The Why behind effective recommenders: user perception and experience

Martijn Willemsen

Technische Universiteit
Eindhoven
University of Technology

Where innovation starts

What are recommender systems about

Rating?

~~ratings~~
preferences

user

Recommendation:
best predicted items

Choose (prefer?)

Experience!

dataset
user-item rating pairs

Chris	2	0	3	2	5
Martijn	0	5	4	0	0
Jaap	4	2	5	4	4
Cees	4	3	4	1	5
Dirk	4	5	3	2	1
Alain	3	0	5	0	3

Accuracy:
compare prediction
with actual values

Technische Universiteit
Eindhoven
University of Technology

Agenda for today

User-centric Evaluation Framework

Understanding and improving algorithm output

- User perceptions of recommendation Algorithms (Ekstrand et al., RecSys 2014)
- Latent feature diversification to improve algorithm output (Willemsen et al., 2011, under review)

Understanding and improving the input of a recommender algorithm: preference elicitation!

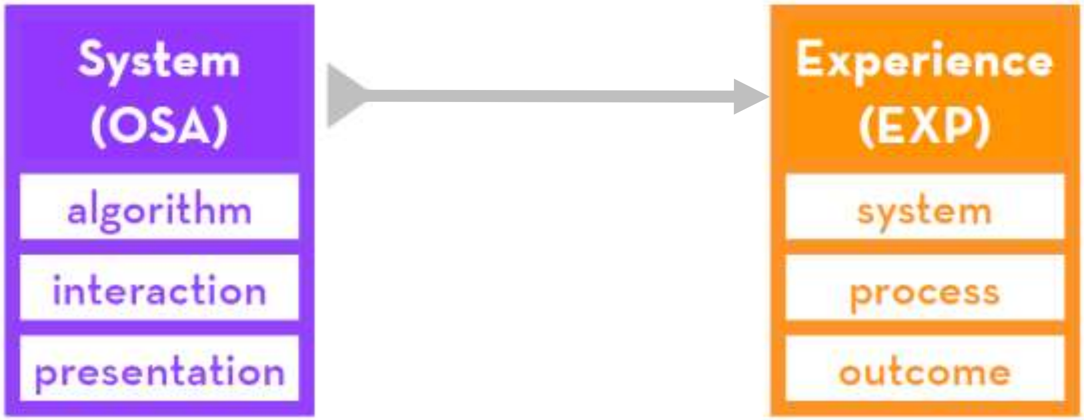
- Comparing choice-based PE with rating-based PE (Graus and Willemsen, RecSys 2015)
- Matching PE-techniques to user characteristics (Knijnenburg et al., Amcis 2014, Recsys 2009 & 2011)

User-Centric Framework

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

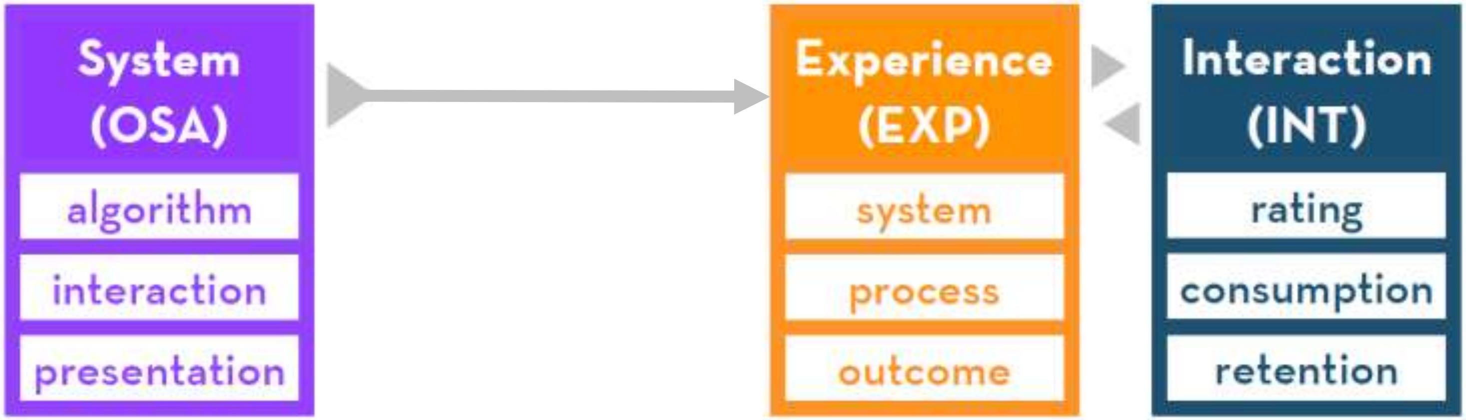
User-Centric Framework

Psychologists and HCI people are mostly interested in experience…



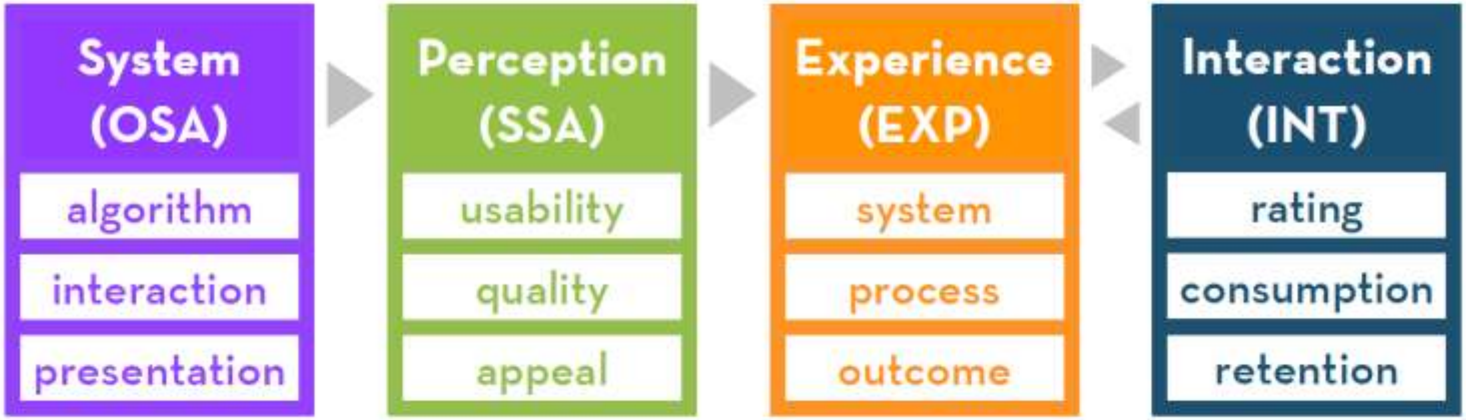
User-Centric Framework

Though it helps to triangulate experience and behavior…



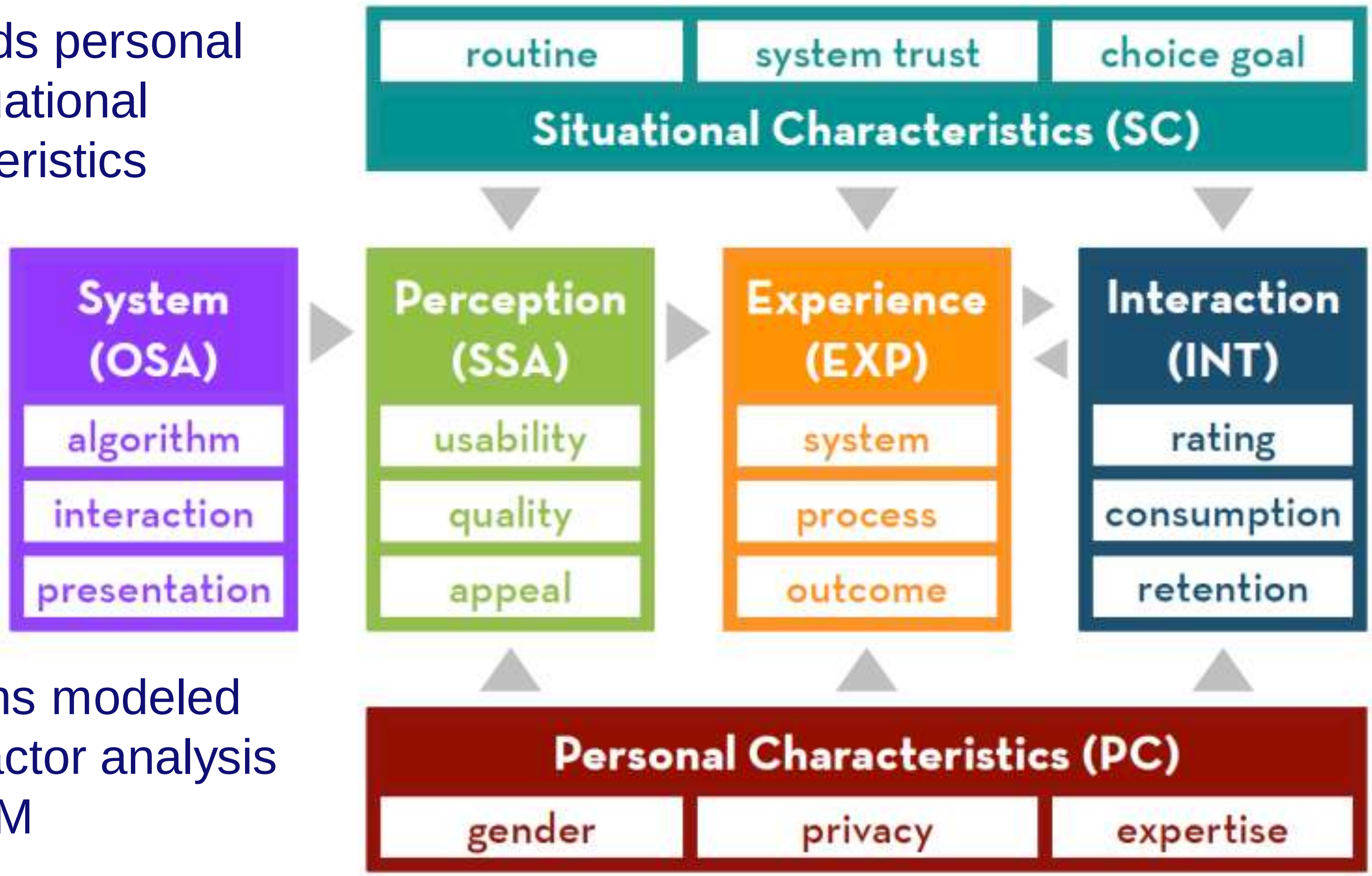
User-Centric Framework

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

User Perceptions of Differences in Recommender Algorithms

Joint work with grouplens
Michael Ekstrand, Max Harper and Joseph Konstan, RecSys 2014

Technische Universiteit
Eindhoven
University of Technology

Where innovation starts

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…

Joint forces between CS (grouplens) and Psy (me)

Goals of this paper

RQ1
How do subjective perceptions of the list affect choice of recommendations?

RQ2
What differences do users perceive between lists of recommendations produced by different algorithms?

