User Experiments

Adaptive Decision Support Systems

User Experiments

Today's goal:

Introduce you to user experiments

Outline:

- What is a user experiment?
- How to run a user experiment
- From research questions to hypotheses
- A number of examples

User experiments

What is a user experiment?

User experiments

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

Systems can be anything:

- Software
- Hardware
- Other people
- Organizations
- Policies

What to ask?

“Is my new travel system 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...

“Does the user interface of my travel system score 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?

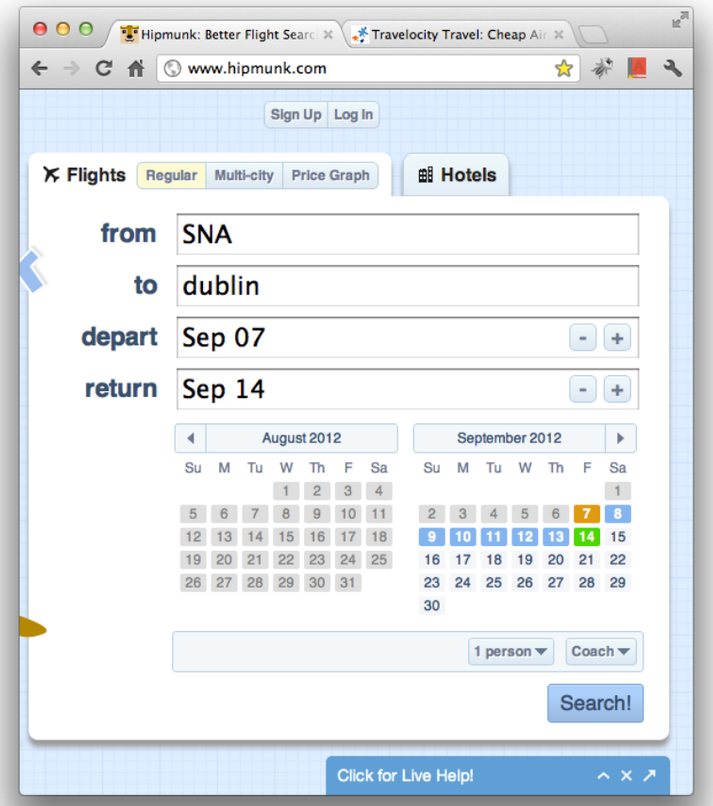
What is the difference between 3.6 and 3.7?

We need to compare the UI against something

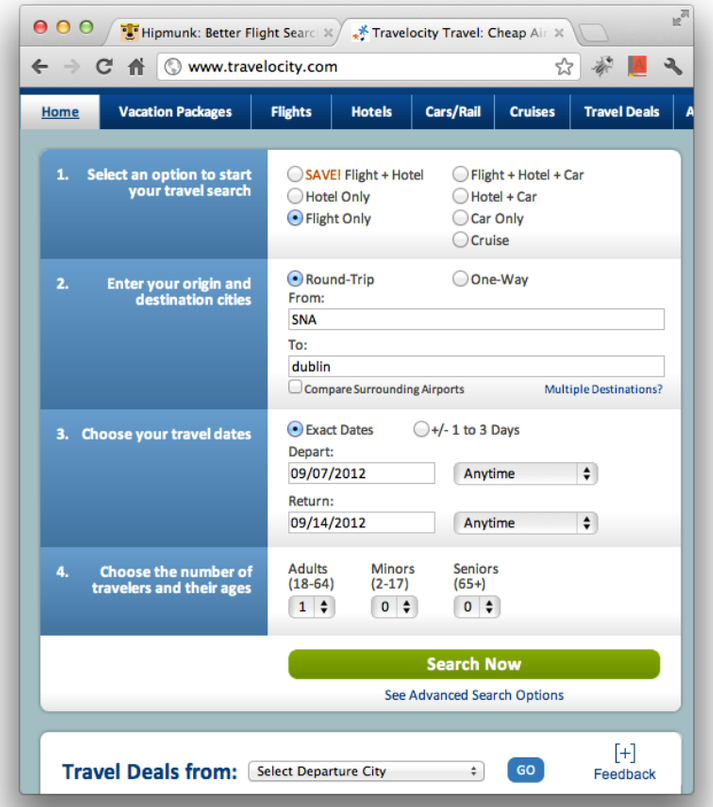
Even better...

“Does the UI of my system score 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?

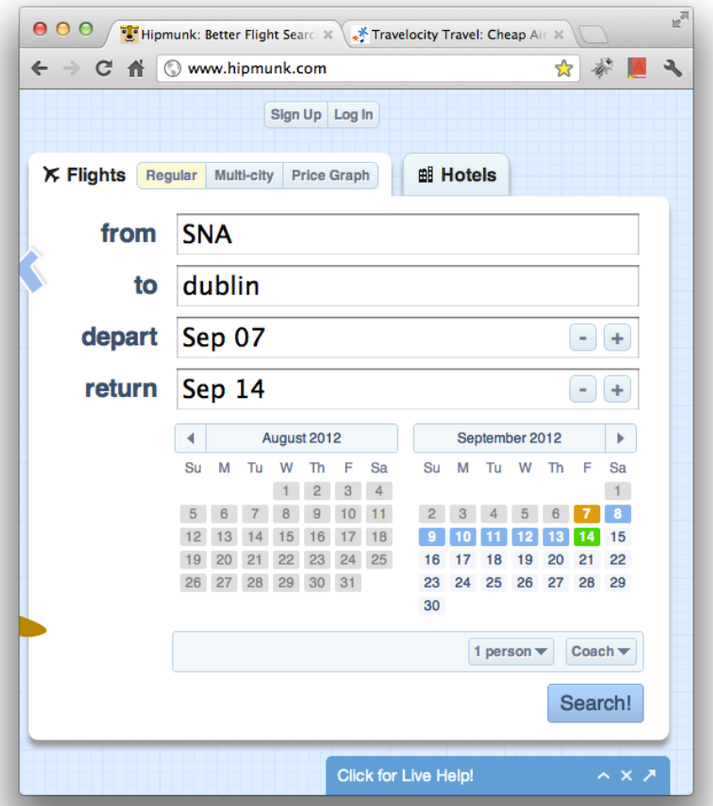
- different date-picker method
- different layout
- different number of options available

