Putting the user in the loop

Martijn Willemsen
M.C.Willemsen@tue.nl

http://www.martijnwillemsen.nl/
@mcwillemsen

Decision Making, Process tracing, Cognition, Recommender Systems, online behavior, e-coaching, Data Science

RecSys Summerschool
Gothenburg
Sept 2019

Recommender LAB @JADS

- PI: Martijn Willemsen, associate professor @JADS / HTI (TU/e)

How can decisions be supported by recommender systems?

The LAB focuses on:

- how insights from decision psychology can improve recommender algorithms
- how to best evaluate recommender systems
- novel recommendation methods that help users with developing their preferences and goals

Domains include movies, music, health-related decisions and recommendation of energy-saving measures.

- http://www.martijnwillemsen.nl/recommenderlab

Recommender systems offer…

Personalized suggestions based on a history of what the user liked and disliked

Main task: predict what items the user would also like…

Algorithmic problem: take a large data set of user data (rating, purchases, clicks, likes) and try to predict the data you don't have

Recommendation task -> predict task

Most popular methods:

Collaborative Filtering (CF) recommenders

- User-based or Item-based CF
- Matrix factorization

This quest for the best algorithm continues…

90% of work in The ACM Conference Series on **Recommender Systems**

But accuracy is not enough…

We need to look at other measures such as **optimize behavior…**

Example: Rows & Beyond

Netflix tradeoffs popularity, diversity and accuracy

AB tests to test ranking between and within rows

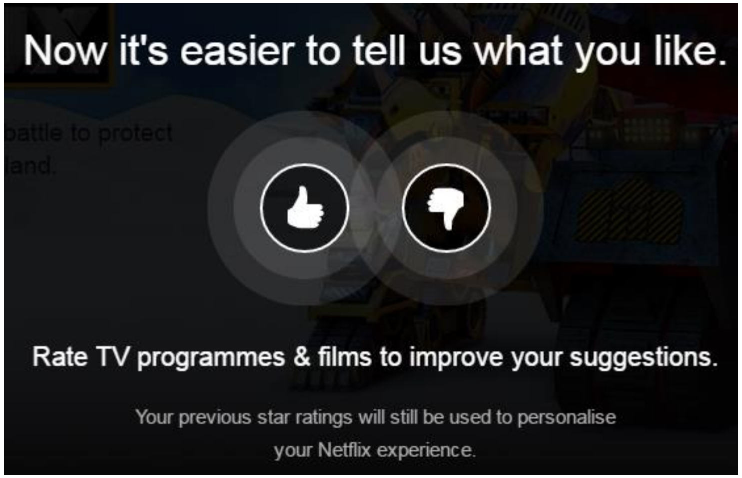
Source: RecSys 2016, 18 Sept: Talk by Xavier Amatriain
http://www.slideshare.net/xamat/past-present-and-future-of-recommender-systems-and-industry-perspective

We don’t need the user: Let’s do AB Testing!

Netflix used 5-star rating scales to get input from users (apart from log data)

Netflix reported an AB test of thumbs up/down versus rating:

Yellin (Netflix VP of product): “The result was that thumbs got 200% more ratings than the traditional star-rating feature.”



We don’t need the user: Let’s do AB Testing!

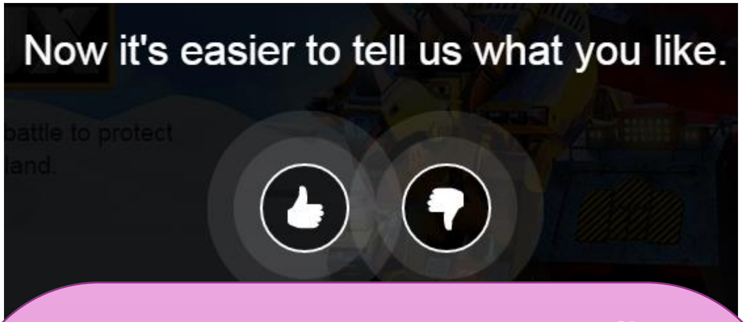
Netflix used 5-star rating scales to get input from users (apart from log data)

Netflix reported an AB test of thumbs up/down versus rating:

Yellin (Netflix VP of product): “The result was that thumbs got 200% more ratings than the traditional star-rating feature.”

So is the 5-star rating wrong? or just different information?

Should we only trust the behavior?



However, over time, Netflix realized that explicit star ratings were less relevant than other signals. Users would rate documentaries with 5 stars, and silly movies with just 3 stars, but still watch silly movies more often than those high-rated documentaries.

http://variety.com/2017/digital/news/netflix-thumbs-vs-stars-1202010492/

Behavior versus Experience

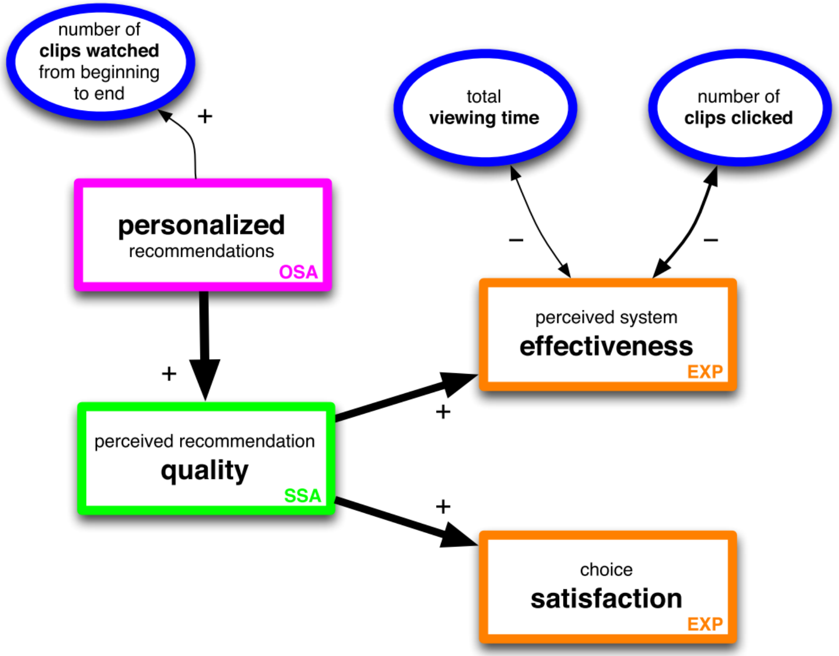
Looking at behavior…

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

Looking at user experience…

- Users found what they liked faster with less ineffective clicks…

Hard to interpret behavior without proper grounding in user experience!



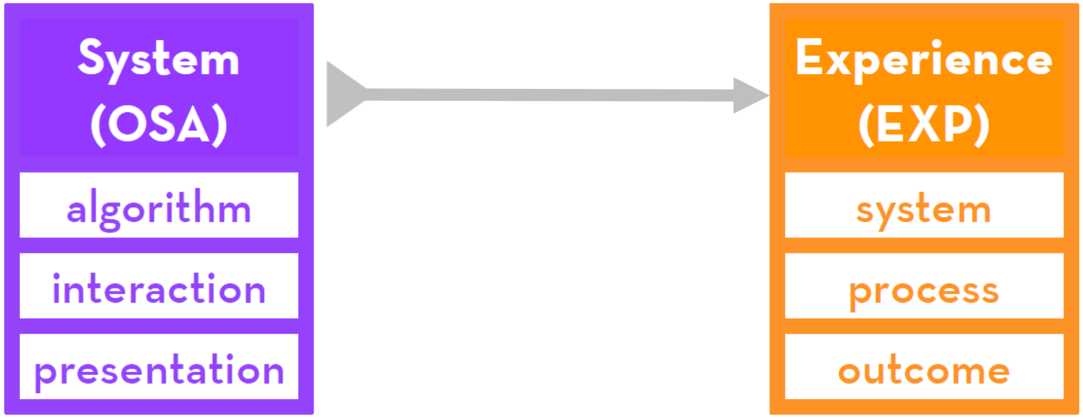
Knijnenburg et al.: "Receiving Recommendations and Providing Feedback", EC-Web 2010

User-Centric Framework

Computers Scientists (and marketing researchers) would study behavior….
(they hate asking the user or just cannot (AB tests))

