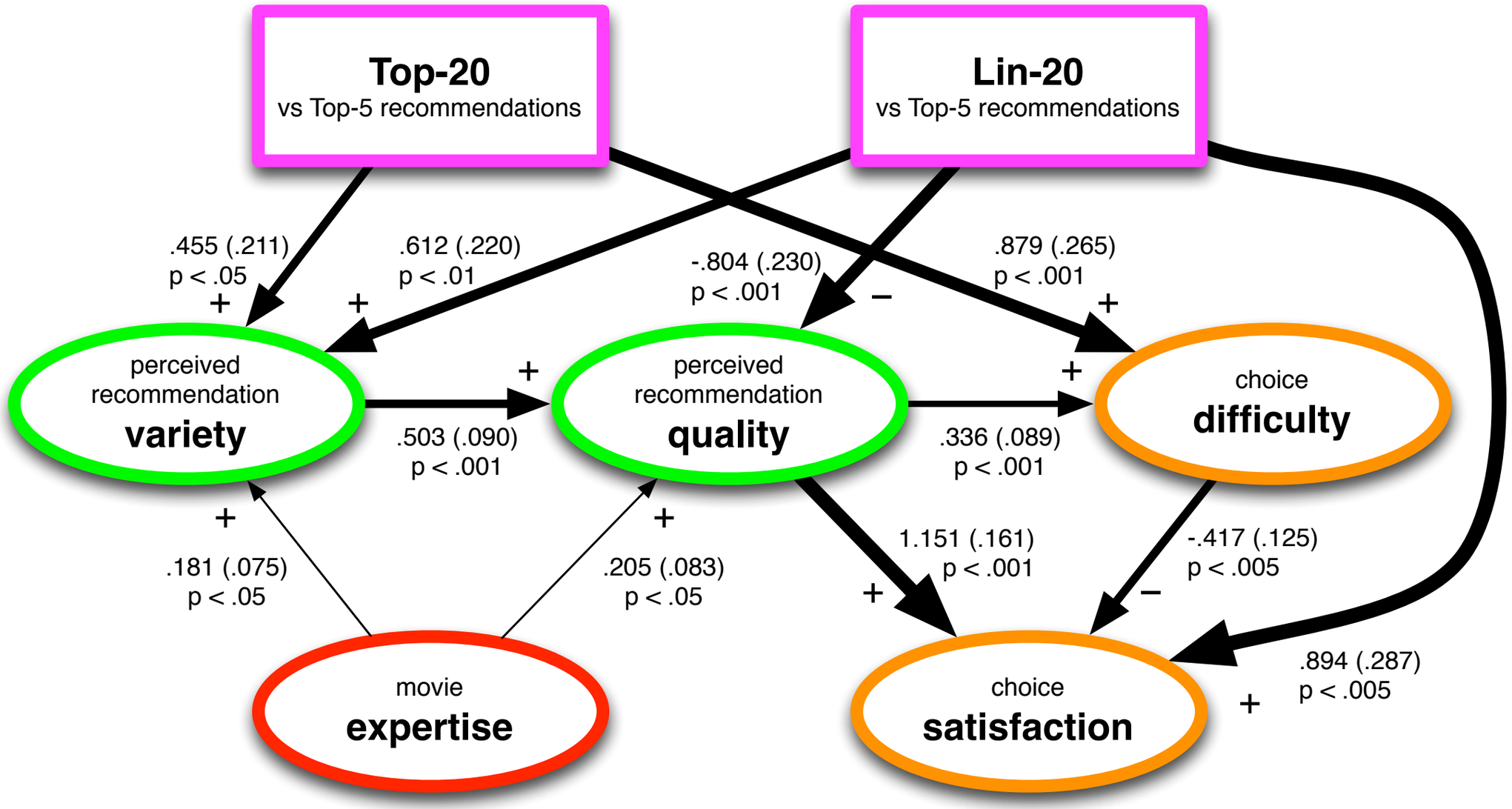
Bollen et al.: “Understanding Choice Overload in Recommender Systems”, RecSys 2010

Example



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

Example



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

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

T H A N K S !

George Bernard Shaw