User Experiments in Recommender Systems
Introduction
Welcome everyone!
Introduction

Bart Knijnenburg

- Current: UC Irvine
  - Informatics
  - PhD candidate

- TU Eindhoven
  - Human Technology Interaction
  - Researcher & Teacher
  - Master Student

- Carnegie Mellon University
  - Human-Computer Interaction
  - Master Student
Introduction

Bart Knijnenburg

- First user-centric evaluation framework
  - UMUAI, 2012
- Founder of UCERSTI
  - workshop on user-centric evaluation of recommender systems and their interfaces
- Statistics expert
  - Conference + journal reviewer
  - Research methods teacher
  - SEM advisor
Introduction

Michael Ekstrand
@elehack

Bizarre, disturbing dream last night. @usabart was going to give a talk, which I missed. Then the hotel turned into a prison camp.
Introduction

“What is a user experiment?”

“A user experiment is a scientific method to investigate how and why system aspects influence the users’ experience and behavior.”
My goal:

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

How? User experiments!

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”
Introduction
Welcome everyone!

Evaluation framework
A theoretical foundation for user-centric evaluation

Hypotheses
What do I want to find out?

Participants
Population and sampling

Testing A vs. B
Experimental manipulations

Measurement
Measuring subjective valuations

Analysis
Statistical evaluation of the results
Evaluation framework
A theoretical foundation for user-centric evaluation
Framework

Offline evaluations may not give the same outcome as online evaluations

Cosley et al., 2002; McNee et al., 2002

Solution: Test with real users
Framework

System
algorithm

Interaction
rating
Higher accuracy does not always mean higher satisfaction

McNee et al., 2006

Solution: Consider other behaviors
Framework

System
algorithm

Interaction
- rating
- consumption
- retention
The algorithm counts for only 5% of the relevance of a recommender system

Francisco Martin - RecSys 2009 keynote

Solution: test those other aspects
Framework

System
- algorithm
- interaction
- presentation

Interaction
- rating
- consumption
- retention
Framework

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

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

Behavior is hard to interpret

Relationship between behavior and satisfaction is not always trivial

User experience is a better predictor of long-term retention

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

Questionnaire data is more robust

Fewer participants needed
Framework

Measure **subjective valuations** with questionnaires
Perception and experience

**Triangulate** these data with behavior
Ground subjective valuations in observable actions
Explain observable actions with subjective valuations

Measure **every step** in your theory
Create a chain of mediating variables
Personal and situational characteristics may have an important impact

Adomavicius et al., 2005; Knijnenburg et al., 2012

Solution: measure those as well
Objective System Aspects (OSA)

These are manipulations
- visual / interaction design
- recommender algorithm
- presentation of recommendations
- additional features

Situational Characteristics
- routine
- system trust
- choice goal

Personal Characteristics
- gender
- privacy
- expertise
Framework

User Experience (EXP)

Different aspects may influence different things

- interface -> system evaluations
- preference elicitation method -> choice process
- algorithm -> quality of the final choice
Framework

Situational Characteristics
- routine
- system trust
- choice goal

Subjective System Aspects (SSA)
- Link OSA to EXP (mediation)
- Increase the robustness of the effects of OSAs on EXP
- How and why OSAs affect EXP

System
- algorithm
- interaction
- presentation

Perception
- usability
- quality
- appeal

Experience
- consumption
- retention

Interaction
- rating

Personal Characteristics
- gender
- privacy
- expertise
## Framework

### Situational Characteristics

<table>
<thead>
<tr>
<th>Routine</th>
<th>System Trust</th>
<th>Choice Goal</th>
</tr>
</thead>
</table>

### Personal and Situational characteristics (PC and SC)

Effect of specific user and task

Beyond the influence of the system

Here used for evaluation, not for augmenting algorithms

### Personal Characteristics

<table>
<thead>
<tr>
<th>Gender</th>
<th>Privacy</th>
<th>Expertise</th>
</tr>
</thead>
</table>