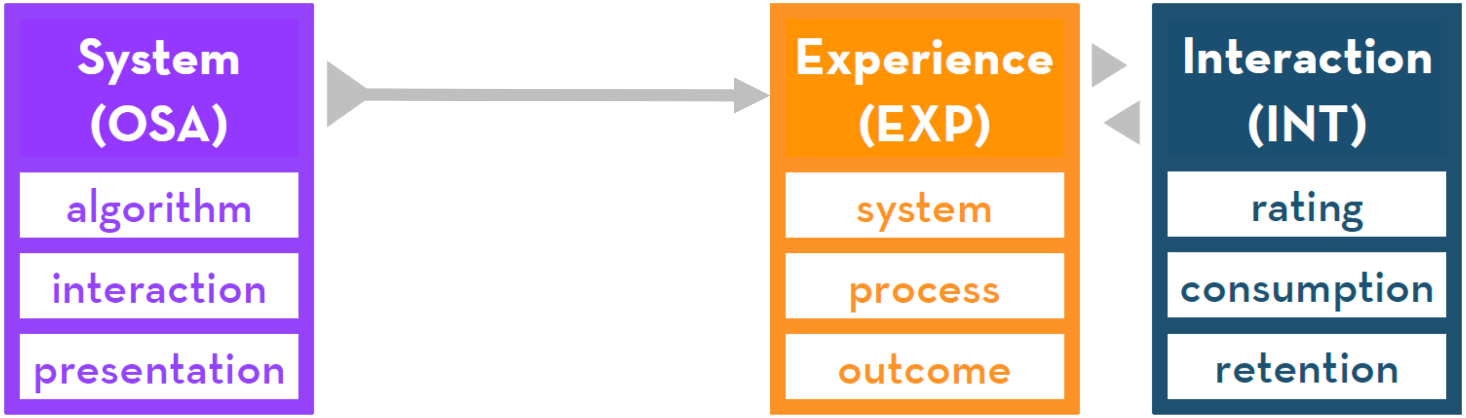
User-Centric Framework

Psychologists and HCI people are mostly interested in experience…



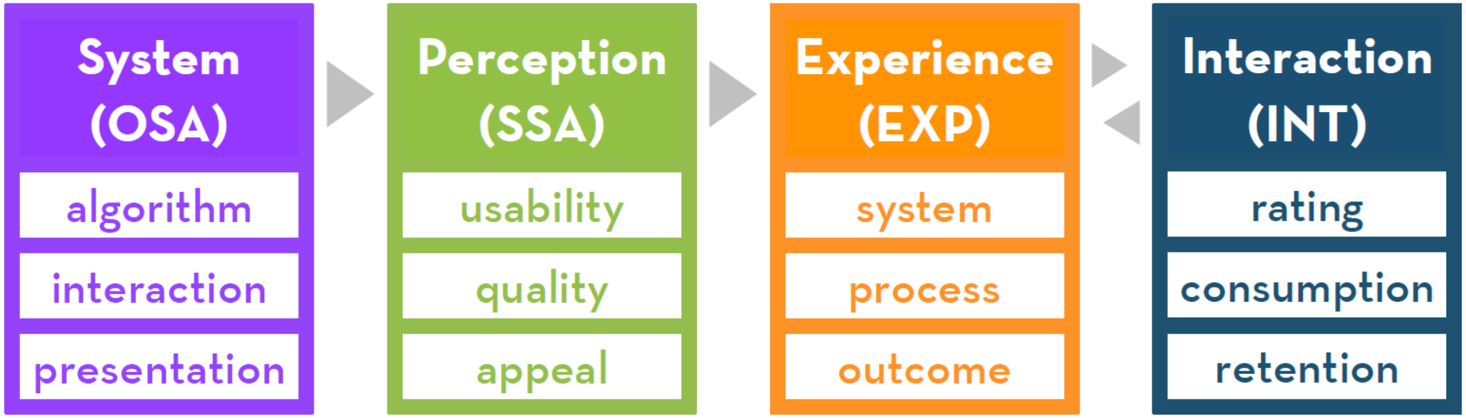
User-Centric Framework

Though it helps to triangulate experience and behavior…



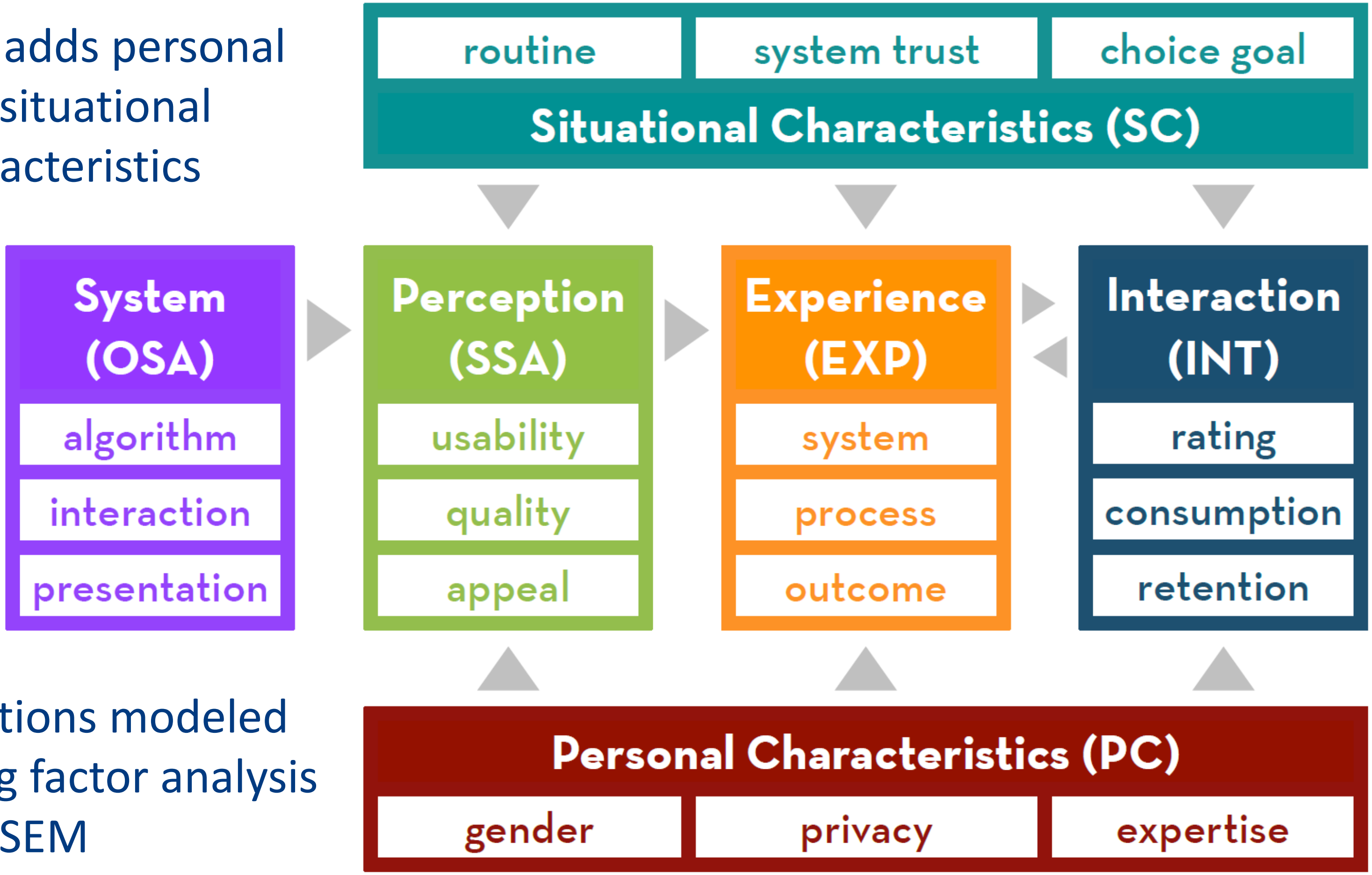
User-Centric Framework

Our framework adds the intermediate construct of perception that explains why behavior and experiences changes due to our manipulations



User-Centric Framework

- And adds personal and situational characteristics



- Relations modeled using factor analysis and SEM

Knijnenburg, B.P., Willemsen, M.C., Gantner, Z., Soncu, H., Newell, C. (2012). Explaining the User Experience of Recommender Systems. *User Modeling and User-Adapted Interaction (UMUAI), vol 22, p. 441-504* *http://bit.ly/umuai*

What should we optimize for?

Objective metrics	Behavior	User Experience
• Historical data (i.e. ratings) • Accuracy, precision/recall • Offline evaluation	• Implicit data • Clickstreams purchases etc. • Online evaluation using AB tests or Bandits	• Explicit data • Subjective perceptions and experiences • Online evaluation using surveys / user experiments

Ex 1:Optimize predict. models using behavior or surveys?

Ex 2: Link objective and subjective measures

Ex 3: Accuracy ≠ satisfaction

comparing objective & subjective measures

- **Ex 1: Online adaptation on hardware.info**
  - Adapting the website to a user segment
  - Predict based on behavior or on survey data?
  - Graus, Willemsen and Swelsen, UMAP 2015
- **Ex 2: Linking objective measures with subjective perceptions**
  - User perceptions of recommender algorithms
  - Ekstrand et al., RecSys 2014
- **Ex 3: Beyond accuracy: increasing diversity and reducing choice difficulty while increasing satisfaction!**
  - Choice difficulty and latent feature diversification
  - Willemsen et al., UMUAI 2016

Online Adaption
behavior versus survey data

Case study based on web log and survey data on Hardware.info

Graus, Willemsen And Swelsen

Graus, M. P., Willemsen, M. C., & Swelsen, K. (2015). Understanding Real-Life Website Adaptations by Investigating the Relations Between User Behavior and User Experience. In F. Ricci, K. Bontcheva, O. Conlan, & S. Lawless (Eds.),*User Modeling, Adaptation and Personalization* (pp. 350–356). Springer International Publishing. Link to springer

Hardware.info

- Aimed at IT/CE-enthusiasts
- Second Biggest IT website in the Netherlands: 8+ mln pageviews/month
- Editorial board, reviews (1500 per year), active community

- hardware components (HC)
  End User Products (EUP)
- Question: can we adapt the sidebar to user interest (HC or EUP)

Log data of the web server

28.177.271 (page) requests
(1 month of data)

-

```
GET /dumprequest HTTP/1.1
Host: djce.org.uk
User-Agent: Mozilla/5.0 (Windows NT 6.1; WOW64; rv:25.0) Gecko/20100101 Firefox/25.0
Accept: text/html,application/xhtml+xml,application/xml;q=0.9,*/*;q=0.8
Accept-Language: en-US,en;q=0.5
Connection: keep-alive
```

```
GET /dumprequest HTTP/1.1
Host: djce.org.uk
User-Agent: Mozilla/5.0 (Win
Accept: text/html,applicatic
Accept-Language: en-US,en;q=
Connection: keep-alive
```

```
GET /dumprequest HTTP/1.1
Host: djce.org.uk
User-Agent: Mozilla/5.0 (Windows NT 6.1; WOW64; rv:25.0) Gecko/20100101 Firefox/25.0
Accept: text/html,application/xhtml+xml,application/xml;q=0.9,*/*;q=0.8
Accept-Language: en-US,en;q=0.5
Connection: keep-alive
```

11.486.008 (40.8%) irrelevant requests
(advertisements, RSS, graphs) dropped

Session

456.233 users

Page 1

Page 2

.....

Page N

1.631.615 sessions

Link categories to product groups

116 different product groups on the website: (processors, main boards, SSDs, but also TVs, phones, game consoles and tablets)

8.818.528 requests for different categories on the website could be linked to a product group

59 product groups (4.148.089 requests) flagged as **HC**

- E.g., OS software, processors, graphic cards and harddrives

57 product groups (3.267.074 requests) flagged as **EUP**

- TVs, tablets, game consoles and laptops

#requests	Category	Percentage
11 486 008	Irrelevant/dropped	40.8%
3 057 930	Product info	10.9%
2 215 879	Newsletter	7.86%
2 189 151	Reviews	7.77%
1 661 115	News	5.90%
1 397 133	Main page	4.96%
1 291 870	Updates	4.58%
1 021 924	Forum	3.63%
730 427	Product group	2.59%

Use a predictive model to alter the website

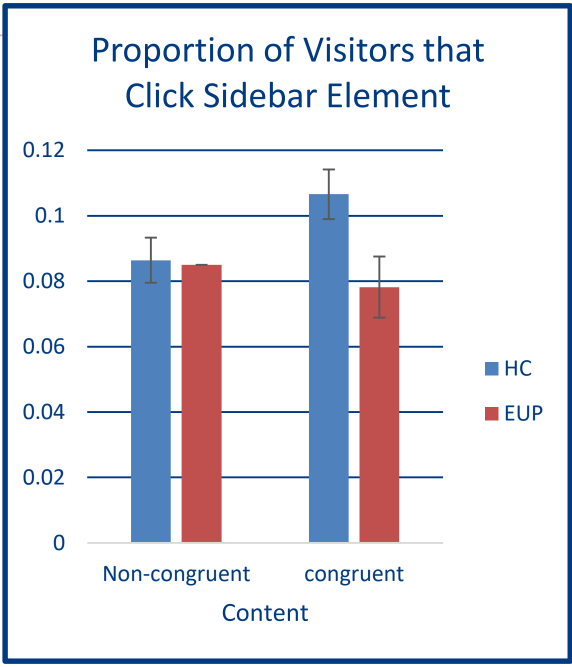
- Classify people based on their previous pageviews:
  Can we predict early on (after 5 pages) what type of user and adapt the side bar to that user?
- During 2 weeks on hardware.info we ran an online experiment

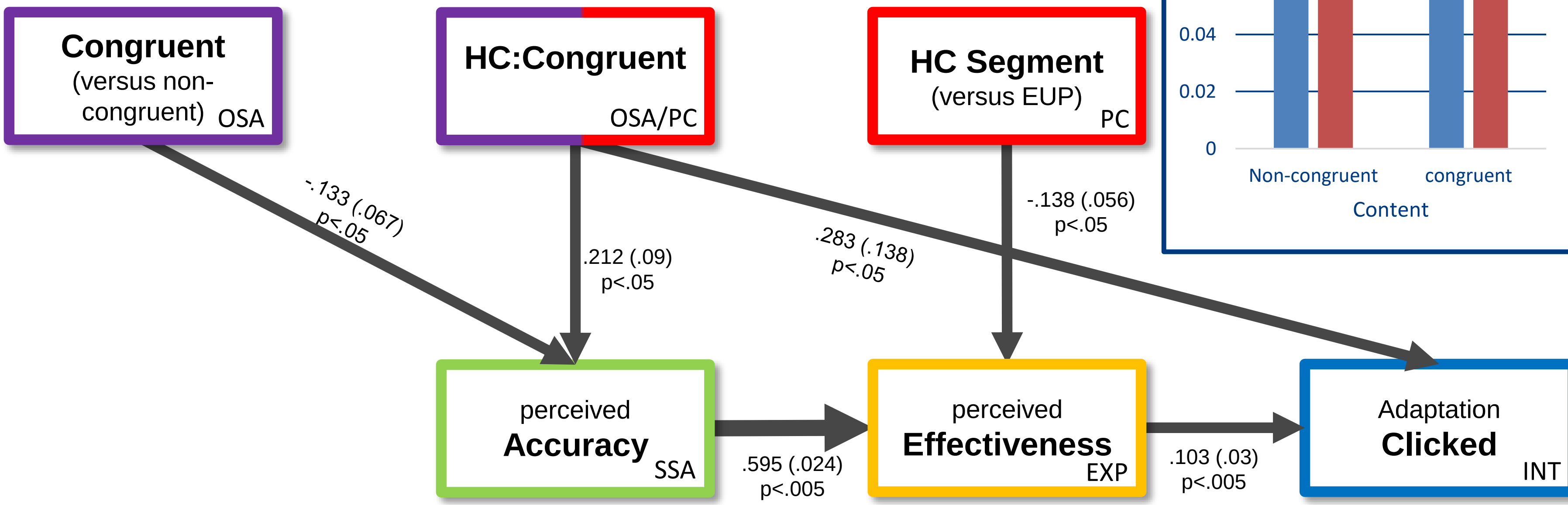
- Collected 2 weeks worth of data
  - 100k unique visitors
  - 3k completed surveys