RQ3
How do objective metrics relate to subjective perceptions?

Taking the opportunity…

Movielens system

- 3k unique users each month

Launching a new version

- Experiment was communicated as an intro for beta testing

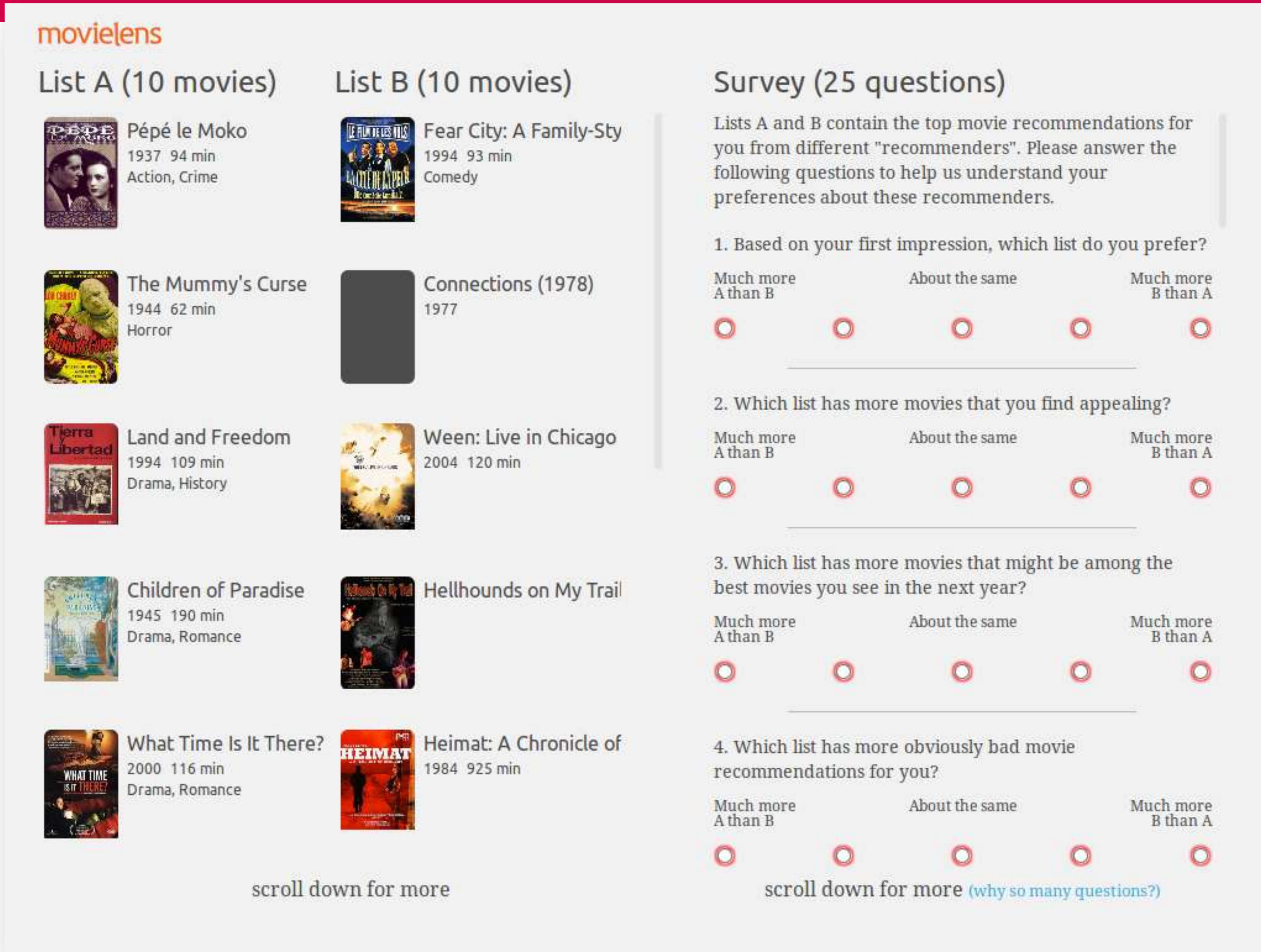
Comparing 3 ‘classic’ Algorithms

- User-user CF
- Item-item CF
- Biased Matrix Factorization (FunkSVD)

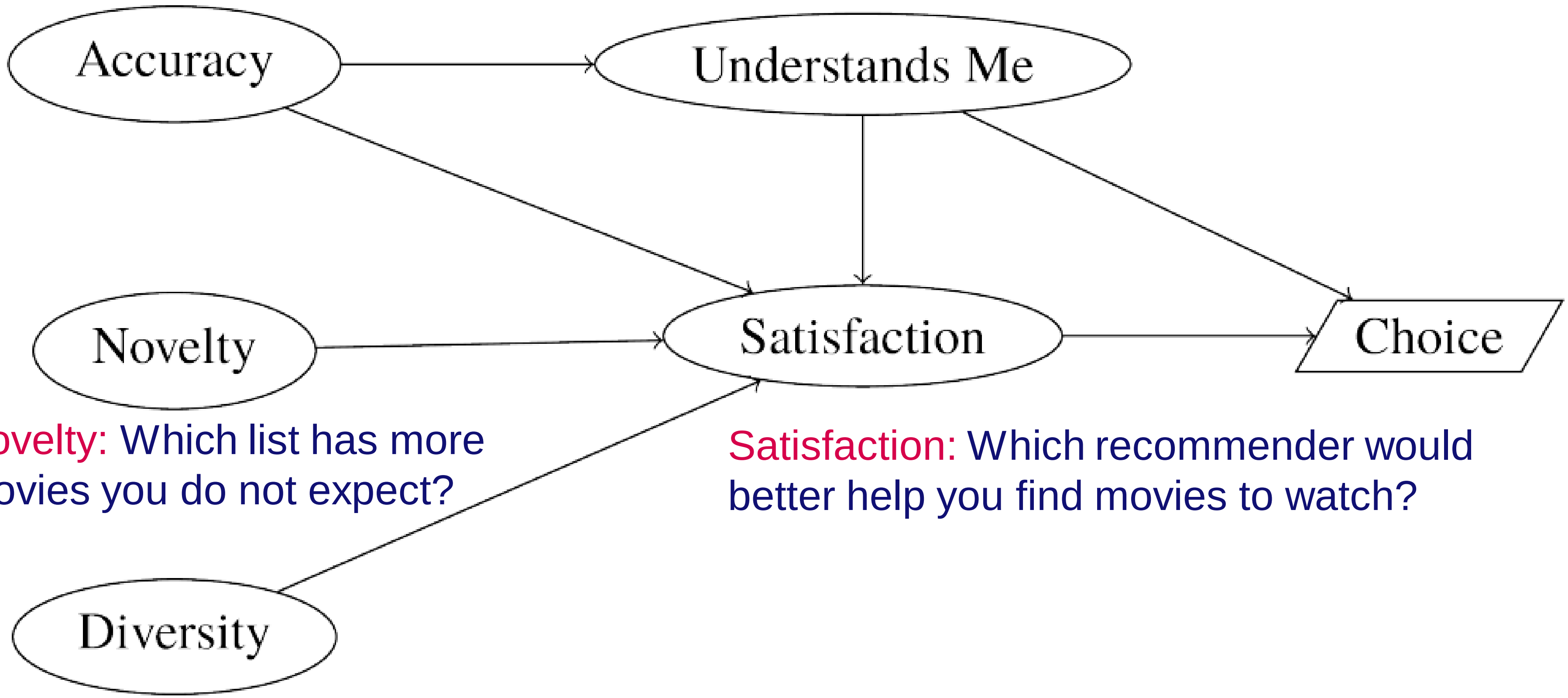
User compares 2 algorithm outputs side by side

- Joint evaluation is more sensitive to small differences…
- And a pain to analyse ☹

The task provided to the user



Concepts and User perception model



Novelty: Which list has more movies you do not expect?

Satisfaction: Which recommender would better help you find movies to watch?

Diversity: Which list has a more varied selection of movies?

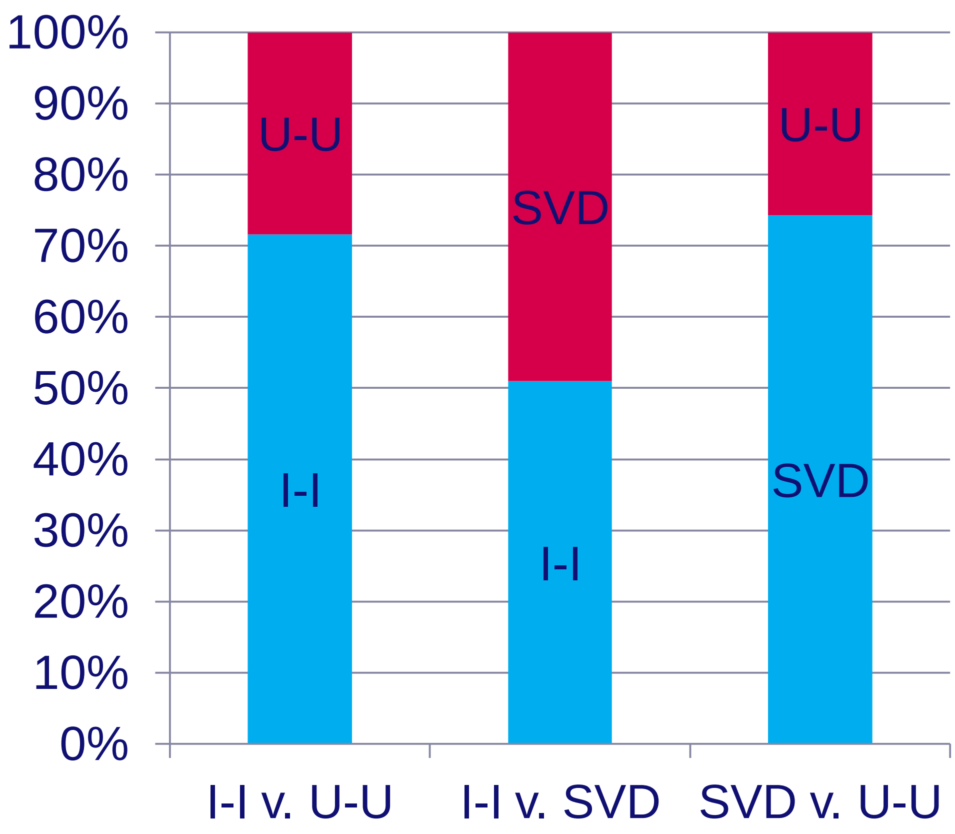
What algorithms do users prefer?