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

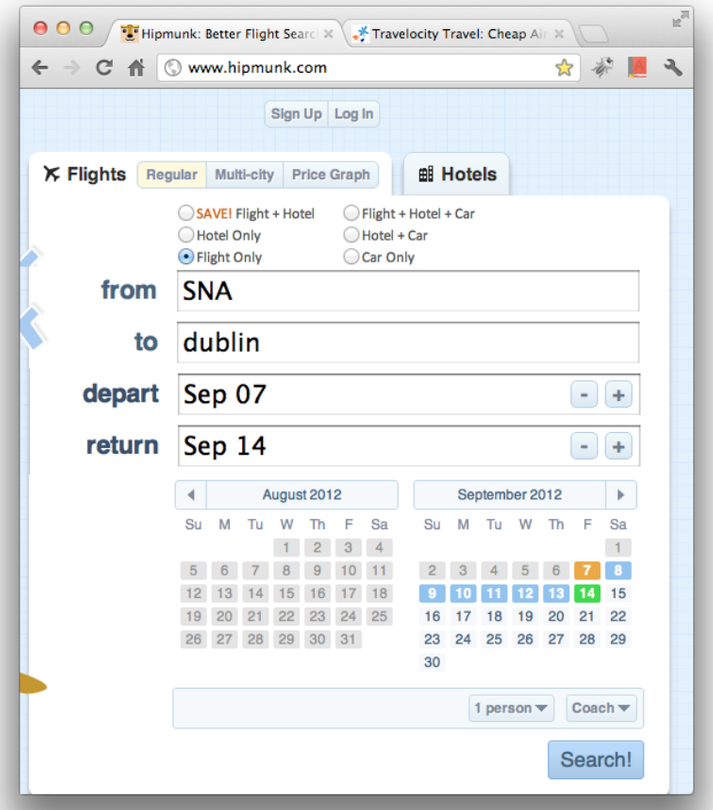
Keep everything the same, except for the thing you want to test (the manipulation)

Any difference can be attributed to the manipulation

Ceteris Paribus



My new travel system



Previous version
(too many options)

How it works

Create multiple versions of your system (intervention and control)

Recruit participants to take part in your study (a sample taken from a population)

You let people use one (or all) of these versions

You measure their behaviors and/or subjective evaluations (outcome)

You statistically evaluate the difference in outcome between intervention and control

Core components

“A user experiment uses a sample of participants to systematically test how different system aspects (manipulations) influence the users’ experience and behavior (observations).”

Participants

Let's say we test our system on our colleagues or contacts...

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

Participants

Let's say we only use data from frequent users...

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

Participants

What's a good **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

Manipulations are the things that you believe will "make a difference"

- Also called "independent variables"

In HCC research, our main manipulations are system aspects

- We are not testing entire systems, but system aspects

Each manipulation consists of multiple conditions (different versions of the system aspect)

- Simplest variant, two conditions: intervention and control

Manipulations

Examples of manipulations:

- Recommendations vs. random items
- Number of recommendations: 5, 10, or 20
- Comic-based privacy policy vs. text-based privacy policy
- Comic length: short, medium, long

Manipulations

You can combine multiple manipulations in a single study!

Type of privacy policy (comic, text) X length of policy (short, medium, long)

In this case, experimental conditions multiply:

Short comic, medium comic, long comic, short text, medium text, long text

Don't go overboard!

Required sample size depends on the number of conditions!

Observations

Observations are the means by which you measure the differences between conditions

Also called “dependent variables” or “outcomes”

In HCC experiments, our observations are either objective or subjective

They are **always** quantitative (more on this next week)

Observations

Examples of observations:

Objective:

- Number of clicks
- Privacy knowledge (# correct answers on a privacy quiz)

Subjective

- Perceived privacy protection
- Perceived system effectiveness

Systematic

If you apply the concept of ceteris paribus...

...and you randomly assign participants to conditions...

...then any difference you find can be attributed to the manipulation!

i.e., the difference between the conditions!

The power of user experiments lies in the ability to make such causal inferences!

Survey/observation

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

Downsides:

Purely correlational

- No manipulations!
- What causes what?
- Third variable problem

No ceteris paribus

- Hard to get rid of confounding variables

Hypotheses

How do we translate research questions into hypotheses?

Research questions

A research question defines the objective of a study

Benefits for you, before doing the study:

- It distills your research topic area into a specific goal
- It motivates the design of your study
- It drives the development of testable hypotheses

Benefits for you, after the study:

- It motivates the analysis of collected data
- It scopes the argumentation of your paper

Research questions

A research question defines the objective of a study

Benefits for others, before the study:

- It explains what question you set out to answer in a study proposal (to IRB, grant reviewers, potential collaborators)

Benefits for others, after the study:

- It explains to readers what they will learn from your paper
- It explains to listeners what they will learn from your presentation

Hypotheses

Experiments can answer causal research questions

i.e., how one variable influences the other

Example: Does a comic-based privacy policy increase privacy awareness compared to a text-based policy?

Hypotheses are predictions regarding the influence of your independent variables (manipulations) on your dependent variables (outcomes)

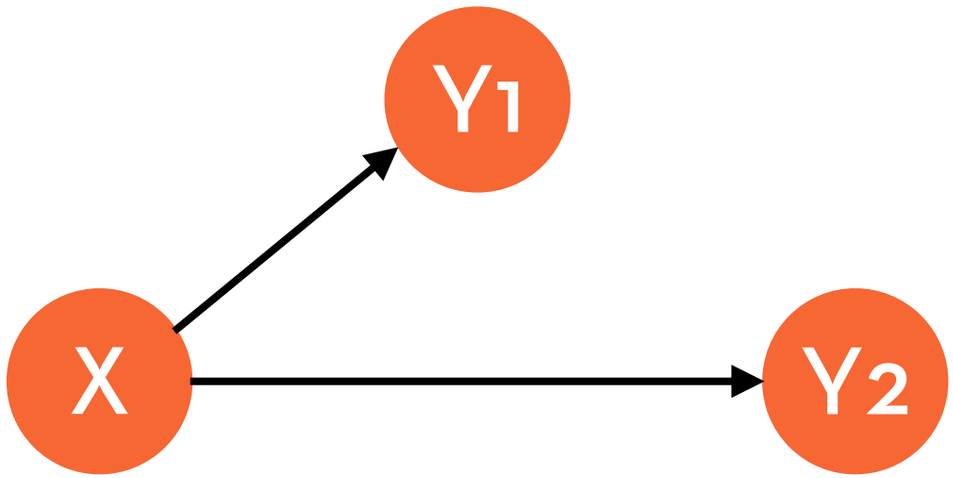
Example: compared to text, comic-based policies increase privacy knowledge

Causal relations

X = comics (vs. text)

Y1 = privacy knowledge
(# correct quiz questions)