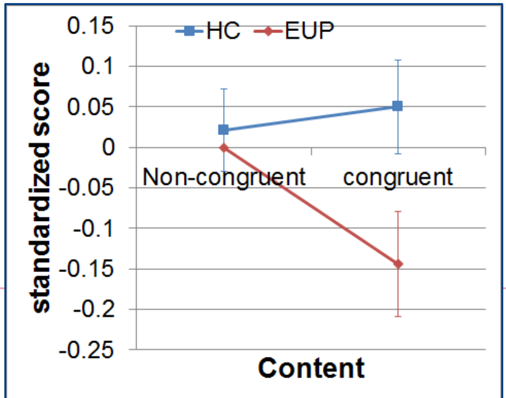
Predicted segment

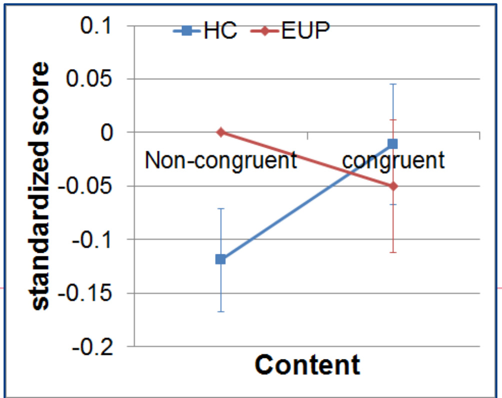
Randomly assigned sidebar

We combined survey and behavioral data…
In a path model





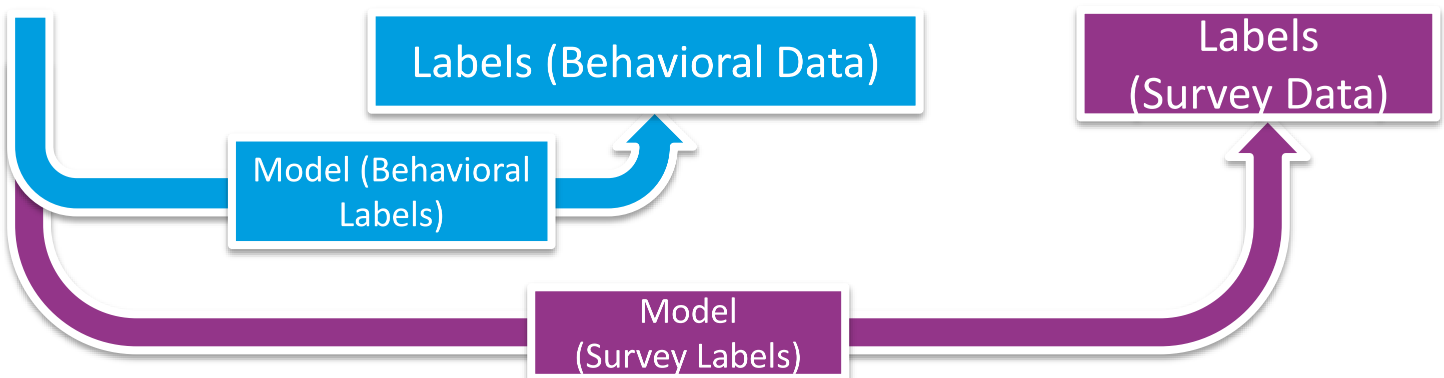




Post-hoc analysis:

Can we build a better predict model if we use the survey data (3K) than the behavioral data (100k)

Data		
5 pageviews	Rest of Visit **Content:** EUP, HC, Mix	Survey Data *To what extent are you interested in HC/EUP?*



- How well do the different models predict behavior?

Model (Survey Data) | Model (Behavioral Data)

Predictions | Predictions

Shown Content	Predicted Segment	Predicted Segment		Rest of Visit
HC	HC	HC	Explain	• Clicks on Sidebar Element
EUP	EUP	EUP		• Pageviews
Mix				• Sessions

- Predict future actions after 5 clicks based on behavioral or survey model using multinomial logistic regression

AIC				
Labels	Clicks on Sidebar	Clicks on Sidebar (Boolean)	Pageviews	Sessions
Behavior	834,821.3	26,910.6	23,362.0	517,453.3
Survey	832,555.5	26,832.5	23,270.2	514,761.0

- Survey-based model provides better predictions for response to the Sidebar Element than models based on Behavioral Data
- Despite less information (3k vs 100k)
- We are predicting segments for 100.000 visitors while using data from only 3,000!

User Perceptions of Differences in Recommender Algorithms

Joint work with grouplens

Michael Ekstrand, Max Harper and Joseph Konstan

Ekstrand, M.D., Harper, F.M., Willemsen, M.C.& Konstan, J.A. (2014). User Perception of Differences in Recommender Algorithms. In *Proceedings of the 8th ACM conference on Recommender systems* (pp. 161–168). New York, NY, USA: ACM

Technische Universiteit Eindhoven University of Technology

Going beyond accuracy…

McNee et al. (2006): Accuracy is not enough

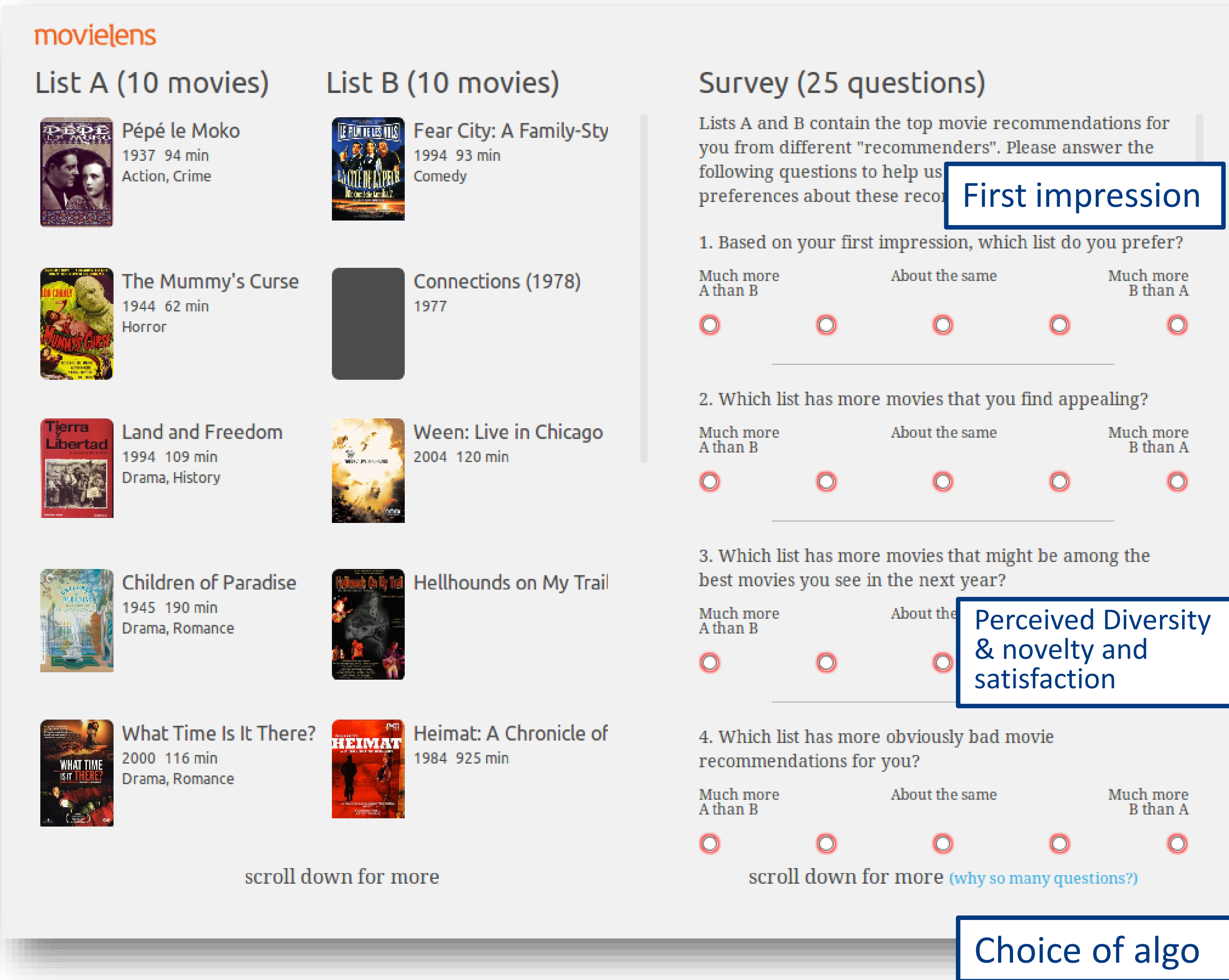
“study recommenders from a user-centric perspective to make them not only accurate and helpful, but also a pleasure to use”

But wait!

we don’t even know how the standard algorithms are perceived… and what differences there are…

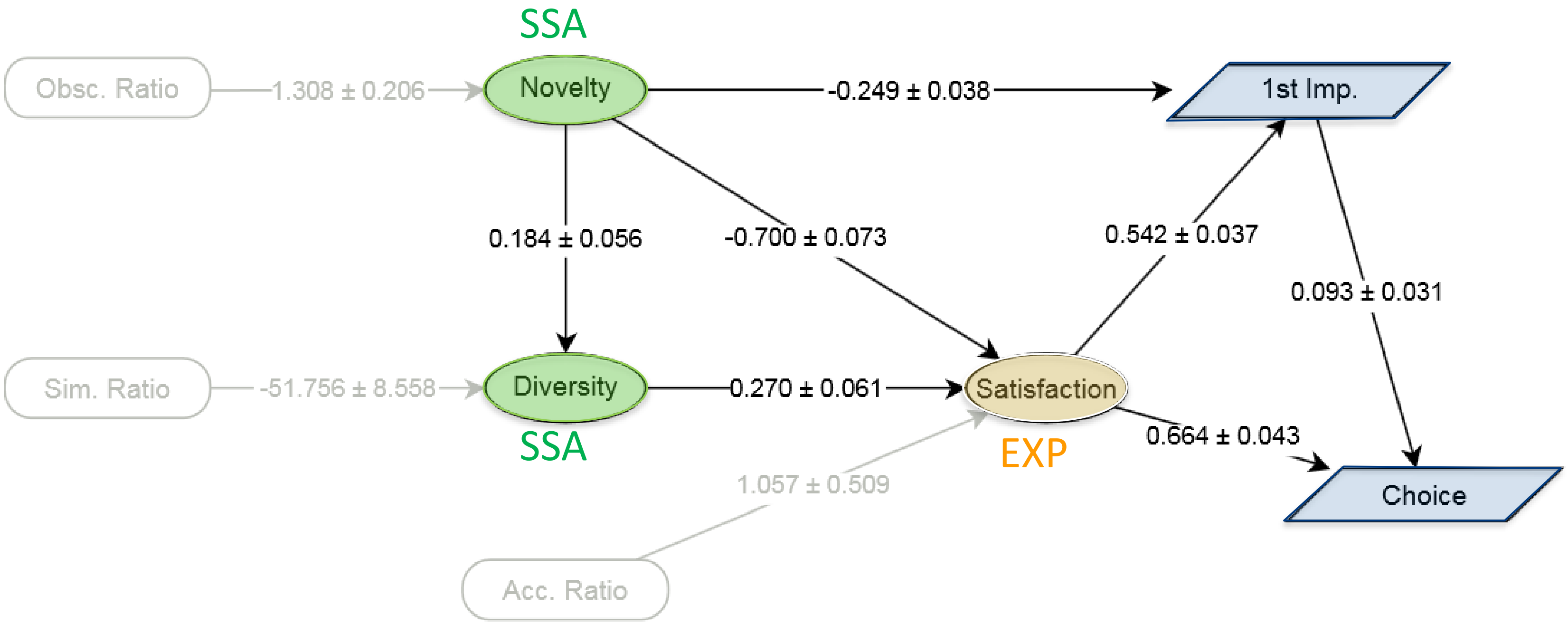
Compare 3 classic algorithms (Item-Item, User-User and SVD) side by side (joint evaluation) in terms of preference and perceptions