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



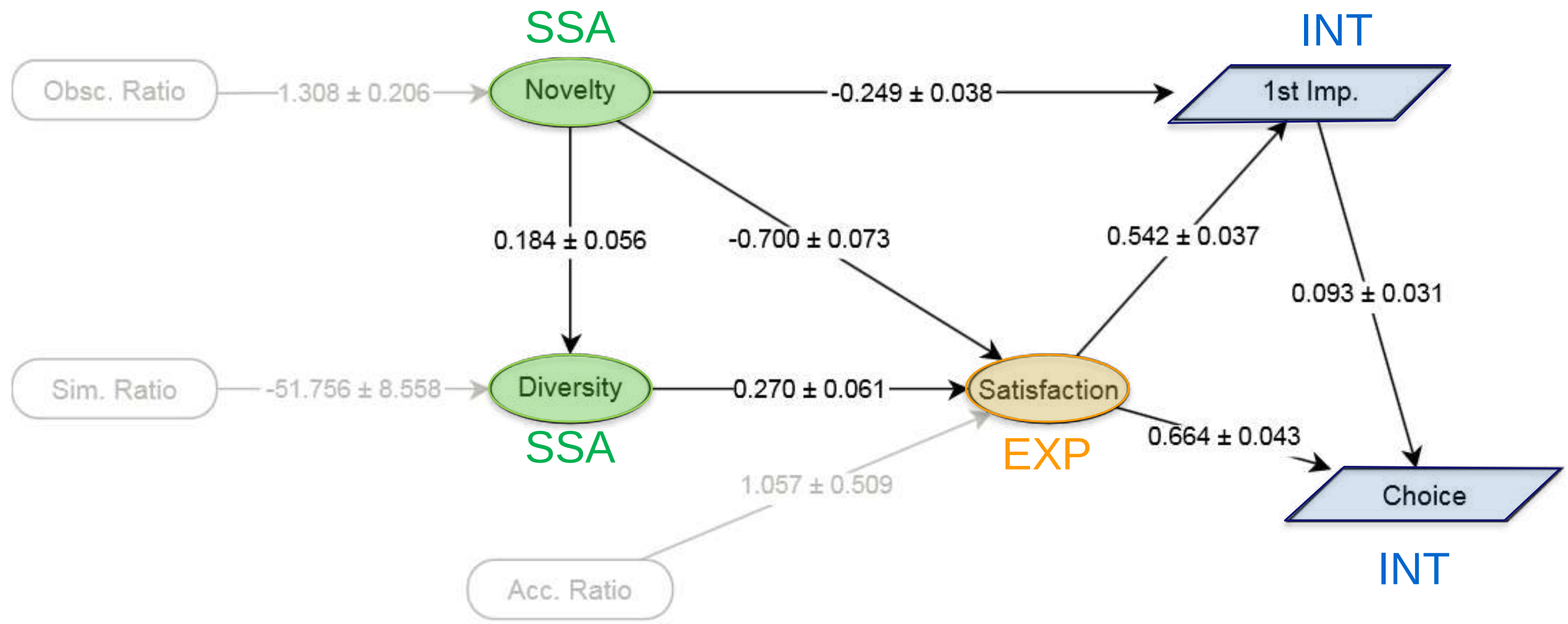
Why? First looking at the measurement model

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

Perceived accuracy and ‘understands me’ not in model



Differences in perceptions between algo’s

RQ2: Do the algorithms differ in terms of perceptions?

Separate models (pseudo-experiments) to check each pair

- 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)

Relate Subjective and Objective measures

RQ3: How do objective metrics relate to subjective perceptions?

Novelty

- obscurity (popularity rank)

Diversity

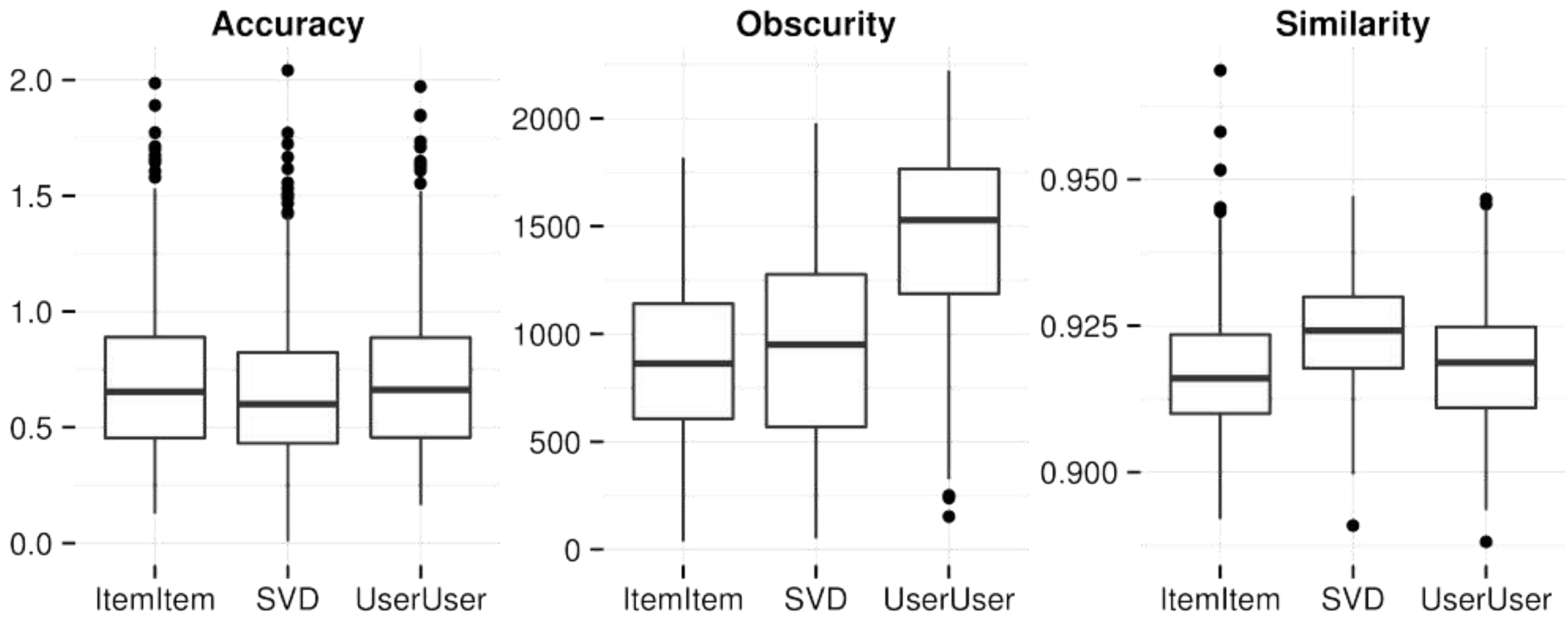
- intra-list similarity (Ziegler)
- Similarity metric: cosine over tag genome (Vig)

Accuracy (~Satisfaction)

- RMSE over last 5 ratings

Objective measures

No accuracy differences, but consistent with subjective data
RQ2: User-user more novel, SVD somewhat less diverse

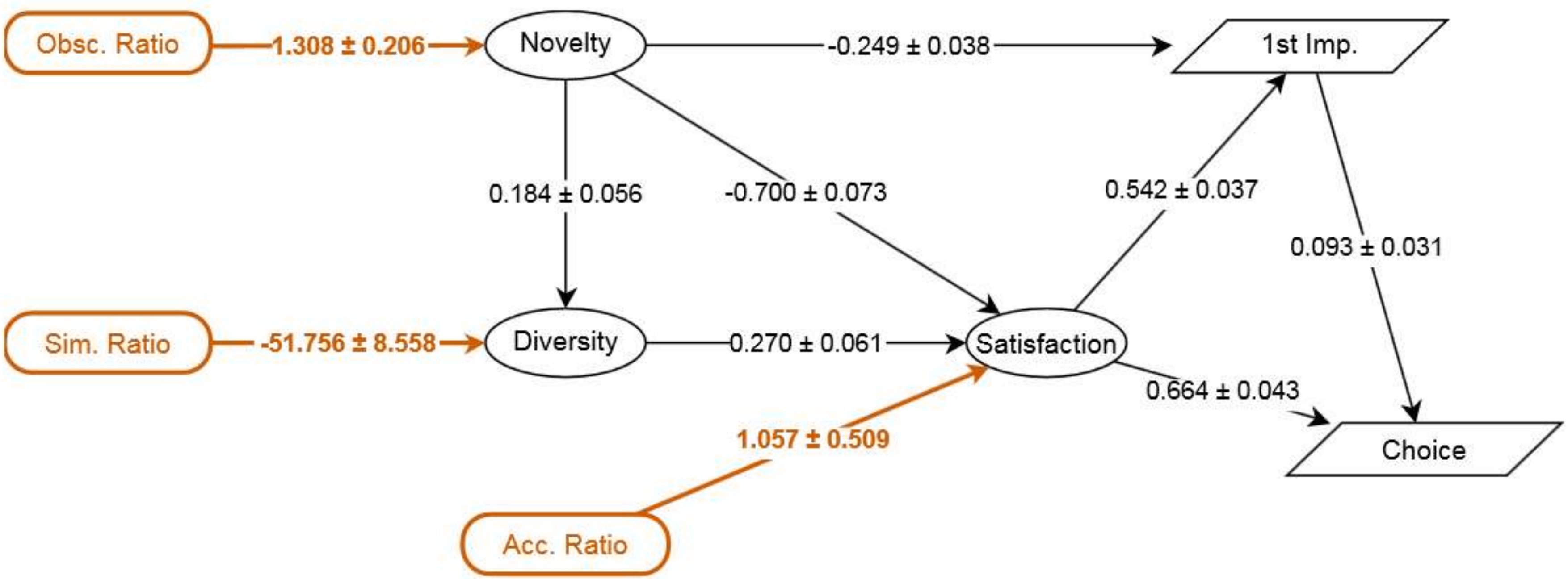


RQ3: Aligning objective with subjective

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…

User-user loses (likely due to obscure recommendations), but users are split on item-item vs. SVD

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 it’s obsure items (which are perceived as novel)

Technische Universiteit Eindhoven University of Technology

Latent feature diversification

from Psy to CS

Joint work with Mark Graus and
Bart Knijnenburg (under review)

Technische Universiteit
Eindhoven
University of Technology

Where innovation starts

Choice Overload in Recommenders

Recommenders reduce information overload…

- But large personalized sets might cause **choice overload**!
- Top-N of all highly ranked items

  What should I choose?
  These are all very attractive!