Y2 = perceived privacy
protection



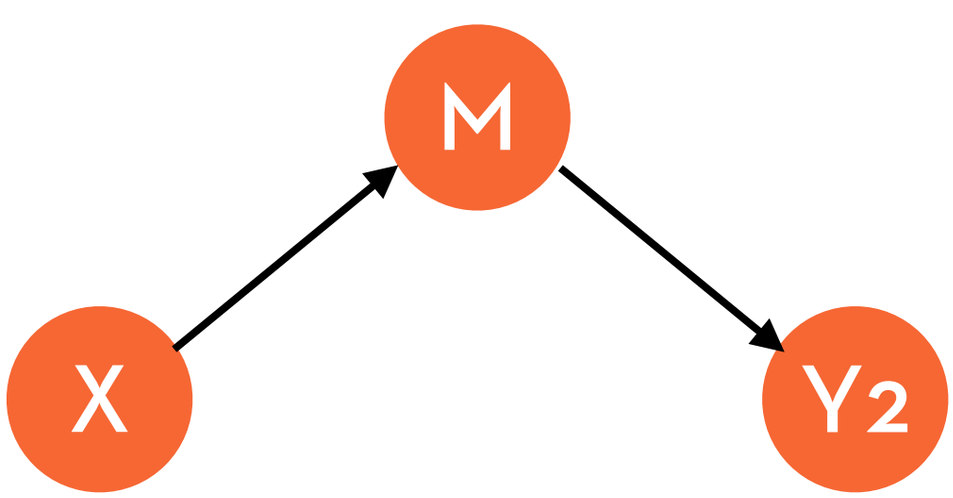
Mediation

What about mediation?

Multiple hypotheses!

Compared to text, comic-based policies (X) increase privacy knowledge (M)

Privacy knowledge (M) is positively associated with perceived privacy protection (Y2)



Example

Research question:

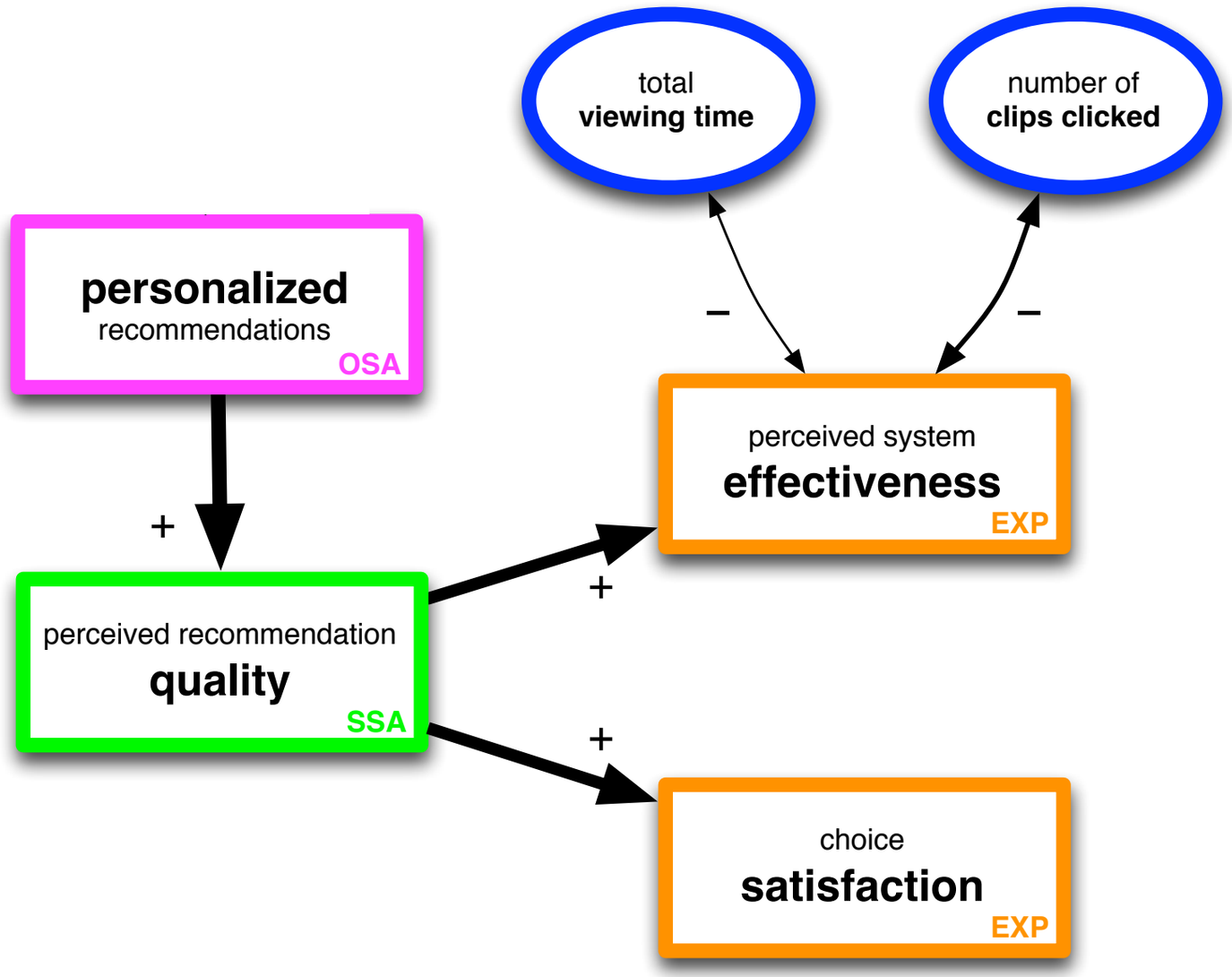
How come personalized video clip recommendations **reduce** users' overall viewing behavior (compared to random clips)?

Example

Hypotheses:

- Users perceive personalized recommendations to have a higher quality than random recommendations
- Recommendation quality is positively associated with perceived system effectiveness
- Recommendation quality is positively associated with choice satisfaction
- Perceived system effectiveness is **negatively** related to users' total viewing time
- Perceived system effectiveness is **negatively** related to the number of clips people watch

Example


total
viewing time
number of
clips clicked
personalized
recommendations
OSA
perceived system
effectiveness
EXP
perceived recommendation
quality
SSA
choice
satisfaction
EXP
+
+
+
–
–

Covariates

What about participant characteristics?

E.g., age, gender, etc.

We are usually interested in them as **covariates**

They change (moderate) the effect of the manipulation

Example

Research question:

> Is there a relationship between domain knowledge and preference elicitation (PE) method in terms of the optimal user experience of a recommender system?