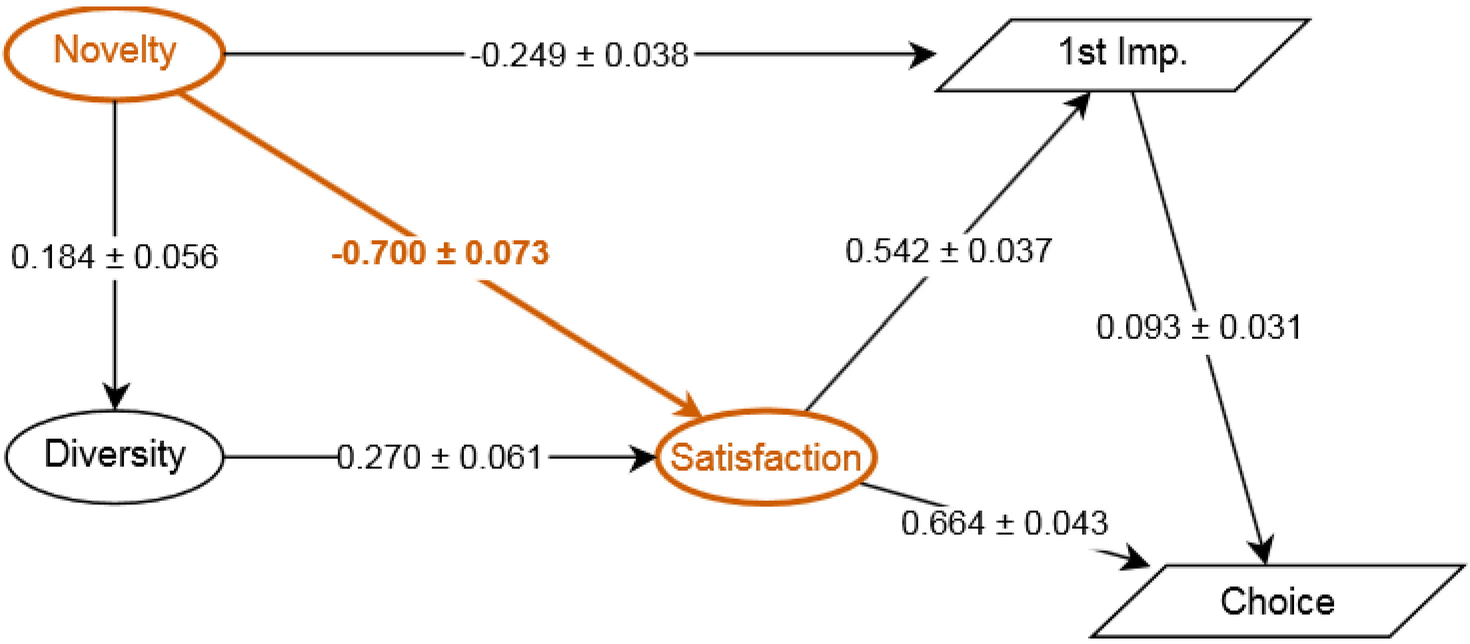
The task provided to the user



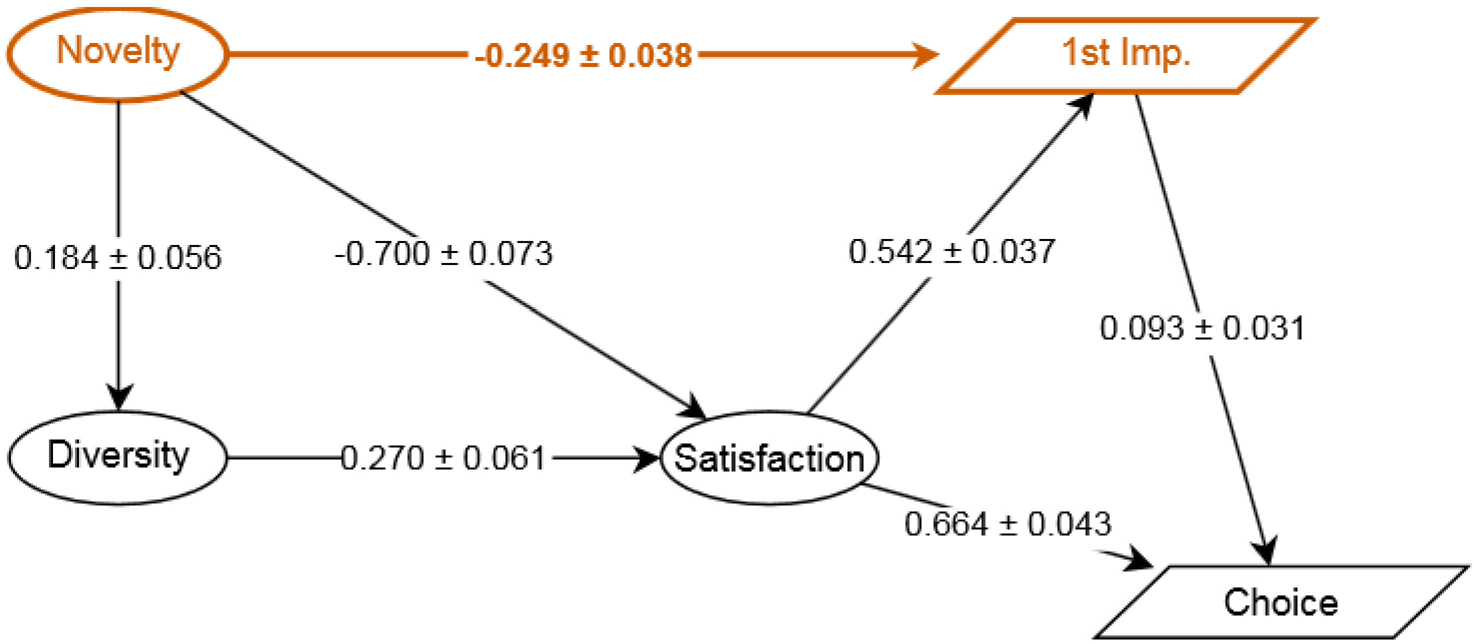
First look at the measurement model

- only measurement model relating the concepts (no conditions)
- All concepts are relative comparisons
  - e.g. if they think list A is more diverse than B, they are also more satisfied with list A than B