Choice Overload

Seminal example of choice overload

Less attractive

30% sales

Higher purchase satisfaction

From Iyengar and Lepper (2000)

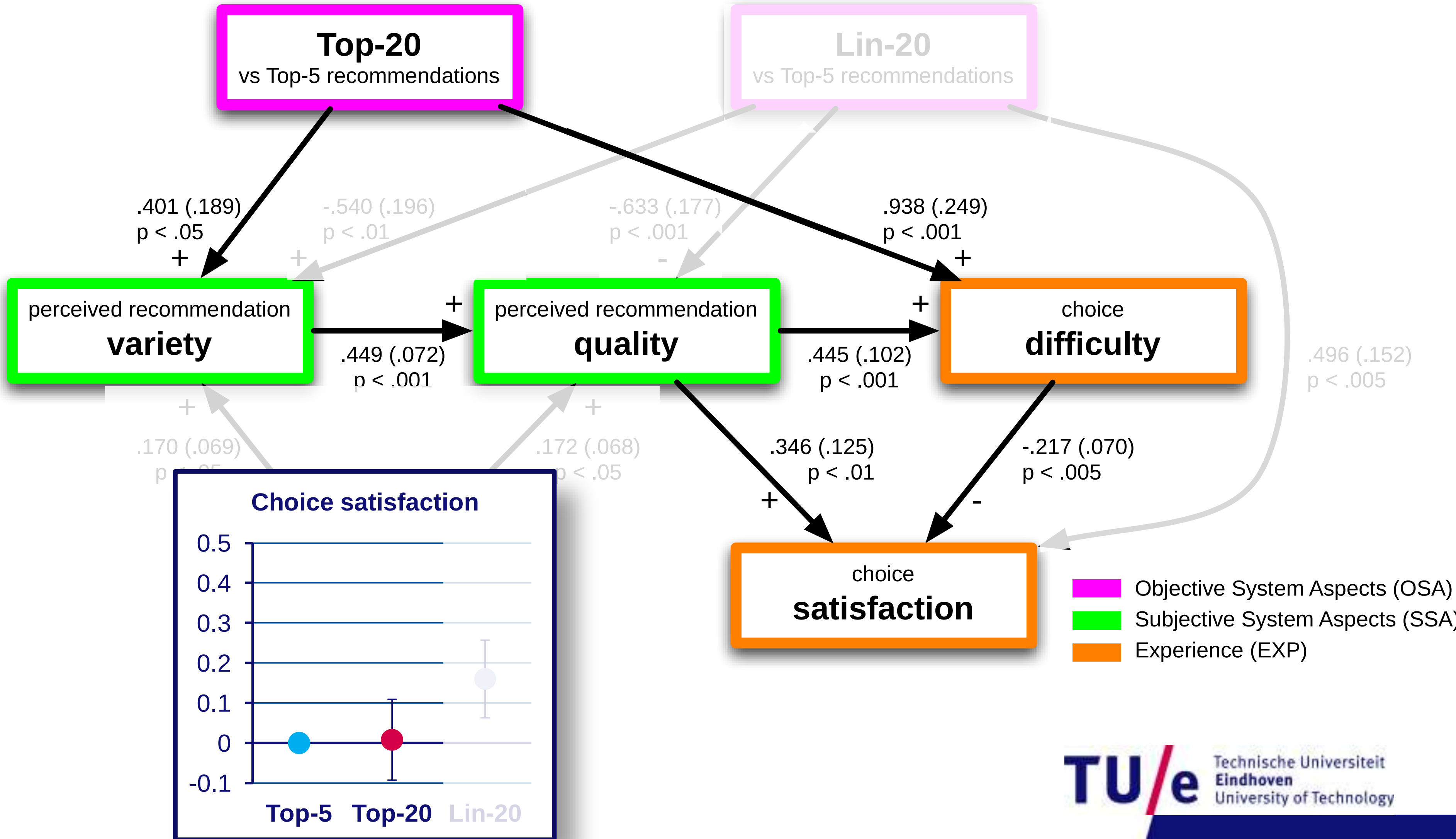
More attractive

3% sales

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

http://www.ted.com/talks/sheena_iyengar_choosing_what_to_choose.html **(at 1:22)**

Choice Overload in Recommenders
(Bollen, Knijnenburg, Willemsen & Graus, RecSys 2010)

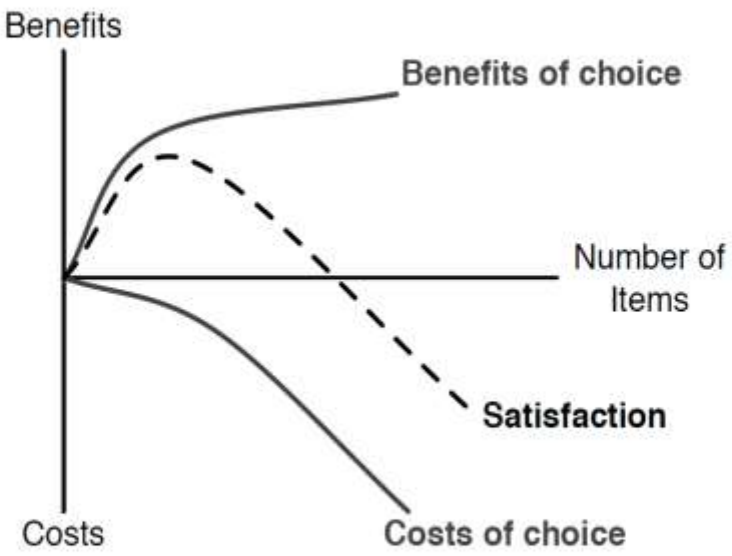


Satisfaction and item set length

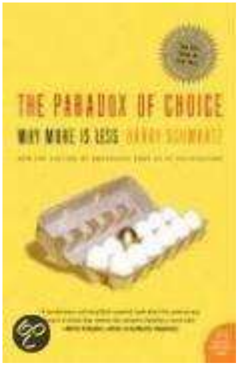
More options provide more benefits in terms of finding the right option…

…but result in higher opportunity costs

- More comparisons required
- Increased potential regret
- Larger expectations for larger sets



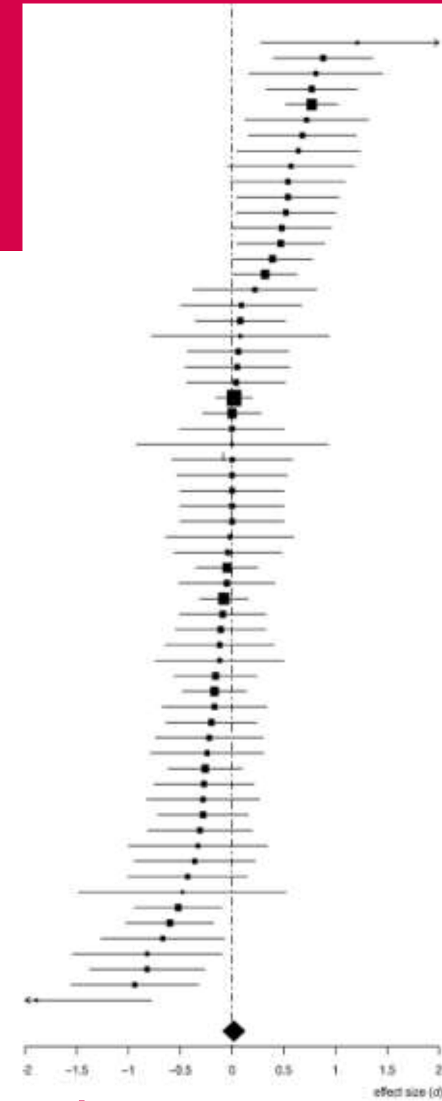
Paradox of choice (Barry Schwartz)



http://www.ted.com/talks/barry_schwartz_on_the_paradox_of_choice.html

Technische Universiteit Eindhoven University of Technology

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…

Technische Universiteit Eindhoven University of Technology

Diversification and attractiveness

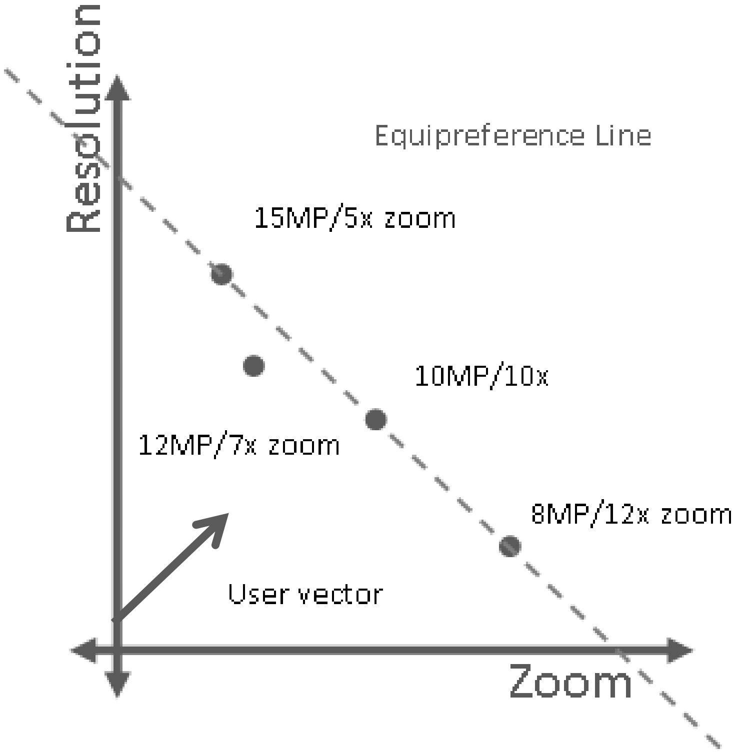
Camera:

Suppose Peter thinks resolution (MP) and Zoom are equally important

- user vector shows preference direction

Equi-preference line:

- Set of equally attractive options (orthogonal on user vector)
- Diversify over the equipreference line!



Matrix Factorization algorithms

	Usual Suspects	Titanic	Die Hard	Godfather
				
Jack	?	★☆☆☆☆	★★★★★	?
Dylan	?	?	★★★★☆	★★★★★
Olivia	★★☆☆☆	★★★★☆	★☆☆☆☆	★★☆☆☆
Mark	★★☆☆☆	?	?	?

$\mathbf{p}_u$	Dim 1	Dim 2
Jack	3	-1
Dylan	1.4	.2
Olivia	-2.5	-.8
Mark	-2	-1.5

Map users and items to a joint latent factor space of dimensionality *f*

$\mathbf{q}_i$	Usual Suspects	Titanic	Die Hard	Godfather
Dim 1	1.6	-1	5	0.2
Dim 2	1	1	.3	-.2

Each item is a vector q_i
each user a vector p_u

Predicted rating r: $\hat{r}_{ui} = q_i^T p_u$

'Understanding' Matrix Factorization

Dimensionality reduction:

- Users and items are somewhere on these dimensions
- Dimensions are latent (have no apparent meaning)
- But they represent some 'attributes' that determine preference
- We can diversify on these attributes!