Example

Hypotheses:

Novices have a higher satisfaction when they use the case-based PE method (compared to the attribute-based PE method)

Experts have a higher satisfaction when they use the attribute-based PE method

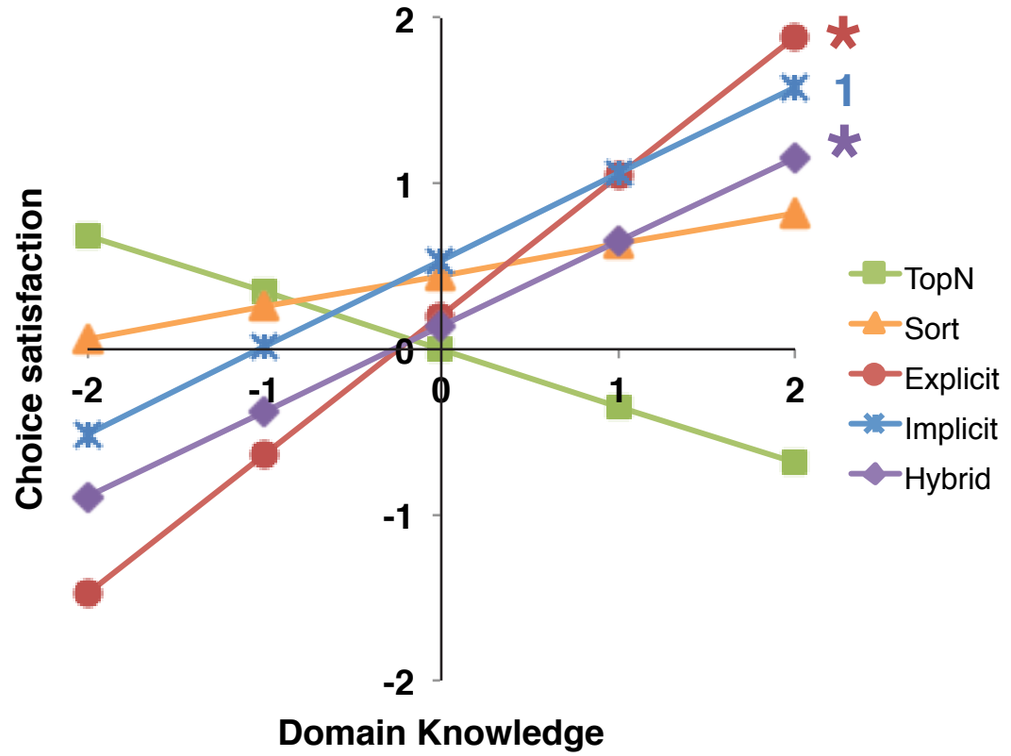
Novices perceive the system with the case-based PE method as more useful

Experts perceive the system with the attribute-based PE method as more useful

Example

The effect of domain knowledge on choice satisfaction is moderated by PE method:

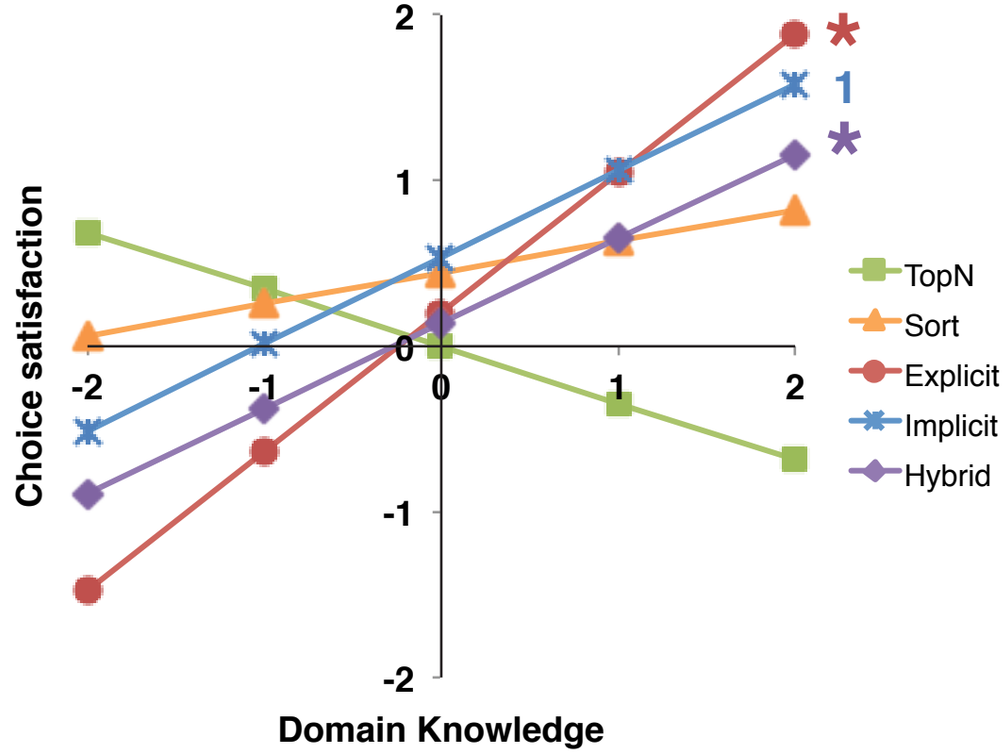
Domain knowledge is negatively associated with choice satisfaction in the TopN condition, but positively associated in the Explicit, Implicit, and Hybrid conditions



Example

The effect of PE method on choice satisfaction is moderated by domain knowledge:

For people with low domain knowledge, TopN performs significantly better than the other conditions; for people with high domain knowledge it performs worse



Interaction effect

Also moderation: Two manipulations at the same time:

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

We call this an interaction effect

Example

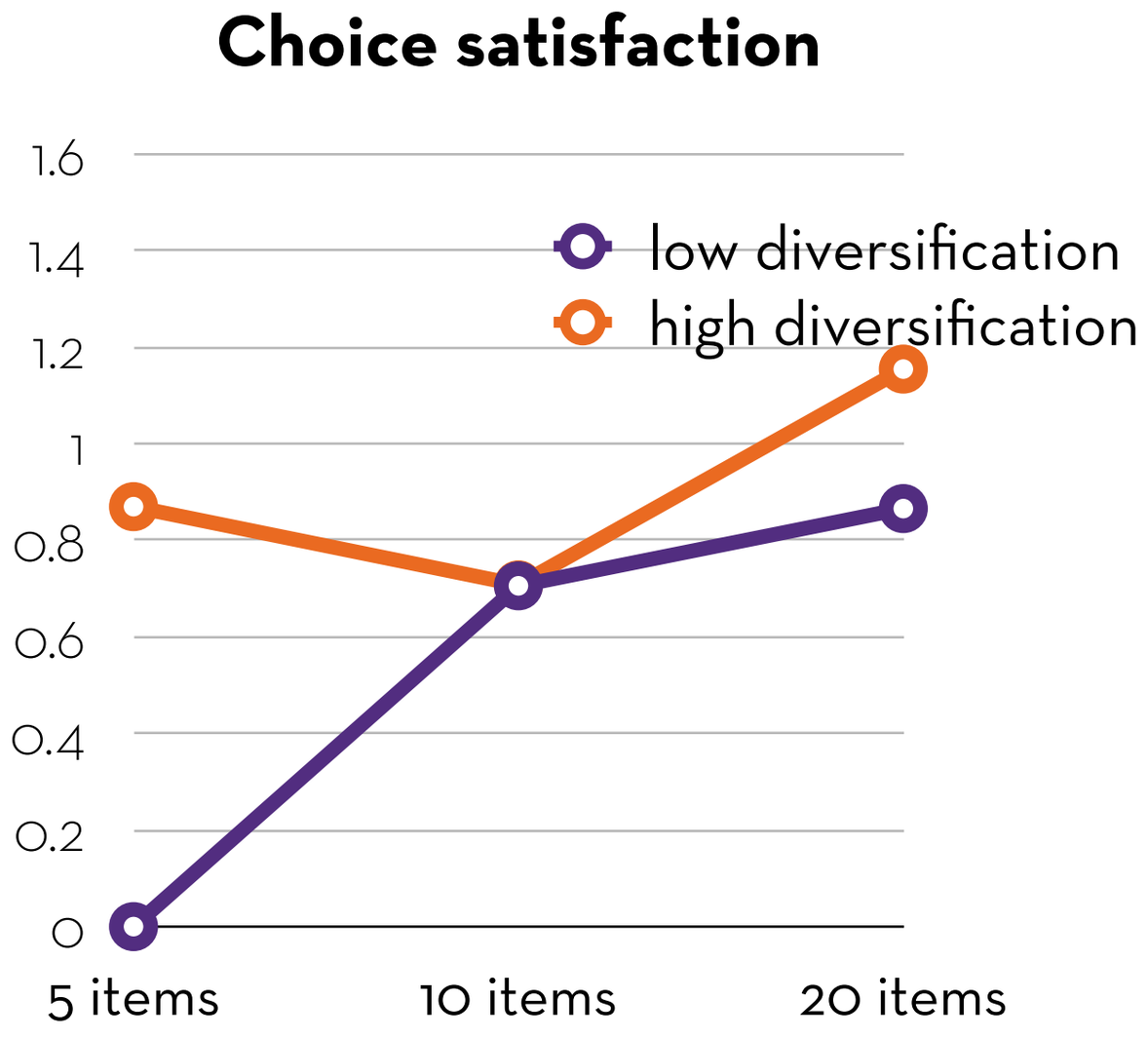
There is an interaction effect between diversification and domain knowledge on choice satisfaction

- High diversification leads to higher choice satisfaction, but only when 5 items are shown

Example

There is an interaction effect between diversification and domain knowledge on choice satisfaction

Or: Choice satisfaction is significantly lower for 5-item lists, but only when diversification is low



Example

Developing hypotheses

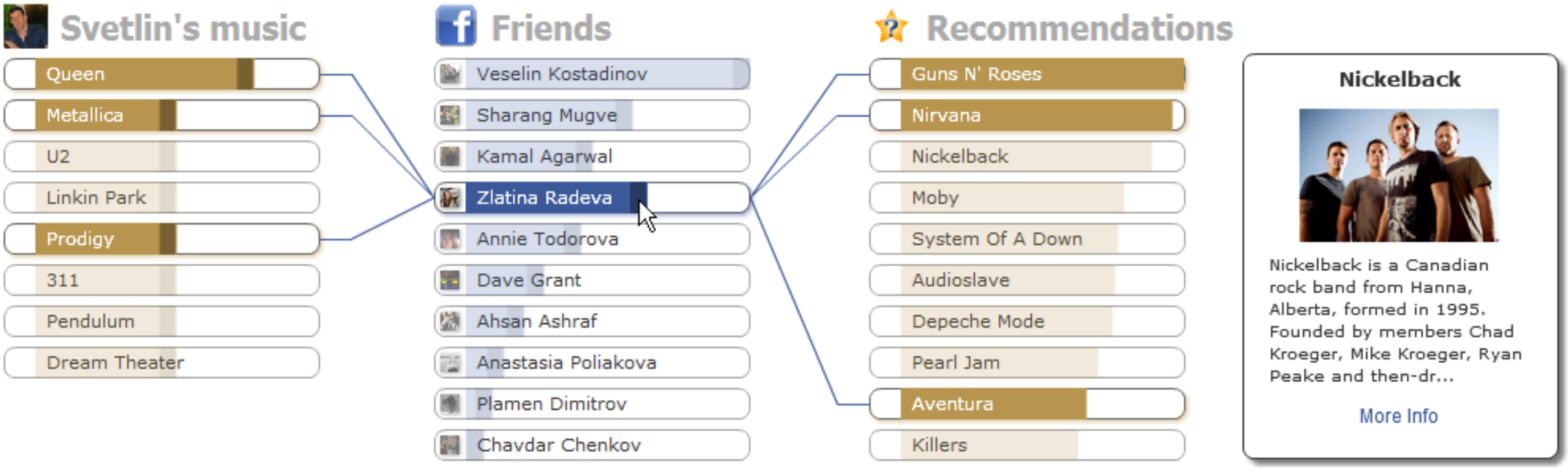
Example

Knijnenburg et al. (2012): "Inspectability and Control in Social Recommenders", *RecSys'12*

The TasteWeights system uses the overlap between you and your friends' Facebook "likes" to give you music recommendations.