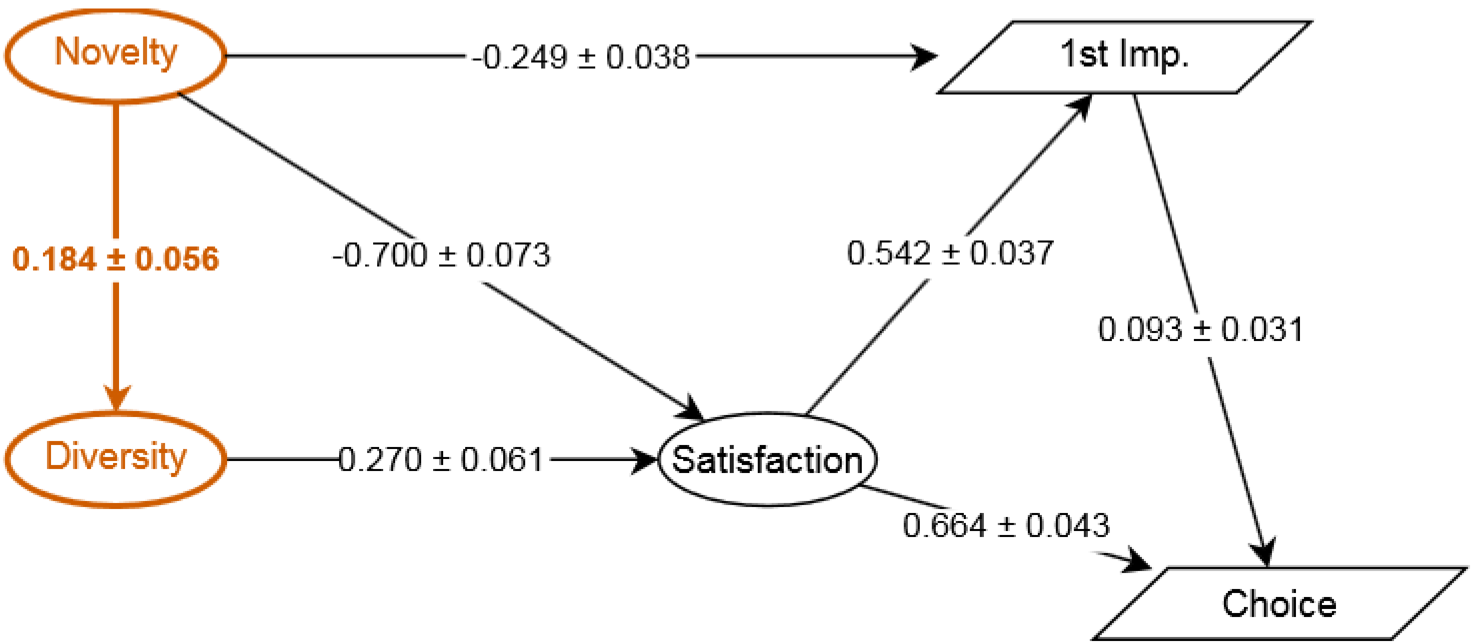




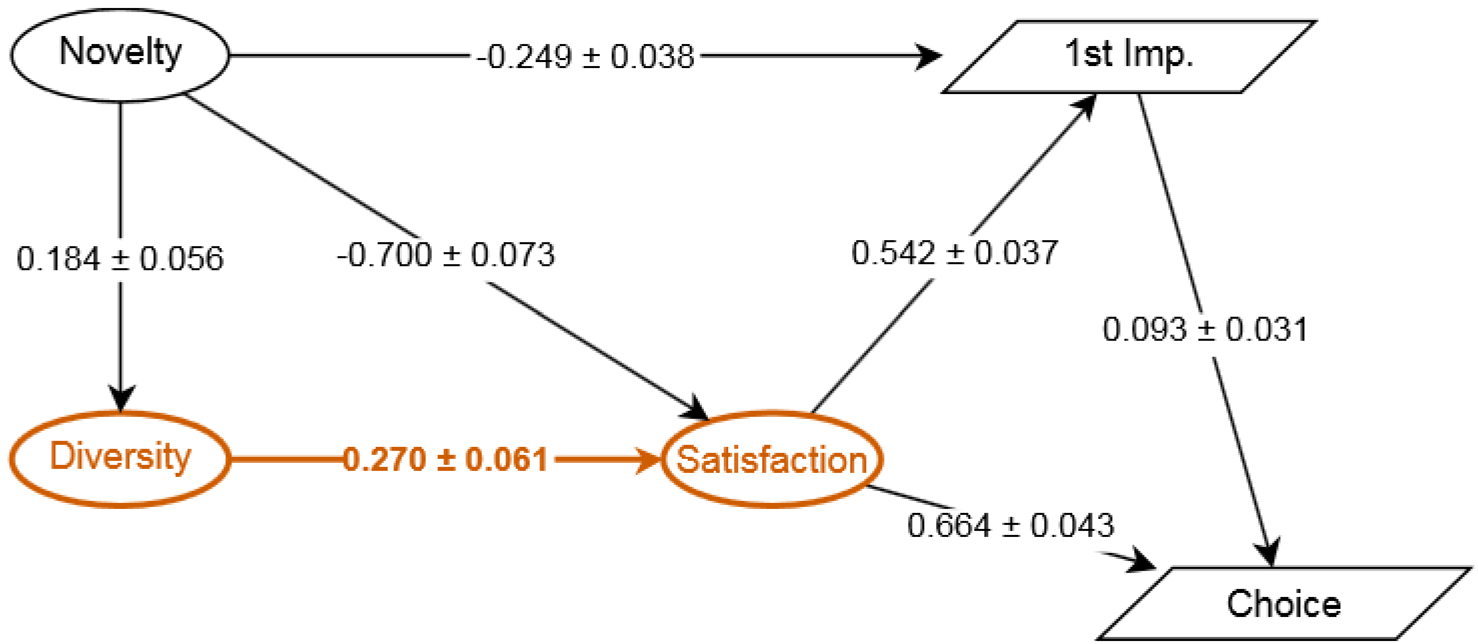
Novelty hurts satisfaction



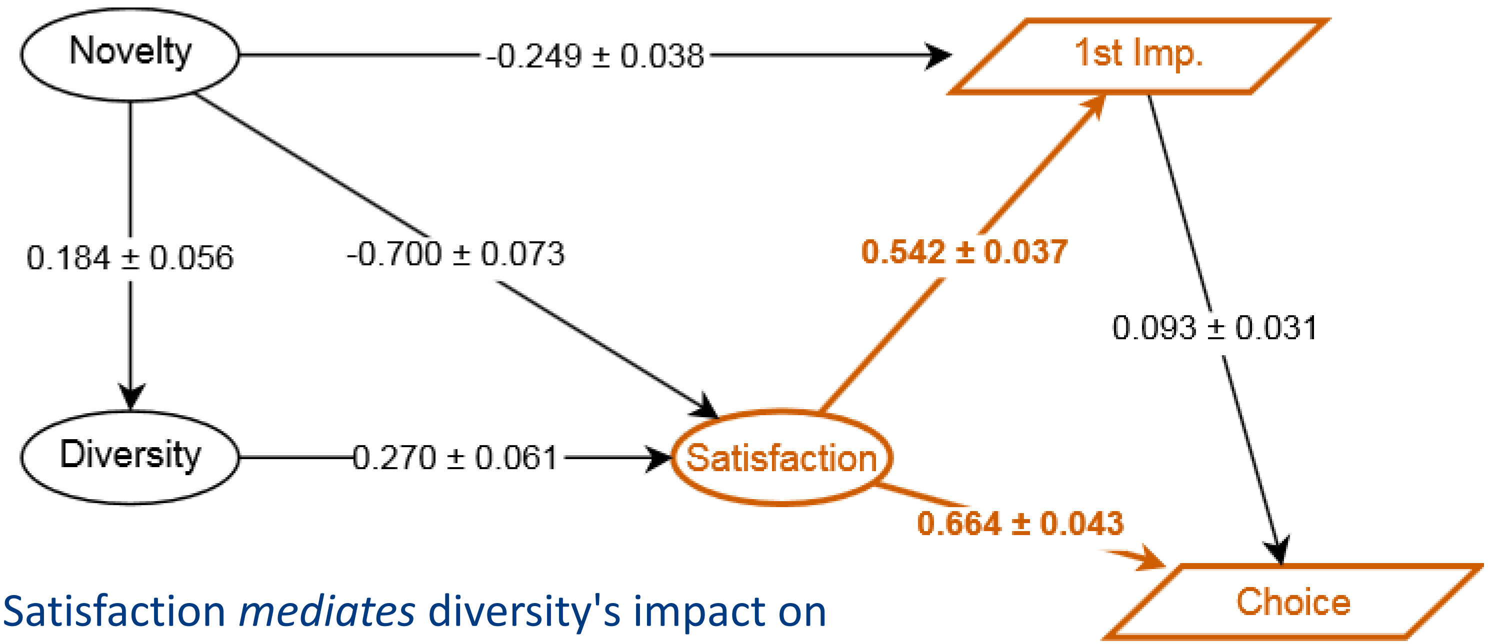
Novelty has direct negative impact on first impression.



Novelty improves diversity (slightly).
outweighed by negative satisfaction effect



Diversity positively influences satisfaction.



Satisfaction *mediates* diversity's impact on preference

No direct effects left of novelty and diversity on choice!

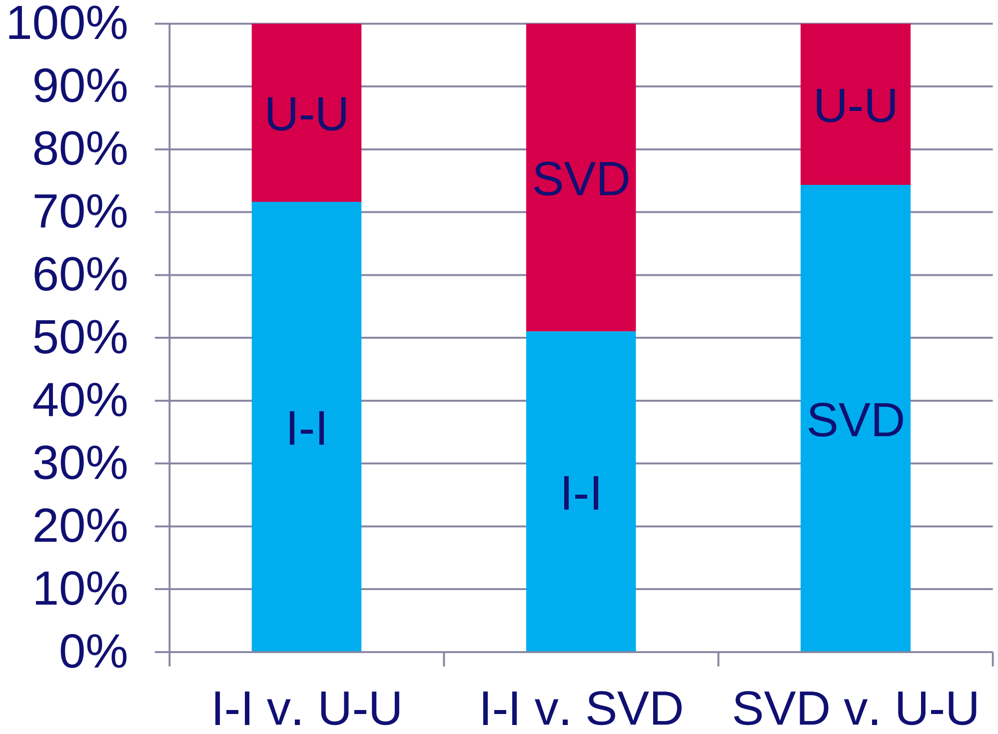
What algorithms do users prefer?

528 users completed the questionnaire

Joint evaluation, 3 pairs of comparing A with B

User-User CF significantly looses from the other two

Item-Item and SVD are on par



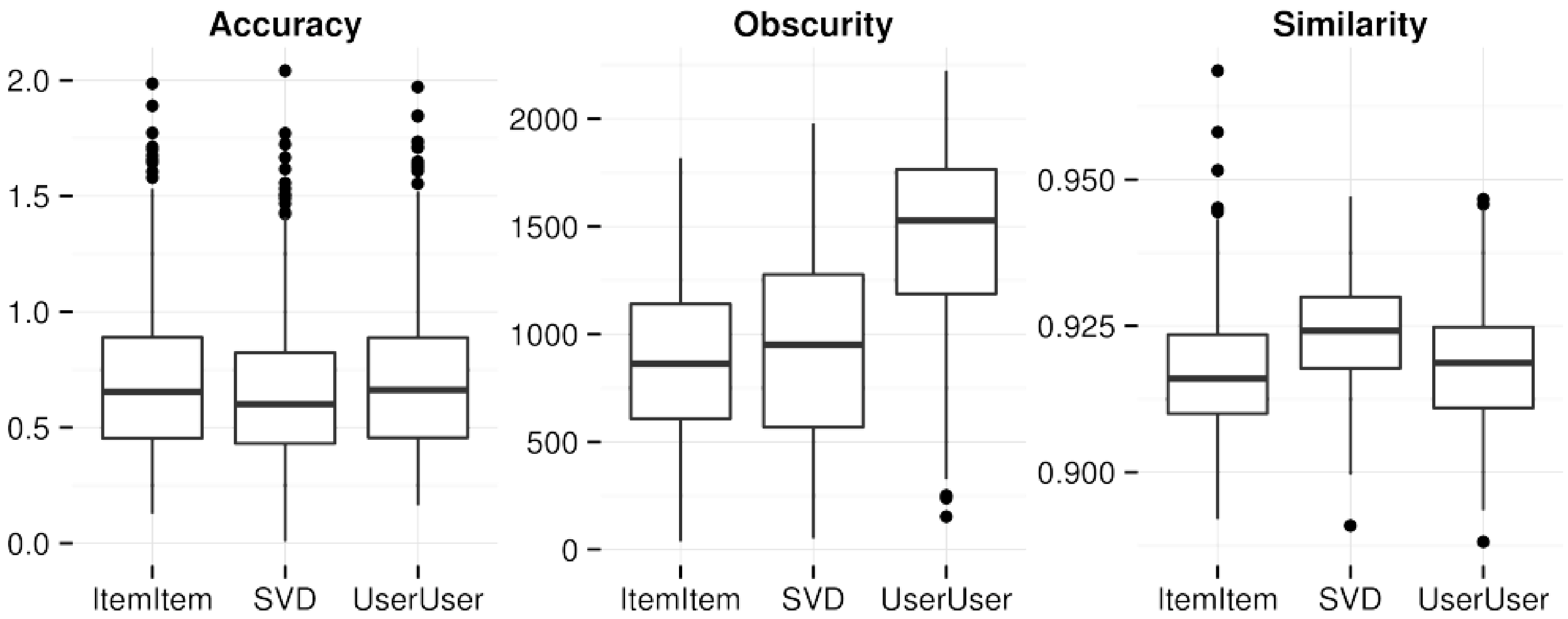
Why?

– User-user more **novel** than either SVD or item-item
– User-user more **diverse** than SVD
– Item-item slightly more diverse than SVD (but diversity didn't affect satisfaction)

Objective measures

No accuracy differences, but consistent with subjective data

RQ2: User-user more novel, SVD somewhat less diverse

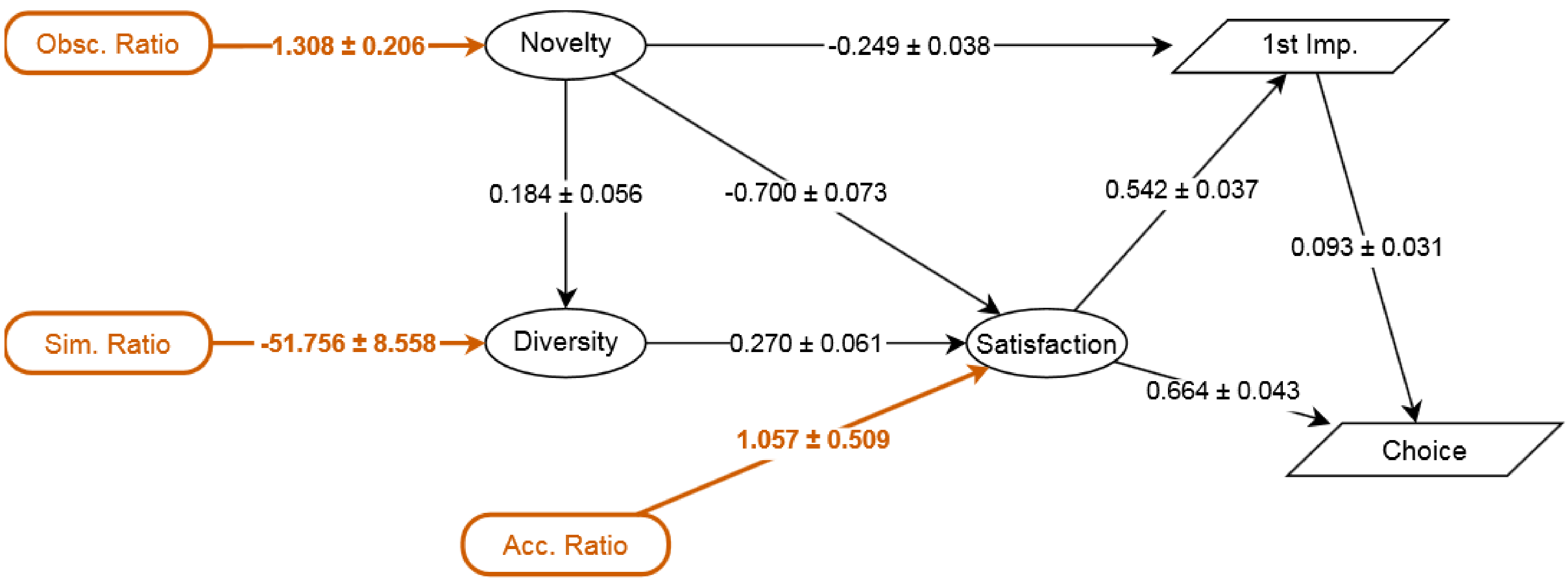


Aligning objective with subjective measures

Objective and subjective metrics correlate consistently

But their effects on choice are mediated by the subjective perceptions!