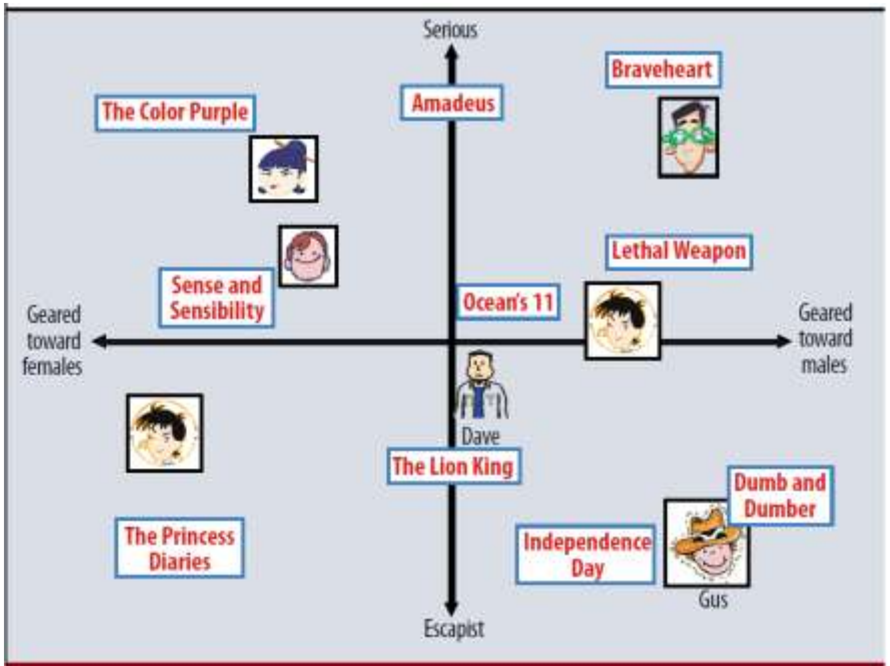
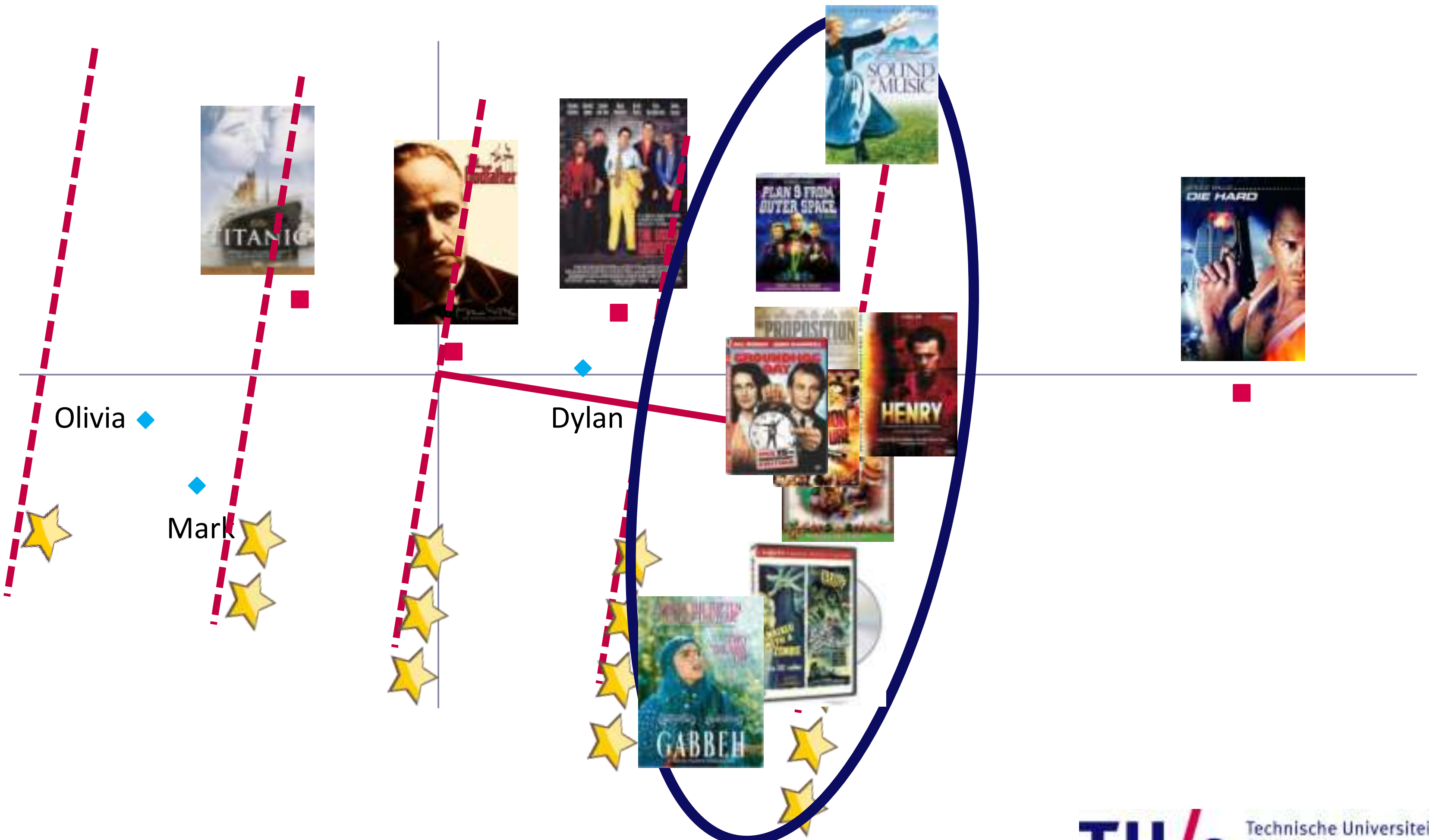


Figure 2. A simplified illustration of the latent factor approach, which characterizes both users and movies using two axes—male versus female and serious versus escapist.

Koren, Y., Bell, R., and Volinsky, C. 2009. Matrix Factorization Techniques for Recommender Systems. *IEEE Computer 42*, 8, 30–37.

Two-dimensional Latent feature space and diversification



Diversity Algorithm

10-dimensional MF model

- Take personalized top-N (200)

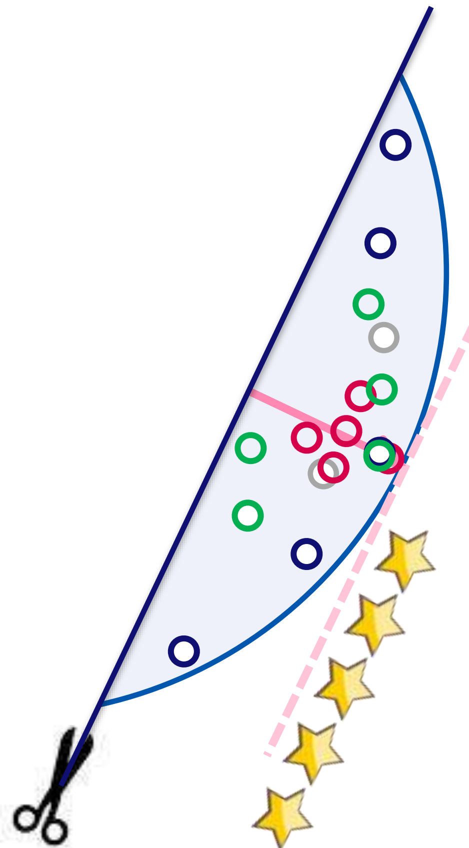
Greedy algorithm

- Select K items with highest inter-item distance (using city-block)

Low: closest to Top-1

High: from all items in top-N

Medium:
weigh item based on distance to other items and predicted rating



System characteristics

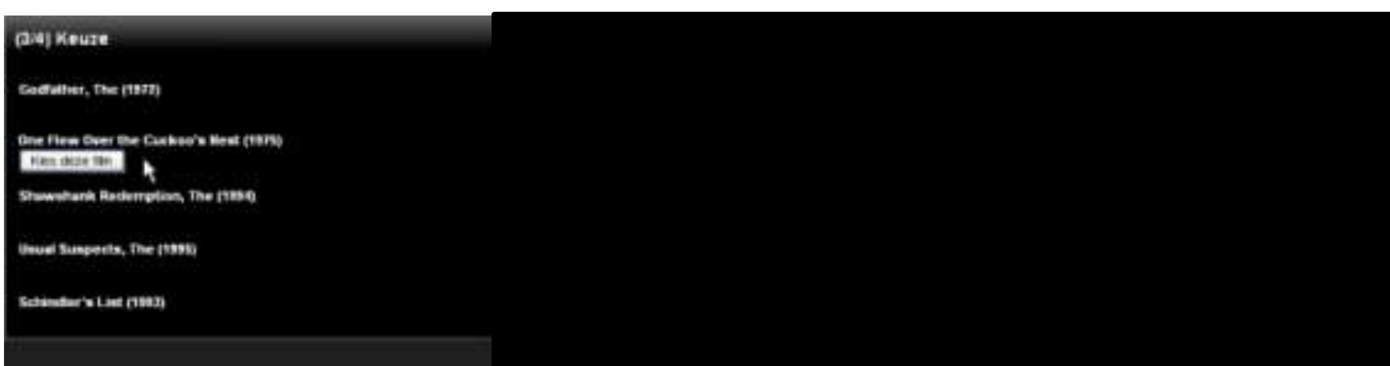
Fully functional Matrix Factorization recommender

10M MovieLens dataset: movies from 1994

- 5.6M ratings for 70k users and 5.4k movies
- RMSE of 0.854, MAE of 0.656

Movies shown with title and predicted rating:

- hovering the mouse over the title reveals additional information: short synopsis, cast, director and image



Study on Choice Satisfaction

Diversification and list length as two factors in a choice overload experiment

- **list sizes:** 5 and 20
- **Diversification:** none (top 5/20), medium, high

Dependent measure: choice satisfaction

- We expect choice overload to be more prominent for standard top-N sets

Design/procedure

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, 2 lengths

Afterwards we measured:

- *Perceptions*: Perceived Diversity & Attractiveness
- *Experience*: Choice Difficulty and Choice satisfaction
- *Behavior:* total views / unique items considered

Questionnaire-items

Perceived recommendation diversity

5 items, e.g. “The list of movies was varied”

Perceived recommendation attractiveness

5 items, e.g. “The list of recommendations was attractive”