- Friends "weights" based on the overlap in likes w/ user
- Friends' other music likes—the ones that are not among the user's likes—are tallied by weight
- Display to the user in a unique graph

Example


Svetlin's music
Queen
Metallica
U2
Linkin Park
Prodigy
311
Pendulum
Dream Theater
Friends
Veselin Kostadinov
Sharang Mugve
Kamal Agarwal
Zlatina Radeva
Annie Todorova
Dave Grant
Ahsan Ashraf
Anastasia Poliakova
Plamen Dimitrov
Chavdar Chenkov
Recommendations
Guns N' Roses
Nirvana
Nickelback
Moby
System Of A Down
Audioslave
Depeche Mode
Pearl Jam
Aventura
Killers
Nickelback
Nickelback is a Canadian rock band from Hanna, Alberta, formed in 1995. Founded by members Chad Kroeger, Mike Kroeger, Ryan Peake and then-dr...
More Info

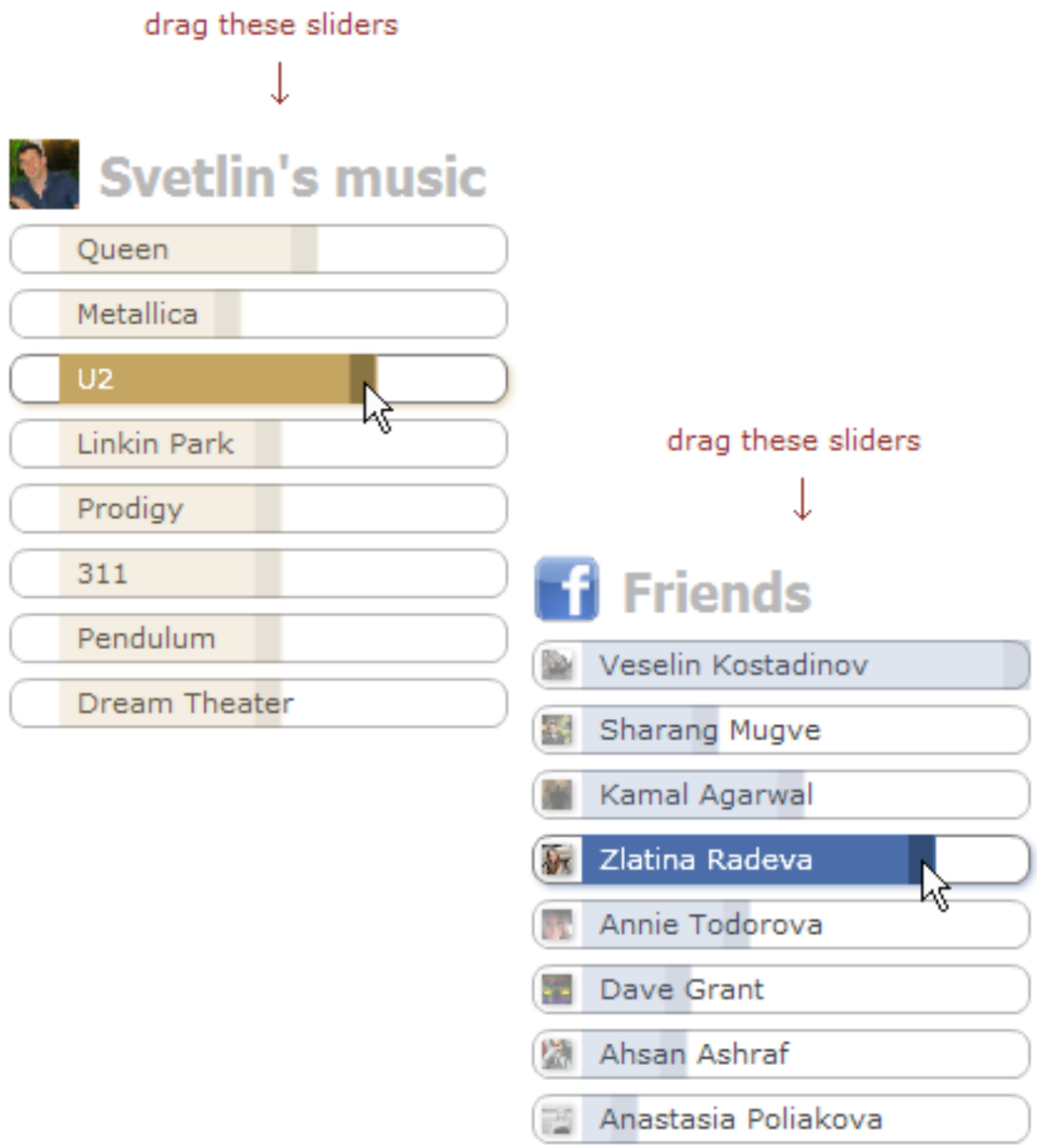
Research question

How do inspectability (the cool graph) and control (the fact that I can set weights) influence the user experience?

Manipulations

3 control conditions:

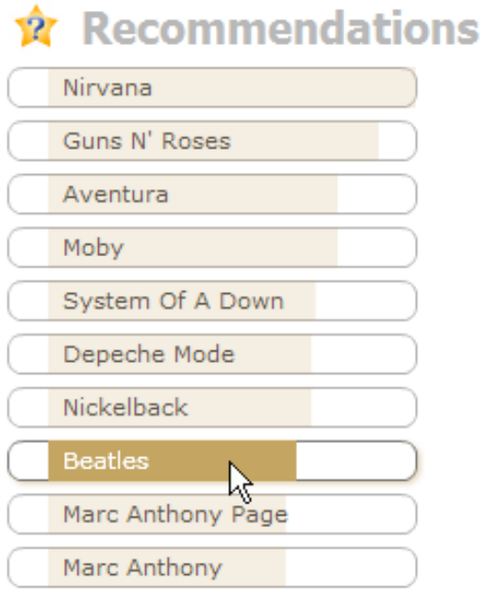
- No control (just use likes)
- Item control (weigh likes)
- Friend control (weigh friends)



Manipulations

2 inspectability conditions:

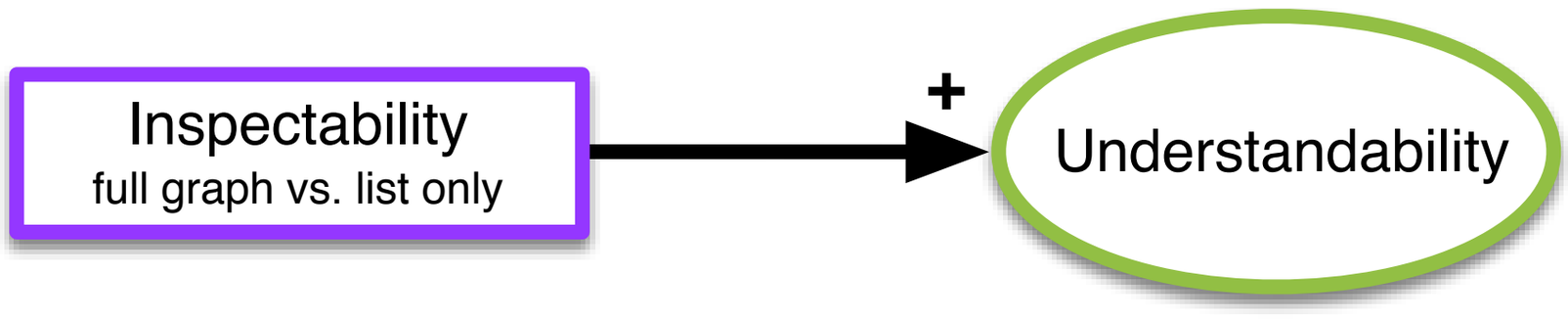
- List of recommendations vs. recommendation graph