- (Objective) obscurity only influences satisfaction if it increases perceived novelty (i.e. if it is registered by the user)



Conclusions

Novelty is not always good: complex, largely negative effect

Diversity is important for satisfaction

- **Diversity/accuracy tradeoff** does not seem to hold…

Subjective Perceptions and experience mediate the effect of objective measures on choice / preference for algorithm

- Brings the 'WHY': e.g. User-user is less satisfactory and less often chosen because of its obscure items (which are perceived as novel)

Technische Universiteit
Eindhoven
University of Technology

Choice difficulty and satisfaction in RecSys

Applying latent feature diversification

User Model User-Adap Inter
DOI 10.1007/s11257-016-9178-6

CrossMark

Understanding the role of latent feature diversification on choice difficulty and satisfaction

Martijn C. Willemsen[1] · Mark P. Graus[2] · Bart P. Knijnenburg[3]

Abstract People like variety and often prefer to choose from large item sets. However, large sets can cause a phenomenon called "choice overload": they are more difficult to choose from, and as a result decision makers are less satisfied with their choices. It

Willemsen, M.C., Graus, M.P, & Knijnenburg, B.P. (2016). Understanding the role of latent feature diversification on choice difficulty and satisfaction. *User Modeling and User-Adapted Interaction (UMUAI), vol 26 (4), 347-389* doi:10.1007/s11257-016-9178-6

Seminal example of choice overload

Less attractive

30% sales

Higher purchase satisfaction

From Iyengar and Lepper (2000)

More attractive

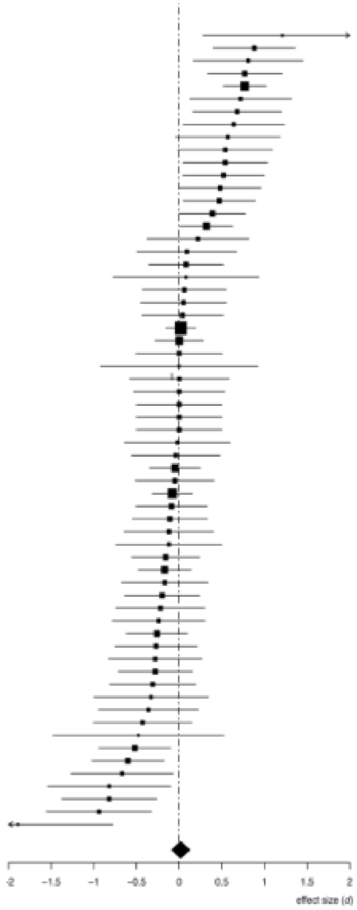
3% sales

Satisfaction decreases with larger sets as increased attractiveness is counteracted by **choice difficulty**

Research on Choice overload

Choice overload is not omnipresent

Meta-analysis (Scheibehenne et al., JCR 2010) suggests an overall effect size of zero



Choice overload stronger when:

No strong prior preferences

Little difference in attractiveness items

Prior studies did not control for the **diversity of the item set**

Can we reduce choice difficulty and overload by using **personalized** diversified item sets?

While controlling for attractiveness…

Latent feature diversification: high diversity/equal attractiveness

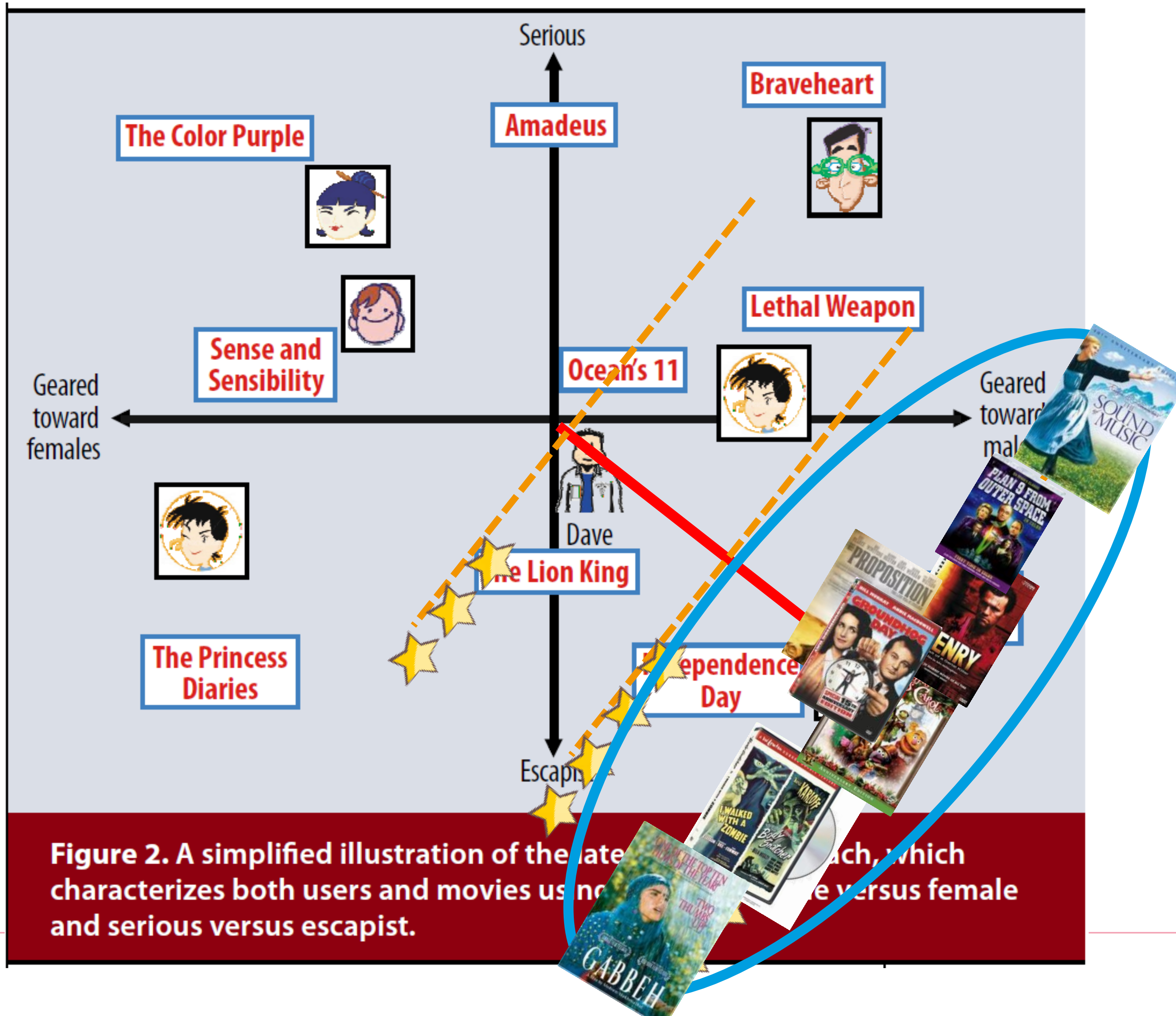
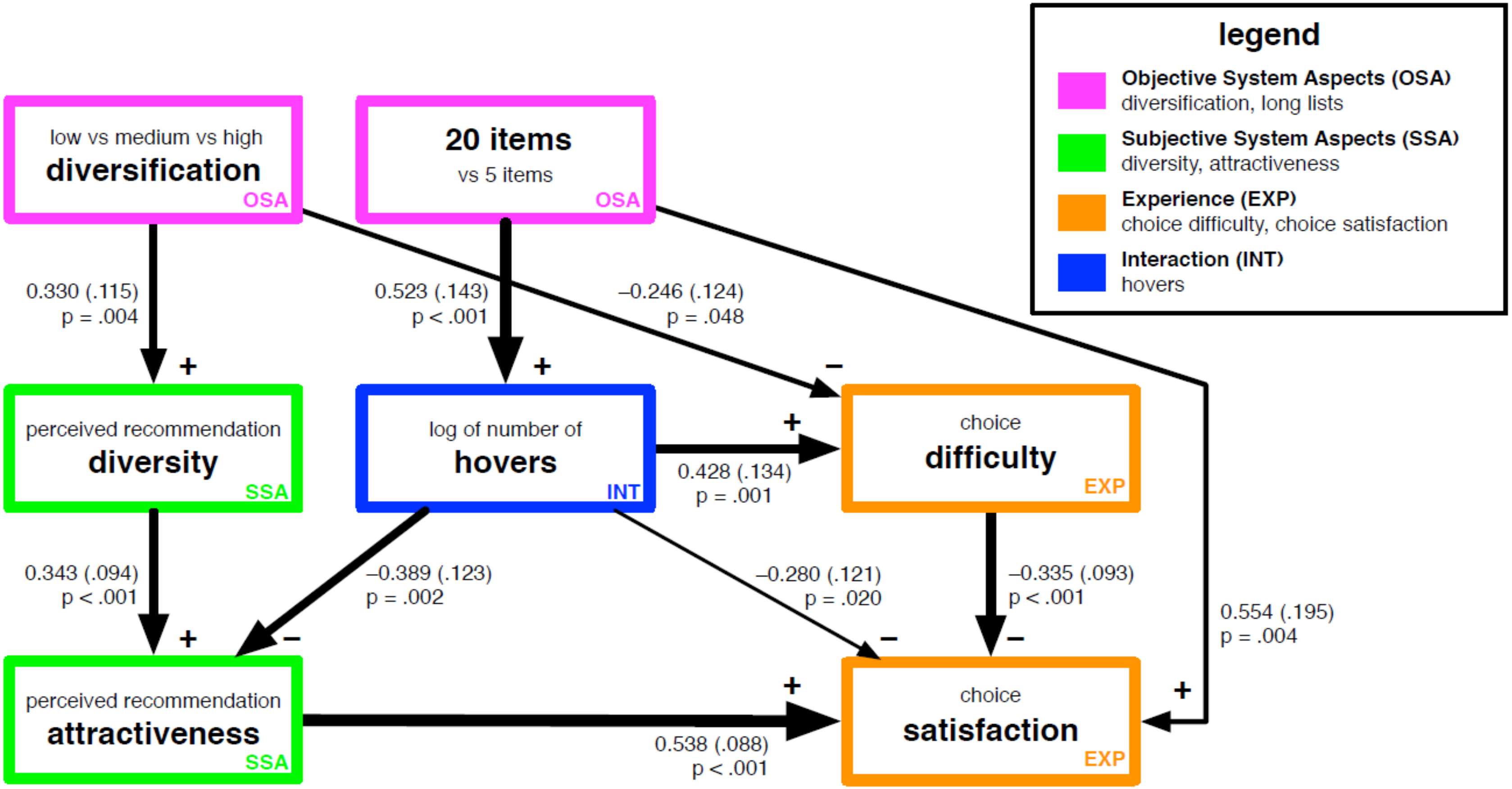


Figure 2. A simplified illustration of the late ... ch, which characterizes both users and movies us ng ... e versus female and serious versus escapist.

Design/procedure of study 2b

- 159 Participants from an online database
- **Rating task to train the system (15 ratings)**
- **Choose one item from a list of recommendations**
  - Between subjects: 3 levels of diversification (none, med, high), 2 lengths: 5 and 20 items **(OSA)**
- **Afterwards we measured:**
  - **Perceived recommendation diversity (Perception, SSA)**
    - 5 items, e.g. “The list of movies was varied”
  - **Perceived recommendation attractiveness (Perception, SSA)**
    - 5 items, e.g. “The list of recommendations was attractive”
  - **Choice satisfaction (experience, EXP)**
    - 6 items, e.g. “I think I would enjoy watching the chosen movie”
  - **Choice difficulty (experience, EXP)**
    - 5 items, e.g.: “It was easy to select a movie”
  - **Behavior (interaction, INT):** total views / unique items considered

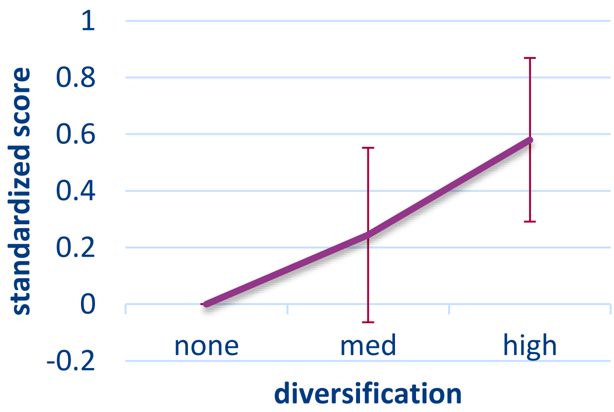
- Full SEM model (for which we won't have time…)



Latent Feature Diversification

System (OSA)	Perception (SSA)	Experience (EXP)	Interaction (INT)
Psychology-informed Diversity manipulation	Increased perceived Diversity & attractiveness	Reduced difficulty & **increased satisfaction**	**Less hovers** More choice **for lower ranked** items

Choice Satisfaction



Diversification	Rank of chosen
None (top 5)	3.6
Medium	14.5
High	77.6

Higher satisfaction for high diversification, despite choice for lower predicted/ranked items