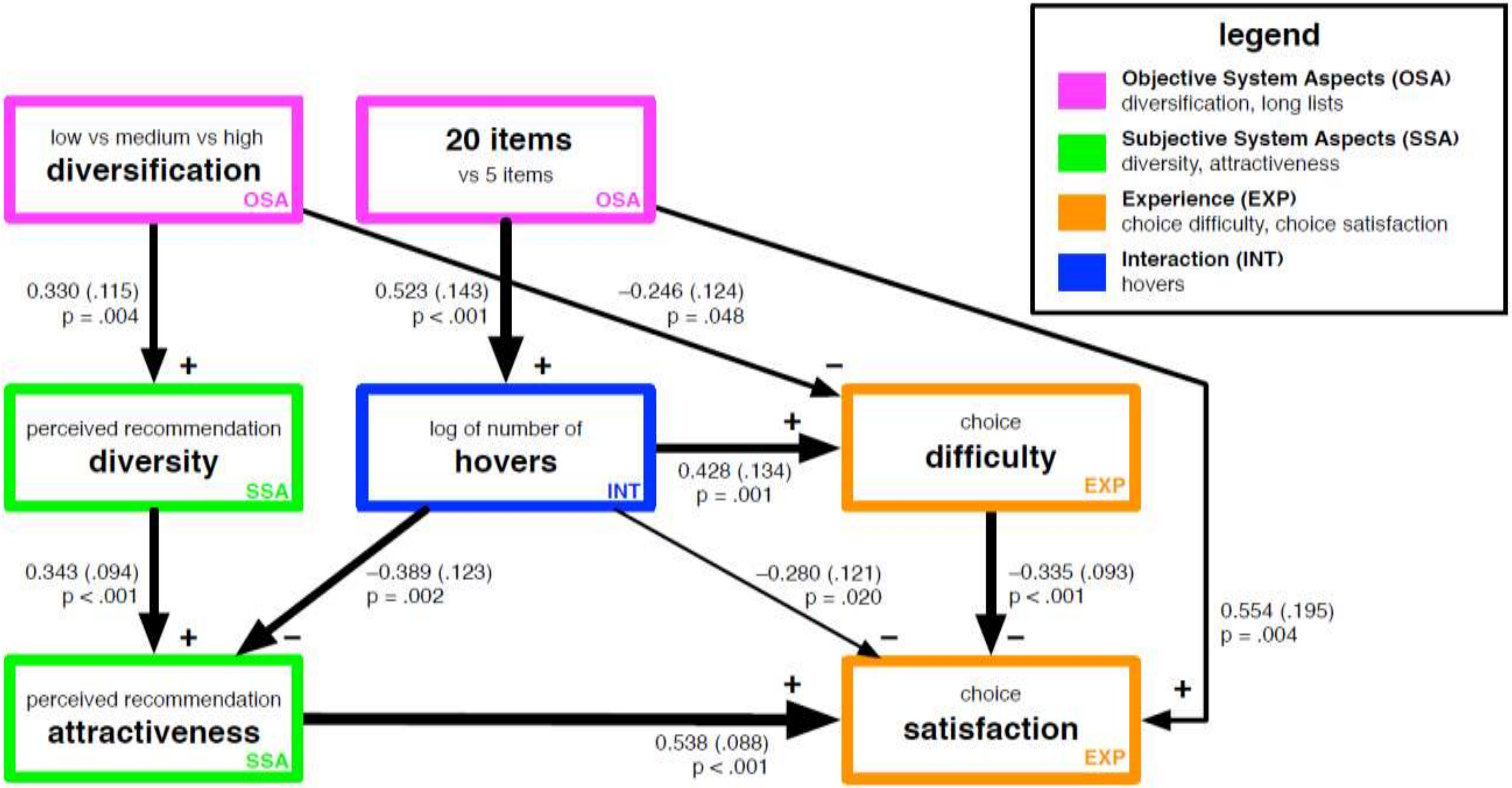
Choice satisfaction

6 items, e.g. “I think I would enjoy watching the chosen movie”

Choice difficulty

5 items, e.g.: “It was easy to select a movie”

Structural Equation Model

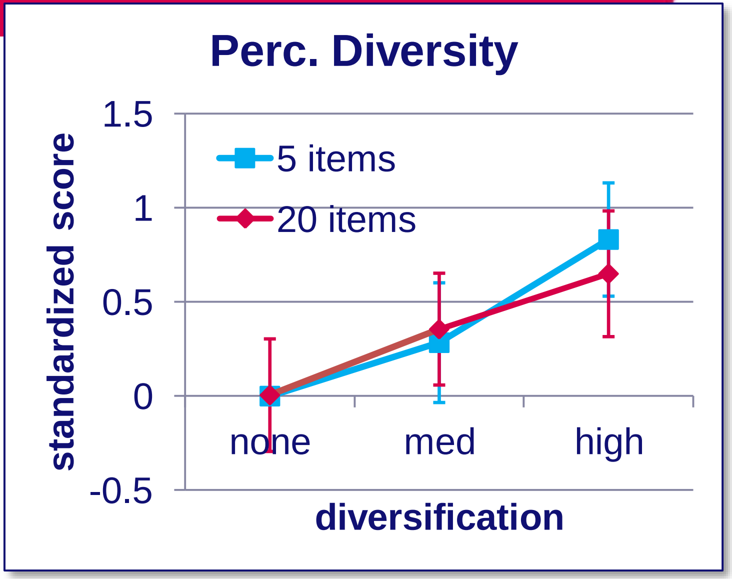


Technische Universiteit Eindhoven University of Technology

Perceived Diversity & attractiveness

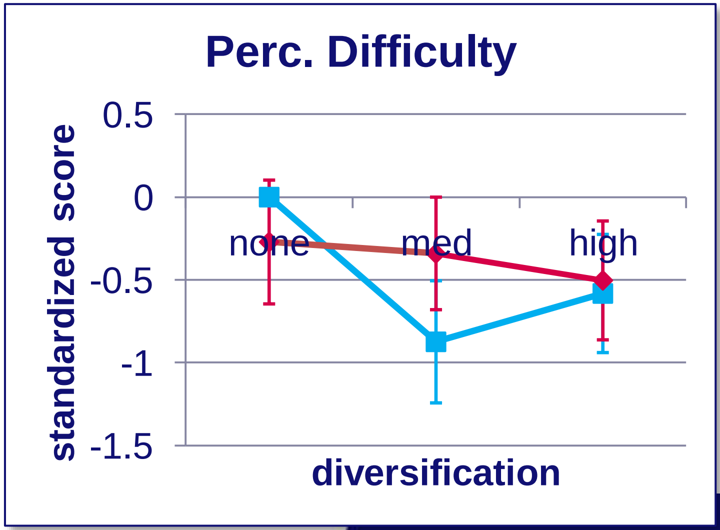
Perceived Diversity increases with Diversification

- Similarly for 5 and 20 items
- Perc. Diversity *increases* attractiveness



Perceived difficulty goes down with diversification

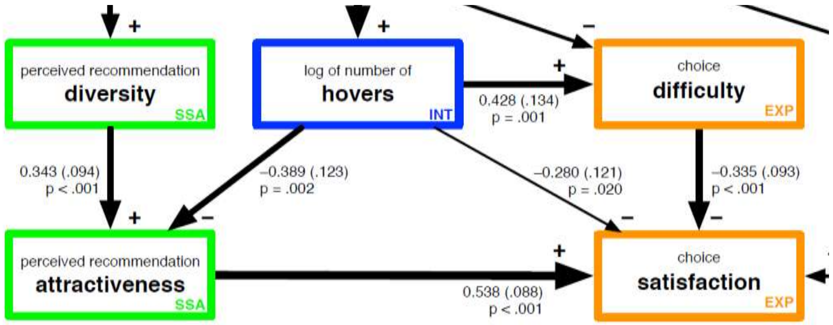
- 5 items lists are affected more by diversification



Difficulty and Satisfaction

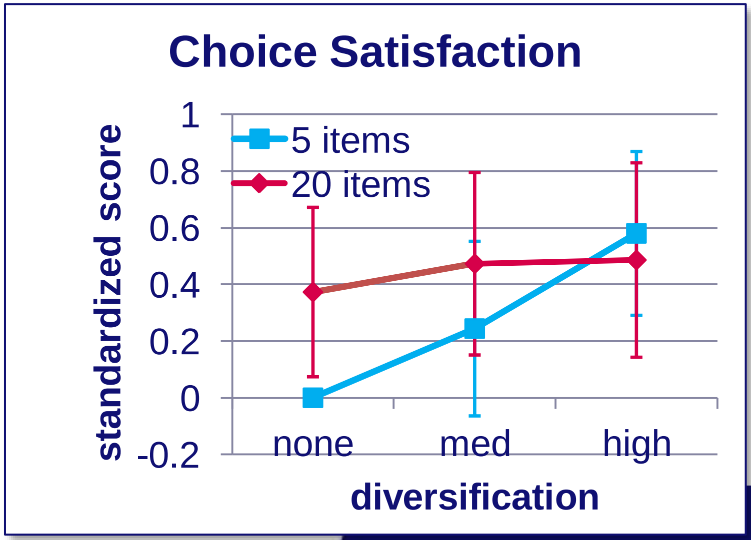
Satisfaction is an interplay between attractiveness and difficulty (as theorized)

Our diversity increases satisfaction especially for short 5 item sets.



Diverse 5 item set excels…

- Just as satisfying as 20 items
- Less difficult to choose from
- Less cognitive load…!



Choice Characteristics

Set	Diversity	Chosen option (mean and std. err) List Position	Rating	Rank
5 items	None (top 5)	3.60 (0.27)	4.51 (0.07)	3.60 (0.27)
	Medium	4.41 (0.59)	4.41 (0.07)	14.52 (5.37)
	High	4.19 (0.27)	4.30 (0.07)	77.59 (12.76)
20 items	None (top 20)	10.15 (0.92)	4.45 (0.05)	10.15 (0.92)
	Medium	10.33 (1.18)	4.40 (0.08)	17.7 (2.68)
	high	9.93 (1.07)	4.16 (0.07)	72.22 (11.84)

With higher diversity, no difference in position of chosen option
Resulting in less 'optimal' choice in terms of predicted rating

Without a reduction in choice satisfaction!

Conclusions

Reducing Choice difficulty and overload

Diversity reduces choice difficulty

Less uniform sets are easier to choose from

Diversity can improve choice satisfaction

Even when the diversified list has movies with lower predicted ratings than standard top-N lists

No need for larger item sets

Offering personalized diversified small items sets might be the key to help decision makers cope with too much choic**e!**

Psychological theory can inform how to improve the output of Recommender algorithms

Intermezzo

We have looked at algorithm output:

- Different perceptions of algorithms that drive satisfaction & choice
- Improve algorithm output based on psychological theory

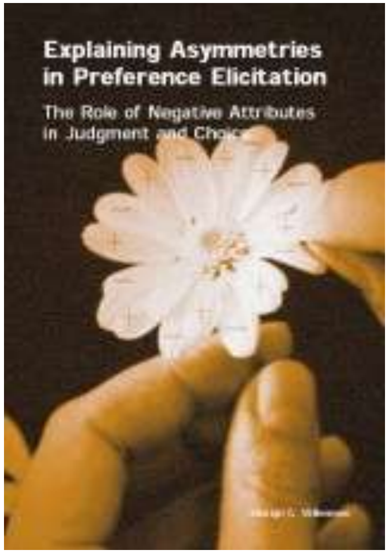
But how do algorithm get their data?

Preference Elicitation (PE)

PE is a major topic in research on Decision Making

- I even did my thesis on it… ;-)

What can Psychology learn us on improving this aspect?