Inspectability

Herlocker argues that explanation provides transparency, "exposing the reasoning behind a recommendation"

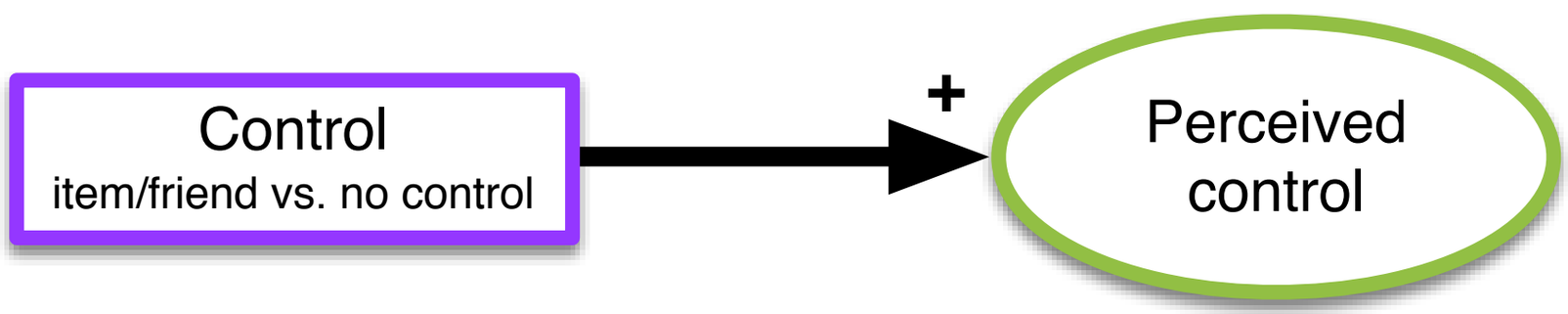
H1: The "full graph" condition results in higher understandability than the "list only" condition



Control

Multiple studies highlight the benefits of interactive interfaces that support control over the recommendation process.

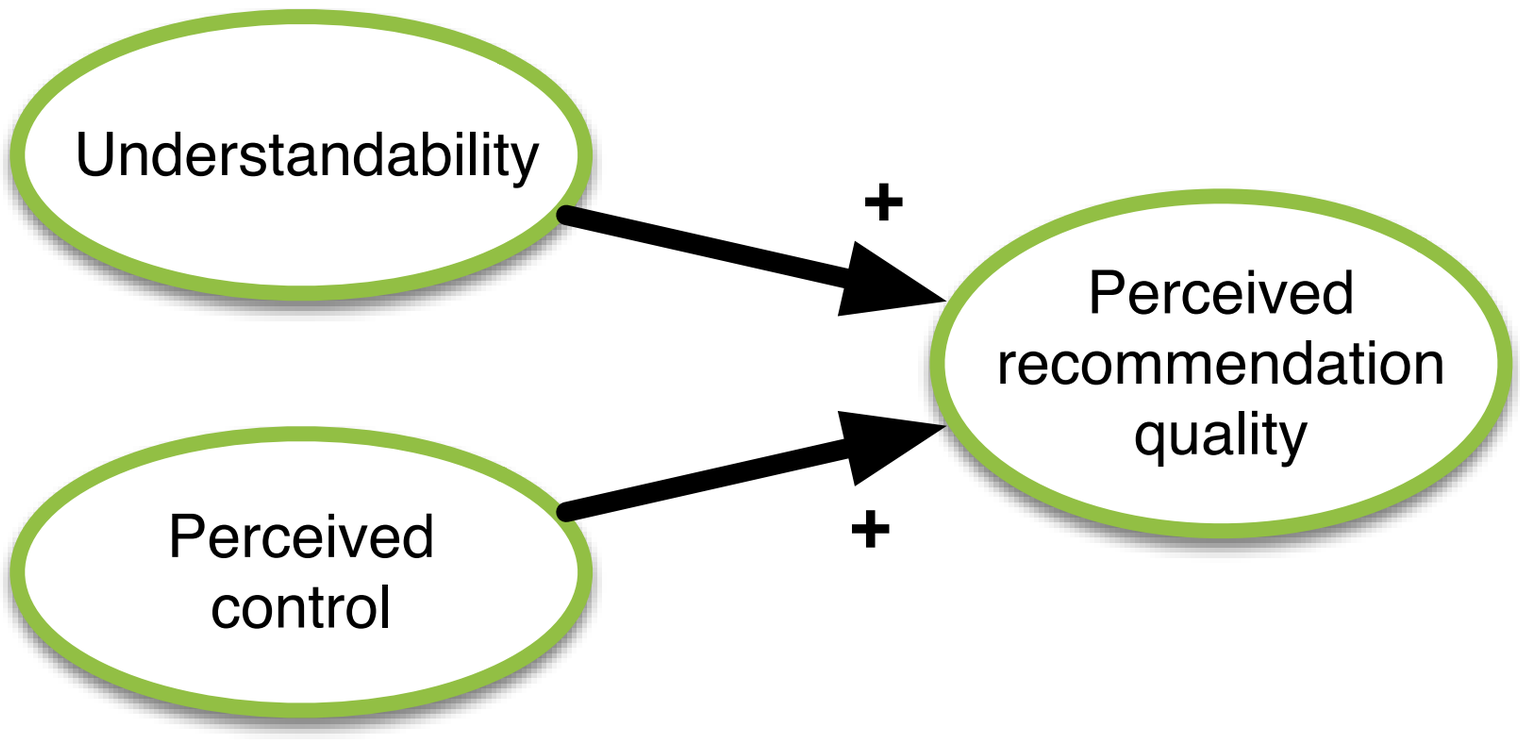
H2: The “item control” and “friend control” conditions lead to a higher level of perceived control than the “no control” condition



Perceived quality

Tintarev and Masthoff show that explanations make it easier to judge the quality of recommendations.