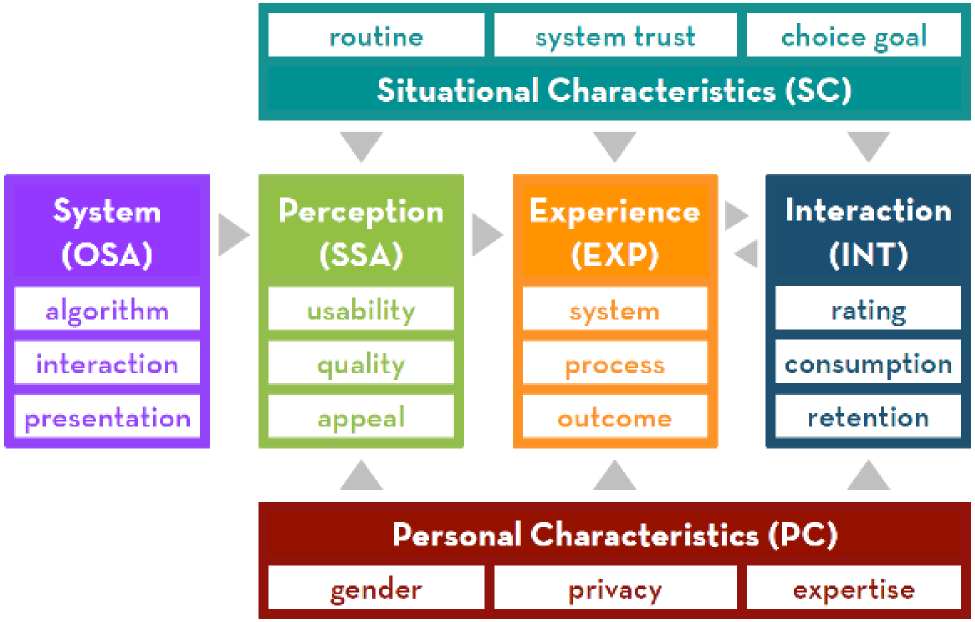
Concluding…

- Objective and subjective measures are both needed to understand what we are trying to improve/optimize
- Interpreting ‘easy to get’ behavioral data might require careful user experimentation to understand the meaning…
- Measuring subjective constructs like perceived diversity, accuracy and satisfaction can help understand WHY things work or not

Technische Universiteit **Eindhoven** University of Technology

Tutorial on user experiments

Using the user-centric evaluation Framework



Martijn Willemsen
Christine Bauer

This tutorial is largely based on

Knijnenburg, B. P., & Willemsen, M. C. (2015). Evaluating Recommender Systems with User Experiments. In F. Ricci, L. Rokach, & B. Shapira (Eds.), Recommender Systems Handbook (pp. 309–352). Springer US.link to springer

And some blatant copying of Bart Knijnenburgs' Tutorial slides (Recsys 2012), see http://bit.ly/recsystutorialhandout

Definition of a user experiment:

A user experiment is a scientific method to investigate how and why system aspects influence the users' experience and behavior.

For this tutorial I wil take it a bit broader: how can you evaluate your recommender algorithm, tool or result with users?

- Could be a large scale user satisfaction experiment, but also a small sale expert evaluation of your new user interface or data visualization!

- We will work in groups of 2-3 to go through the steps of designing a user experiment!

Assignment

- Team up with a group of 2-3 persons (one with a Spotify account!)
- Test our genre exploration tool https://spotify.vlab.nl/explore
- Take it seriously, generate playlist with a particular setting of the slider, check it and press (save playlist) to save it to your Spotify account. After that you will get a short questionnaire.
  - 1. * Are you familiar with the selected genre?
  - 2. * How often do you listen to songs from that genre?
  - 3. * How satisfied would you be with the generated playlist?

- Write down for yourself what dependent measures we have (both experience and interaction measures)

The 5 Steps for today (see practical guidelines in the chapter)

1. **Research Model:** what are you going to test, what question do you want to answer and to what will you compare?
2. **Participants:** considerations about your sample
3. **Experimental setup**: what conditions to test and how?
4. **Measurement:** develop scales
5. **Statistical Evaluation:** t-tests or structural equation models?

Step 1: Building a research model

When is your algorithm or system good/successful?

Define success: accuracy, CTR, usability, satisfaction?

NOT: Can we test if our new algorithm scores high on satisfaction?”

What is high? 3.6 on a 5 point scale?

BETTER: Does the new algorithm scores high on satisfaction compared to this other system?

Apply the concept of ceteris paribus to get rid of confounding variables: **keep everything else the same**

Building your own research model:

- Determine the outcome measure, is it **EXP** or **INT** or both (remember the clip recommender!)
  - Are you able to survey the users?
  - Are you able to get good user data (does the system log ?)

- Determine what aspect you want to test (which **OSA**?) is there theory/evidence that supports that OSA?

- Do you have theory that explains why the effect might happen: **SSA**?
  - Are there mediating constructs that can explain?

Step 2: Participants

Test on an unbiased sample…

At least test on a population of representative users

- these are typically not your colleagues…why?
- These are typically not you facebook friends… why?

Sample size:

Don't underestimate the size of the sample needed…

Perhaps use within designs (step 3)

Anticipated effect size	Needed sample size
Small	385
Medium	54
Large	25

DIY: step 1 & 2

Determine what you want to test (when will you be successful?)

How to measure it (INT/EXP)?

(What are the users, and can you sample enough?)

What potential manipulations (OSA)?

Are there explaining constructs (SSA)?

Genre exploration tool has two degrees of freedom

Genre selection

- Users pick their own genre to explore

Slider to explore genre-typical versus more personalized lists

- Users can get songs that are mostly representative for that genre or more personalized

Potential research questions? -> potential measures!

Dependent measure (INT/EXP):

- subjective: liking, usefulness, helpfulness
- objective/behavioral: slider usage, checking items, save playlist

SSA: What intermediate variables van explain the experience or interaction?

- personalization, recommendation quality, perceived control…

Step 3: Experimental setup

What is the right **baseline** to test your treatment (OSA) against?

Test against a reasonable alternative!

- Non-personalized or random system: might be a too easy win…
- Test against state-of-the-art (but small effects?)

Randomize assignment to conditions

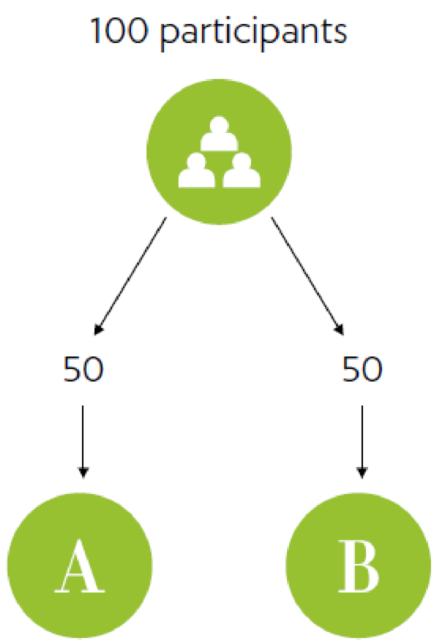
- **Bad:** first 10 users get system A, the second 10 users get system B

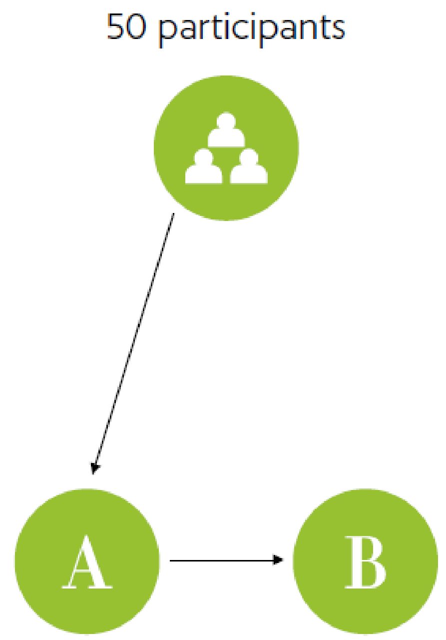
Randomization neutralizes (but doesn't eliminate) participant variation

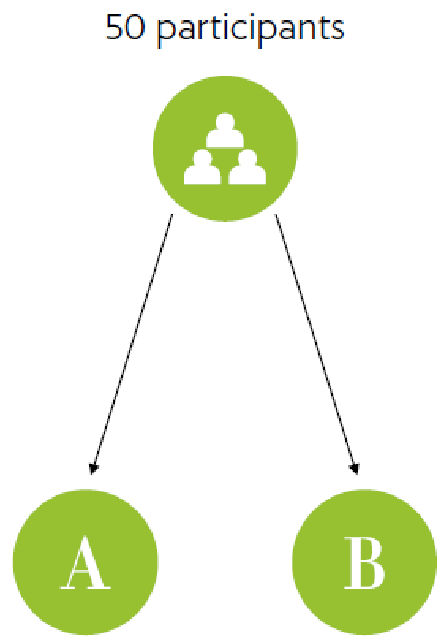
Within or between designs?

- Within designs have more power, but can be unrealistic…
- (life is a between-subjects experiment, D. Kahneman)

Between subject	Within-subjects design	Within-subjects design
Randomly assign half the participants to A, half to B • Realistic interaction • hidden from user • Many participants	Give participants A first, then B • Remove subject variability • Manipulation may be visible • Spill-over effect	Show participants A and B simultaneously • Remove subject variability • Participants can compare conditions: subtle differences detectable • Not a realistic interaction







DIY: step 3

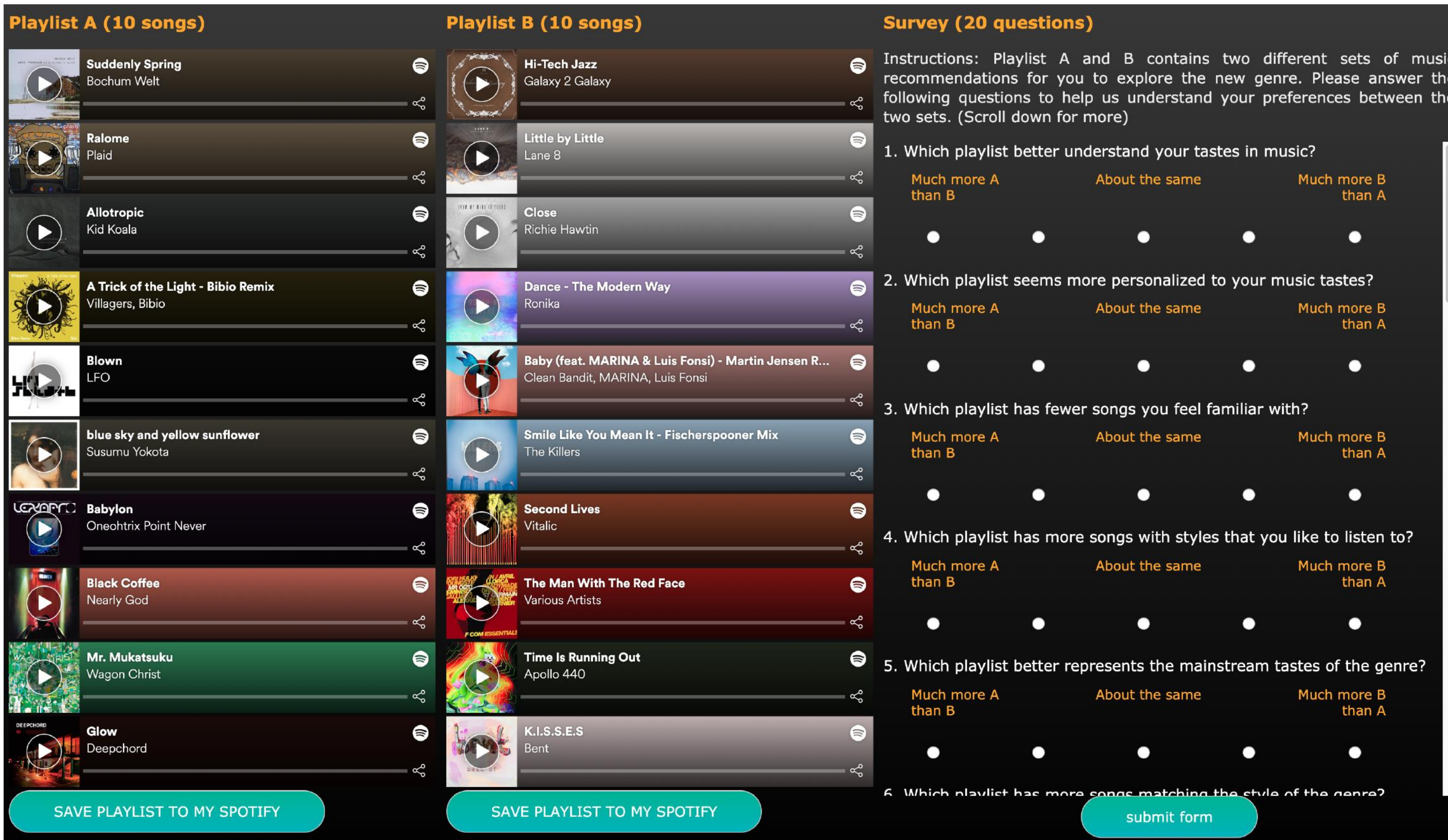
Think about a reasonable baseline…

Do you have normal or expert users?