Technische Universiteit Eindhoven University of Technology

Beyond ratings…

Choice-based PE

Martijn Willemsen

with Mark Graus

Technische Universiteit Eindhoven University of Technology

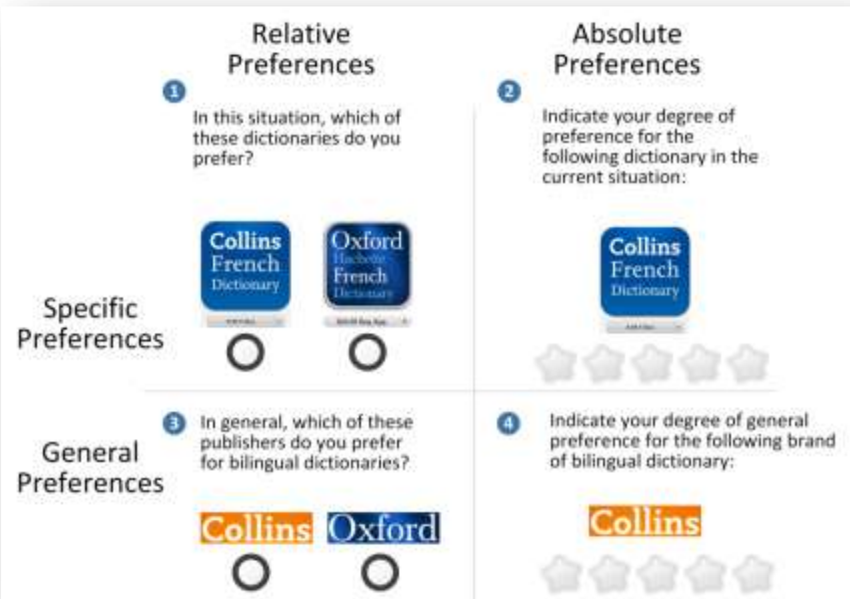
Where innovation starts

What are preferences?

Ratings are absolute statements

Preference is a relative statement!

- I like Grand Budapest hotel more then King's Speech



Which do you prefer?

Jameson et al., chapter in 2nd RecSys handbook

Choice-based preference elicitation

Choices are relative statements that are easier to make

- Better fit with final goal: finding a good item rather than making a good prediction

In Marketing, conjoint-based analysis uses the same idea to determine attribute weights and utilities based on a series of (adaptive) choices

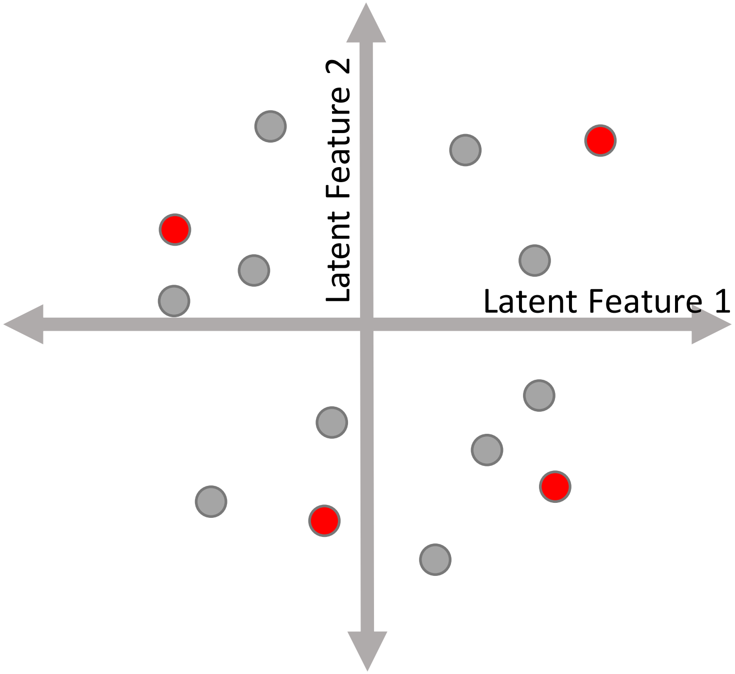
Can we use a set of choices in the matrix factorization space to determine a user vector in a stepwise fashion?

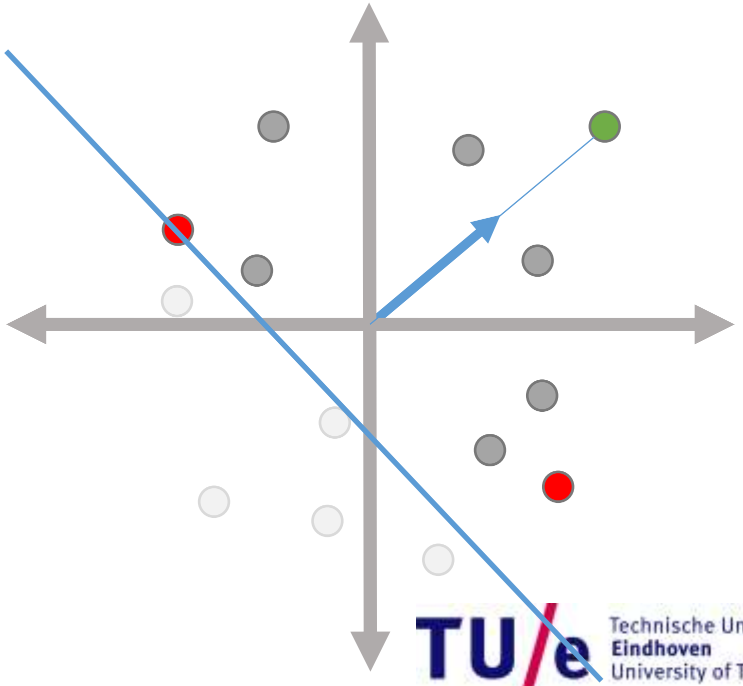
- Users make 10 successive choices out of sets of 10 movies.
- Choice set is adaptively calculated from a matrix factorization model
- Each choice is used to update the user vector and discard the least relevant items.

How does this work? Step 1

Iteration 1a: Diversified choice set is calculated from a matrix factorization model (red items)

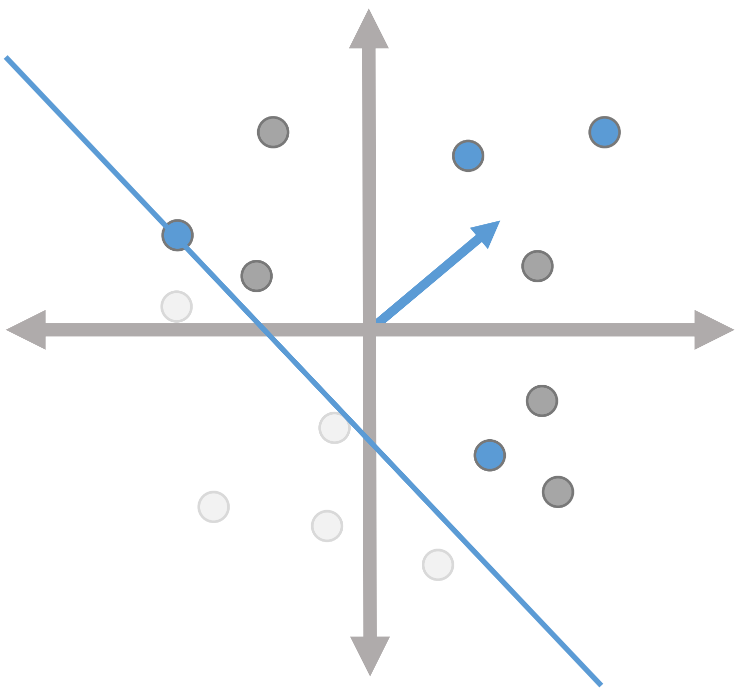
Iteration 1b: User vector (blue arrow) is moved towards chosen item (green item), items with lowest predicted rating are discarded (greyed out items)



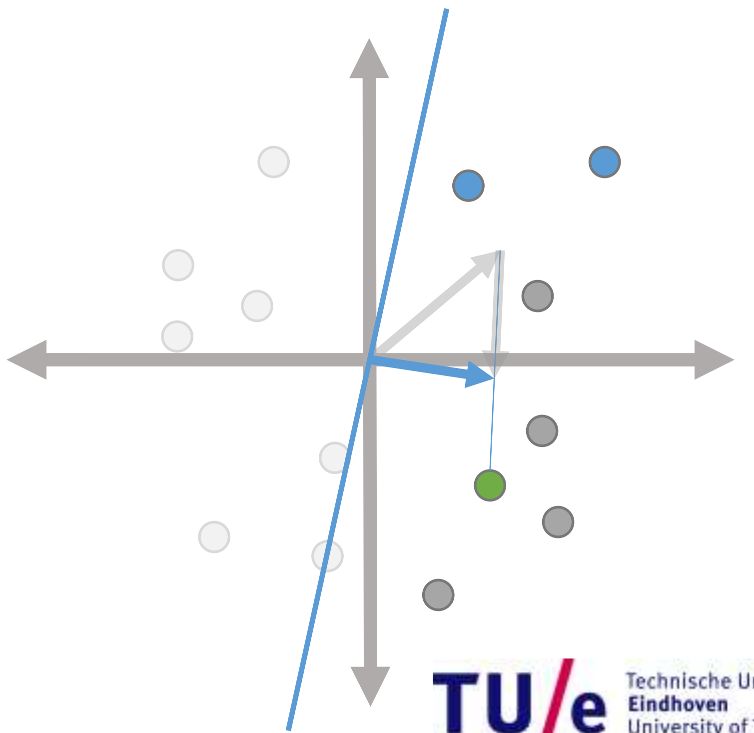


How does this work? Step 2

Iteration 2: New diversified choice set (blue items)



End of Iteration 2: with updated vector and more items discarded based on second choice (green item)



User study

103 users compared and evaluated **choice-based PE** and standard **rating-based PE** in a user-centric study.

We evaluate the interaction (Q1), the perception (Q2, cf. Ekstrand et al. 2014) and the recommendation lists (Q3)

1. Choice-based PE and Evaluation (Q1)
2. Rating-based PE and Evaluation (Q1)

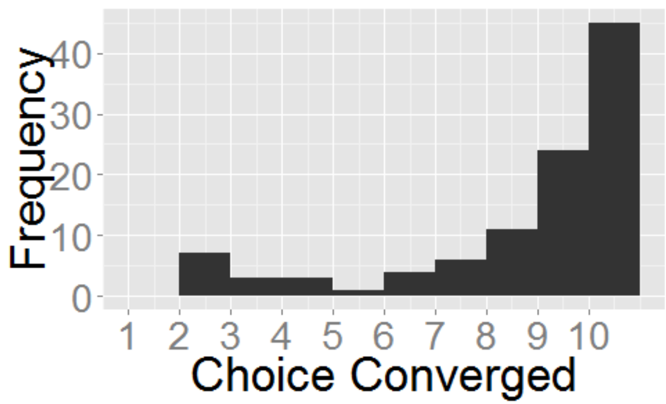
counter-balanced

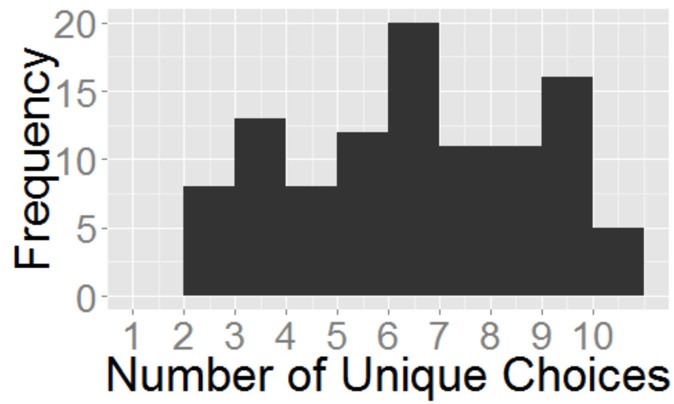
3. Calculation of Recommendations for both tasks
4. Recommendation Lists Side-By-Side Comparison (Q2)
5. Choice Based Recommendation List Evaluation (Q3)
6. Rating-Based Recommendation List Evaluation (Q3)

counter-balanced

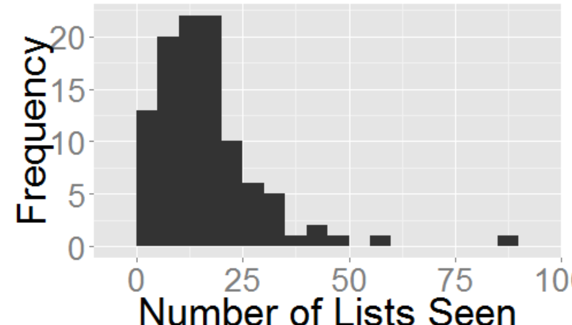
Behavioral data of PE-tasks

Choice-based PE: most users find their perfect item around the 8th / 9th item and they inspect quite some unique items along the way





Rating-based: user inspect many lists (Median = 13), suggesting high effort in rating task.