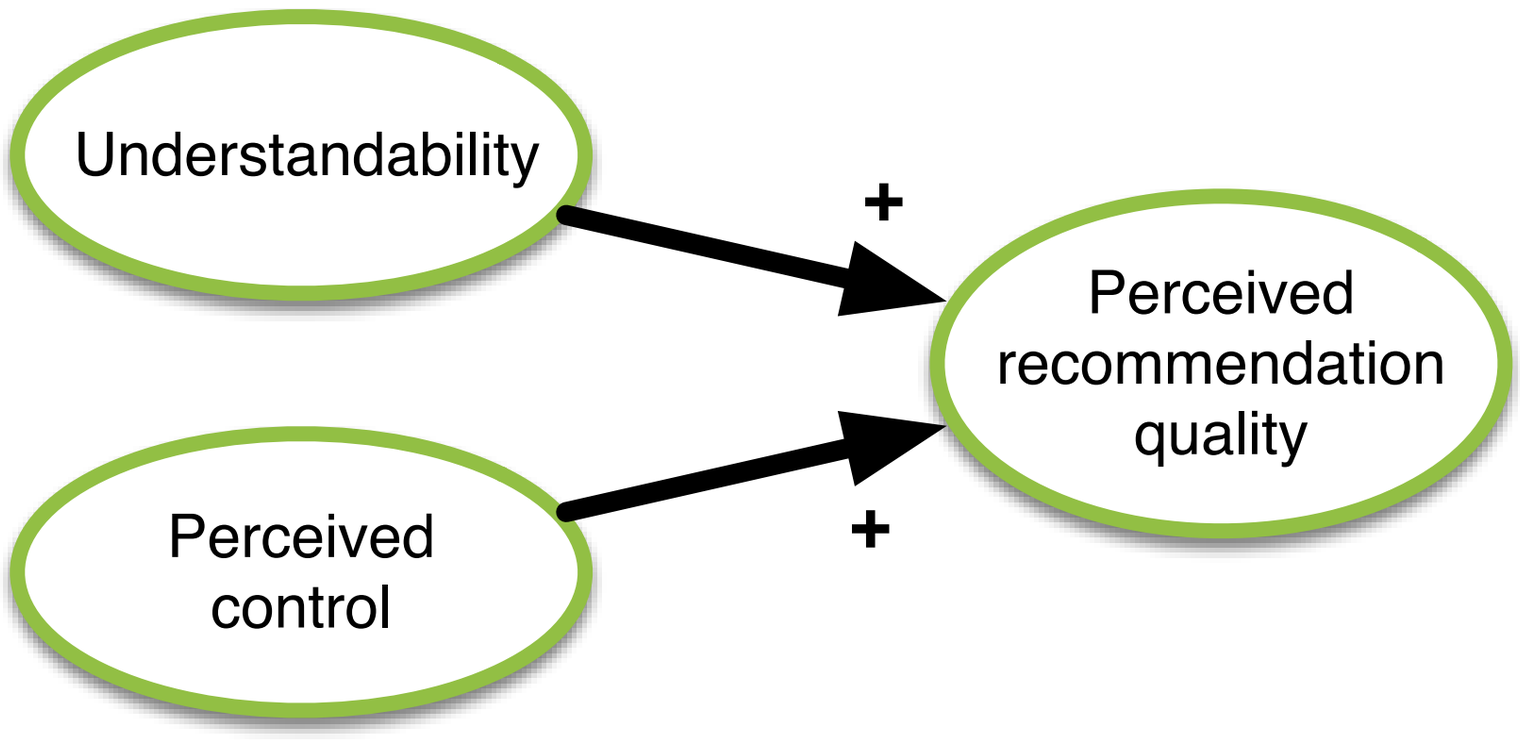
H3: Understandability is positively associated with perceived recommendation quality



Perceived quality

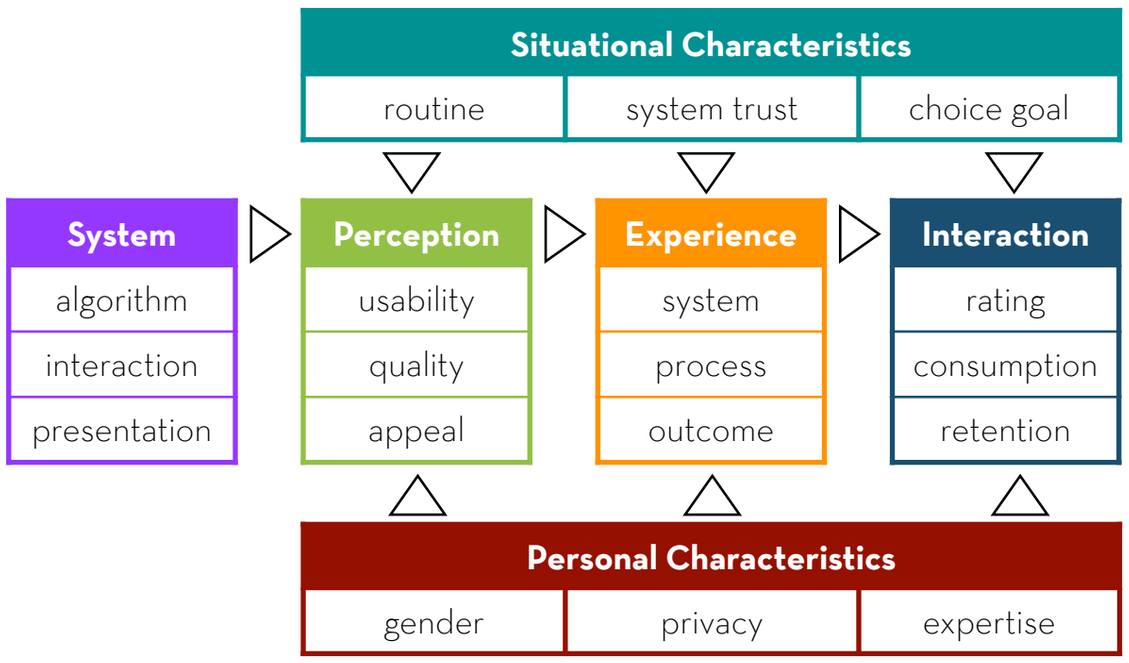
McNee et al. found that study participants preferred user-controlled interfaces because these systems “best understood their tastes”

H4: Perceived control is positively associated with perceived recommendation quality



Satisfaction

Knijnenburg et al. developed a framework that describes how certain manipulations influence subjective system aspects (i.e. understandability, perceived control and recommendation quality), which in turn influence user experience (i.e. system satisfaction).



Satisfaction

Understandability (H5), perceived control (H6) and perceived recommendation quality (H7) are positively associated with system satisfaction