Can you randomize conditions?

Within or between design?

- Liang & Willemsen, UMAP 2019



Step 4: Measurement

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

Does the question mean the same to everyone?

- John likes the system because it is convenient, Mary it because it is easy to use, Dave likes it because the recommendations are good

We need a multi-item measurement scale…

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!

Choose simple over specialized words,

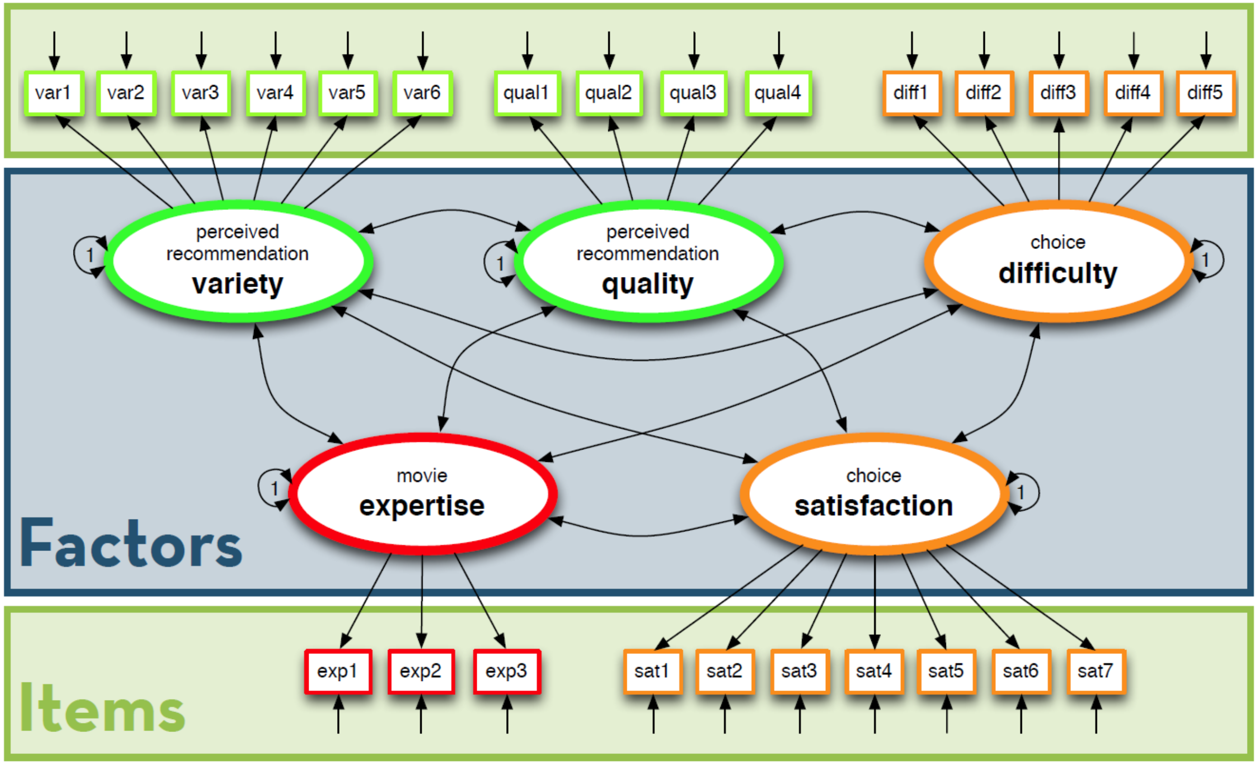
Avoid double-barreled questions

Use existing (validated) scales as much as possible

Factor analysis:

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

- This makes sure the scales are unidimensional



DIY: step 4

Try to construct a set of questions for a subjective measure in your study

Define the concept

Think of positive and negative items

Use existing scales for inspiration

Framework paper: http://Bit.ly/umuai

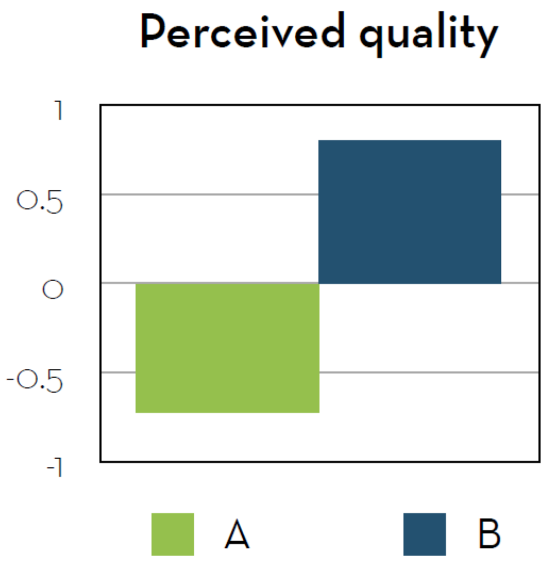
- Concepts we used: perceived accuracy (quality), perceived personalization, representative for genre, helpfulness and diversity

Considered aspects	Items	SEM Coef.
Accuracy	Which playlist has more songs that you find appealing?	0.949
Alpha: 0.96	Which playlist has more songs that you might listen to again?	0.942
AVE: 0.87	Which playlist has more obviously bad songs for you?	
	Which playlist has more songs that are well-chosen?	
Personalization (formerly)	Which playlist better understands your tastes in music?	0.933
	Which playlist seems more personalized to your music tastes?	0.876
	Which playlist has fewer songs you feel familiar with?	
	Which playlist has more songs with styles that you like to listen to?	0.947
Representativeness	Which playlist better represents the mainstream tastes of the genre?	
Alpha: 0.81	Which playlist has more songs matching the style of the genre?	0.818
AVE:0.65	Which playlist has fewer songs you would expect from the genre?	−0.772
	Which playlist seems less typical of the genre?	−0.779
Helpfulness	Which playlist better supports you to get to know the new genre?	0.716
Alpha: 0.77	Which playlist motivates you more to delve into the new genre?	
AVE: 0.61	Which playlist is more useful to explore a new genre?	0.626
	Which playlist has more songs that helps you understand the new genre?	0.402
Diversity	Which playlist has more songs that are similar to each other?	
Alpha: N.A.	Which playlist has a more varied selection of songs within the genre?	
AVE: N.A.	Which playlist would suit a broader set of tastes?	
	Which playlist has songs that match a wider variety of moods?	

Step 5: Statistical Evaluation

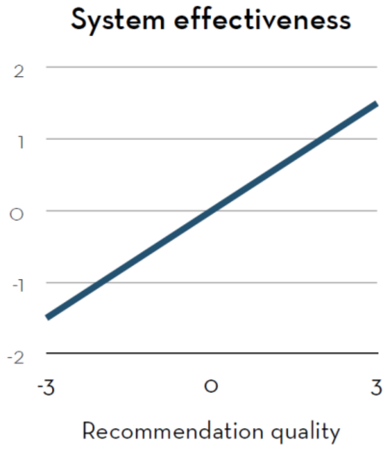
T-tests for simple one-factor designs:

Do these two algorithms lead to a different level of perceived quality?



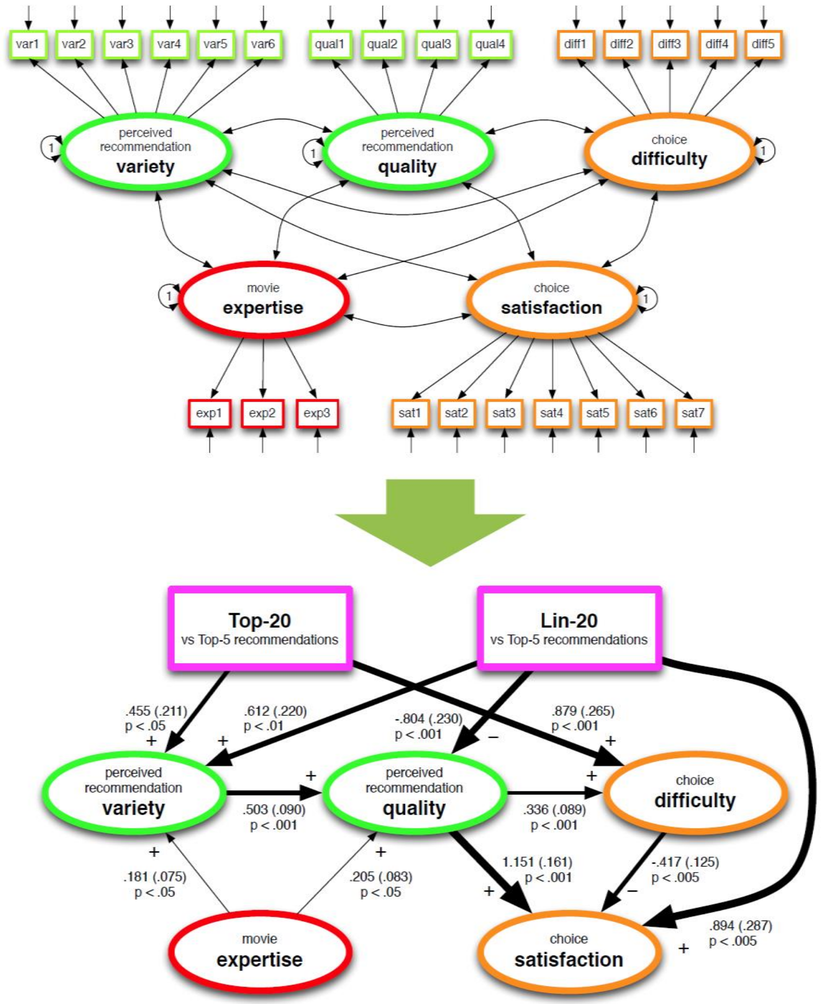
Regression for linear relations

Does perceived quality influence system effectiveness?



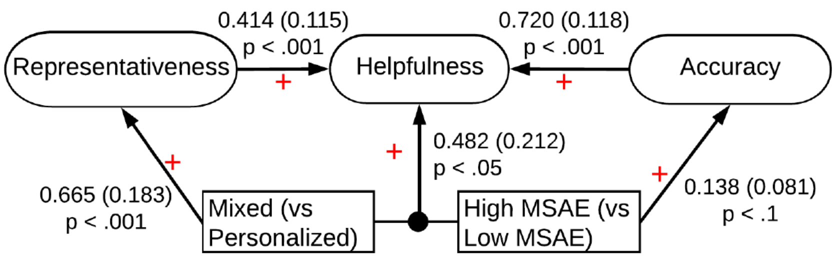
Structural equation models

- Combines factor analysis and path models
- Complex analysis requires dedicated software and knowledge (mplus/stata/R etc.)
- Allows for answering ‘Why’ effects via mediation

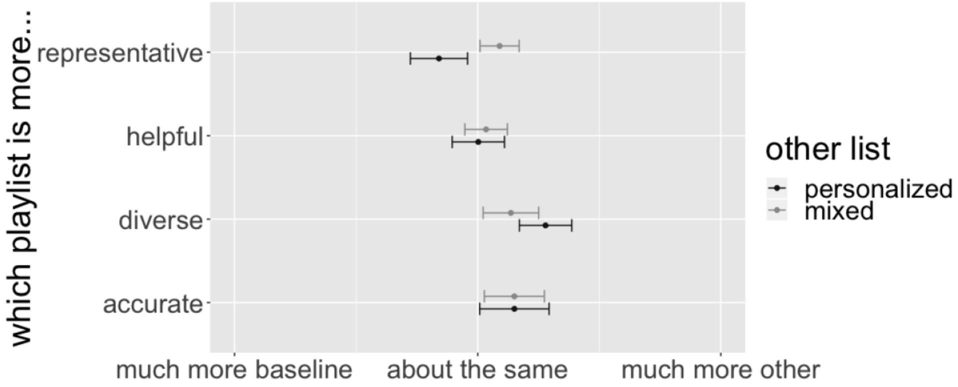


DIY: step 5

Let's have a look at the models/results from the UMAP paper



SEM Model to understand the relations between concepts



Direct comparison of the concepts between baseline and other list (paired t-tests)

Further inspection of the interaction of condition and MSAE (expertise)