Q1 – Evaluation of Preference Elicitation

Choice-based PE: choosing 10 times from 10 items
Rating-based PE: rating 15 items

After each PE method they evaluated the interface on

interaction usability in terms of ease of use

e.g., “It was easy to let the system know my preferences”

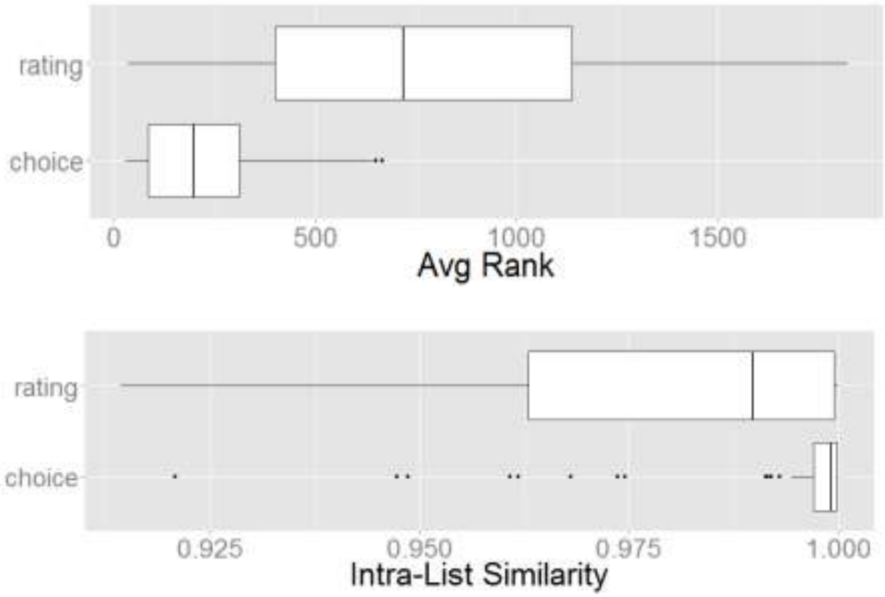
Effort: e.g., “Using the interface was effortful.”

effort and usability are highly related (r=0.62)

Results: less perceived effort for choice-based PE
perceived effort goes down with completion time

Objective measures

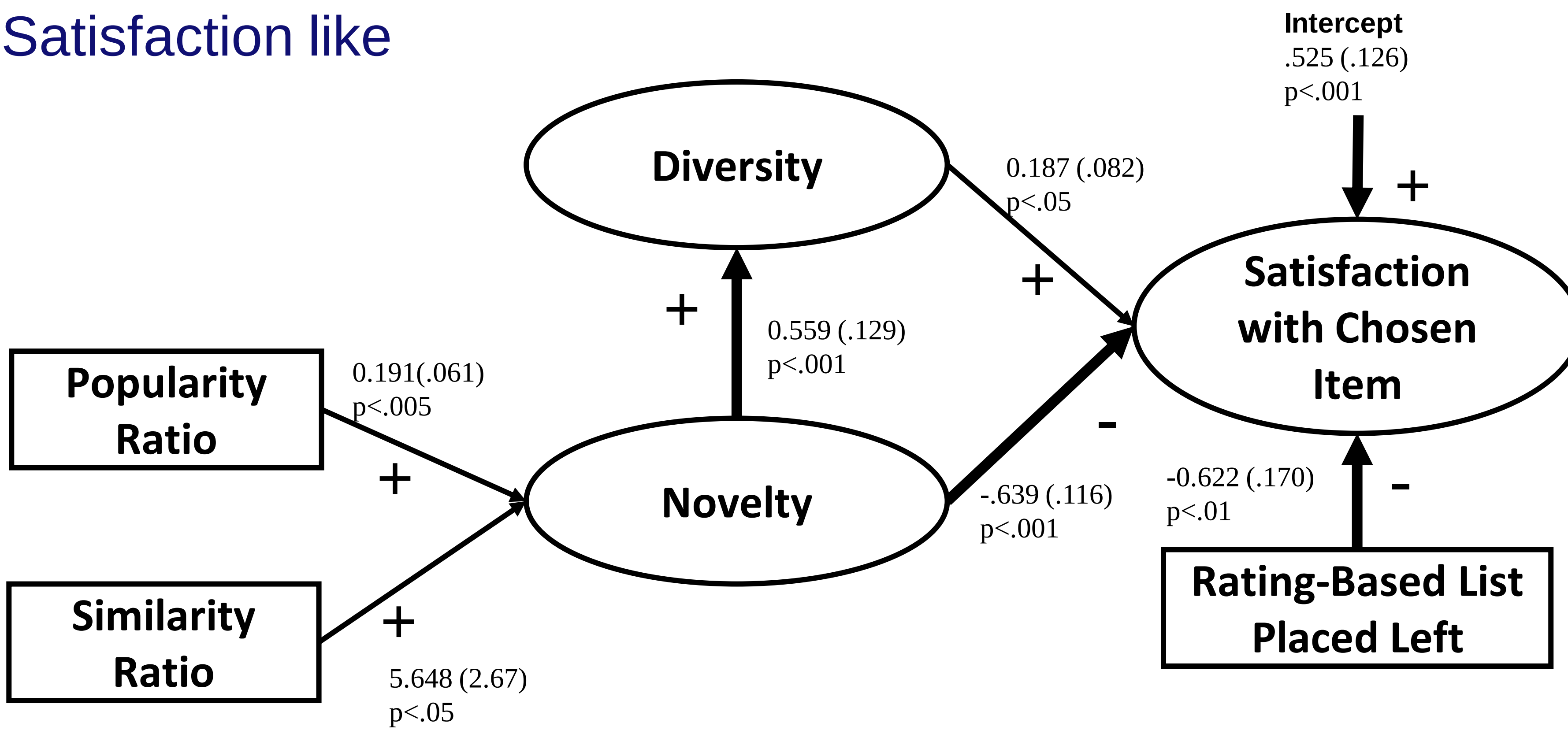
Recommendations coming from choice-based PE contain more popular and more similar items than from the rating-based PR



Technische Universiteit Eindhoven University of Technology

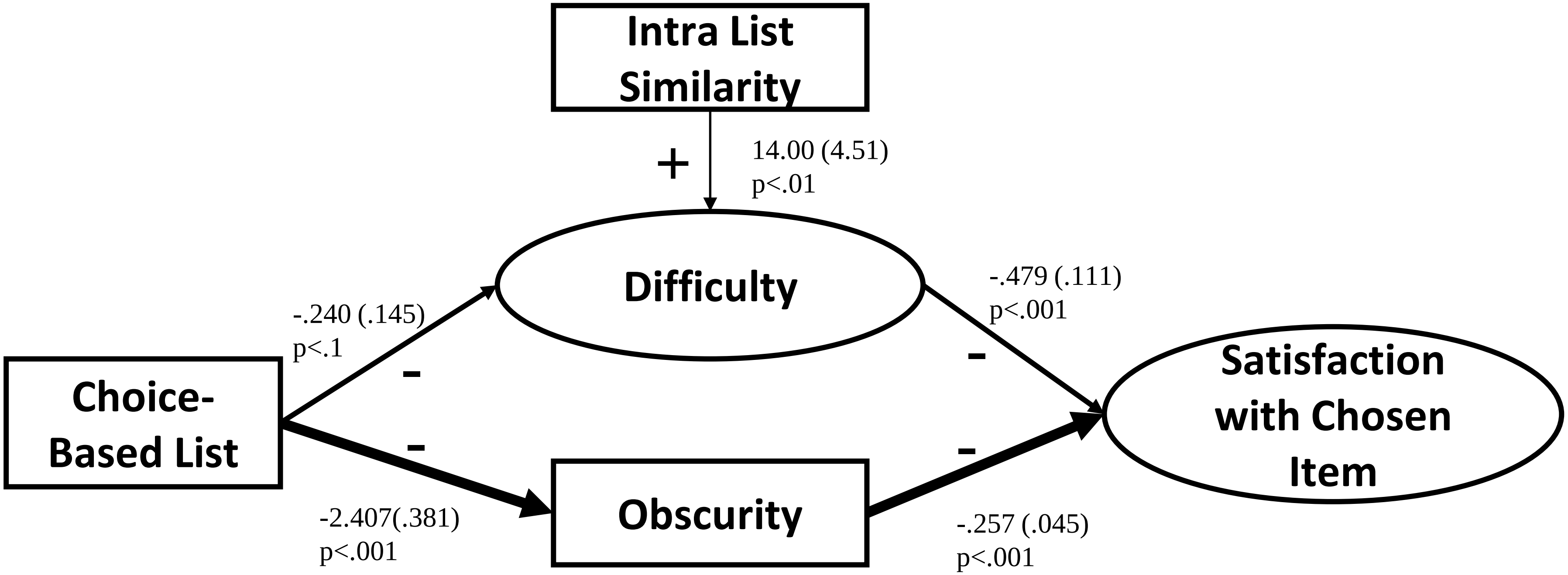
Q2 – Comparison of Recommendation Lists

side-by-side comparison on Diversity, Novelty and Satisfaction like



Q3 – Perception of Recommendation List

Participants evaluated the recommendation lists separately on Choice Difficulty and Choice Satisfaction



Conclusion

Participants experienced reduced effort and increased satisfaction for choice-based PE over rating-based PE

- relative (choice) rather than absolute (rating) PE could alleviate the cold-start problem for new users

Further research needed:

- the parameterization of the choice task
- strong effect of choice on the popularity of the resulting list
- novelty effects might have played a role

Task might help to adapt recommendations to the specific context a user is in!

What you should take away…

- Psychological theory can inform new ways of diversifying algorithm output or eliciting preferences
  - But we can reverse the argument: working with recommenders and algorithms we could enhance psychological theory
- User-centric evaluation helps to assess the effectiveness
  - Lot of work…
  - Linking subjective to objective measures might help future studies that cannot do user studies
- User-centric framework allows us to understand WHY particular approaches work or not
  - Concept of mediation: user perception helps understanding..