---

**System**
- Algorithm
- Interaction
- Presentation

**Perception**
- Usability
- Quality
- Appeal

**Experience**
- System
- Process
- Outcome

**Interaction**
- Rating
- Consumption
- Retention

---

**Framework**

- [Diagram]

**Effect of specific user and task**

Beyond the influence of the system

Here used for evaluation, not for augmenting algorithms
Framework

Interaction (INT)

Observable behavior
- browsing, viewing, log-ins

Final step of evaluation

Grounds EXP in “real” data

Situational Characteristics

<table>
<thead>
<tr>
<th>routine</th>
<th>system trust</th>
<th>choice goal</th>
</tr>
</thead>
</table>

Personal Characteristics

<table>
<thead>
<tr>
<th>gender</th>
<th>privacy</th>
<th>expertise</th>
</tr>
</thead>
</table>

Experience

<table>
<thead>
<tr>
<th>system</th>
<th>process</th>
<th>outcome</th>
</tr>
</thead>
</table>

Interaction

<table>
<thead>
<tr>
<th>rating</th>
<th>consumption</th>
<th>retention</th>
</tr>
</thead>
</table>

Framework

Interaction (INT)

Observable behavior
- browsing, viewing, log-ins

Final step of evaluation

Grounds EXP in “real” data
Framework

Has suggestions for **measurement** scales
  e.g. How can we measure something like “satisfaction”?

Provides a good starting point for **causal relations**
  e.g. How and why do certain system aspects influence the user experience?

Useful for **integrating** existing work
  e.g. How do recommendation list length, diversity and presentation each have an influence on user experience?
Framework

“Statisticians, like artists, have the bad habit of falling in love with their models.”

George Box
Hypotheses

What do I want to find out?
Can you test if my recommender system is good?
Hypotheses

What does good mean?
- Recommendation accuracy?
- Recommendation quality?
- System usability?
- System satisfaction?

We need to define measures
Hypotheses

“Can you test if the user interface of my recommender system scores high on this usability scale?”
Hypotheses

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
Hypotheses

“Can you test if the UI of my recommender system scores high on this usability scale compared to this other system?”
$h_0$ Hypotheses

My new travel recommender

Travelocity
Hypotheses

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
Hypotheses

My new travel recommender

Previous version
(too many options)
**Hypotheses**

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

For industry, this may suggest further improvements to the system

For research, this makes the work generalizable

Measure *mediating variables*

Measure understandability (and a number of other concepts) as well

Find out how they mediate the effect on usability
An example:
We compared three recommender systems
  Three different algorithms
Ceteris paribus!
Hypotheses

The mediating variables show the entire story

Knijnenburg et al.: “Explaining the user experience of recommender systems”, UMUAI 2012
Hypotheses

Matrix Factorization recommender with explicit feedback (MF-E) (versus generally most popular; GMP)

Matrix Factorization recommender with implicit feedback (MF-I) (versus most popular; GMP)

perceived recommendation
gerated effectiveness

perceived recommendation

perceived recommendation

Hypotheses

Knijnenburg et al.: “Explaining the user experience of recommender systems”, UMUAI 2012
Hypotheses

“An approximate answer to the right problem is worth a good deal more than an exact answer to an approximate problem.”

John Tukey
Participants
Population and sampling
Participants

“We are testing our recommender system on our colleagues/students.”

-or-

“We posted the study link on Facebook/Twitter.”
Participants

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

“We only use data from frequent users.”
Participants

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

“We tested our system with 10 users.”
Participants

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

<table>
<thead>
<tr>
<th>Anticipated effect size</th>
<th>Needed sample size</th>
</tr>
</thead>
<tbody>
<tr>
<td>small</td>
<td>385</td>
</tr>
<tr>
<td>medium</td>
<td>54</td>
</tr>
<tr>
<td>large</td>
<td>25</td>
</tr>
</tbody>
</table>
Testing A vs. B
Experimental manipulations
“Are our users more satisfied if our news recommender shows only recent items?”
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”
Manipulations

“The first 40 participants will get the baseline, the next 40 will get the treatment.”
AB Manipulations

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
Manipulations

Between-subjects design:

Randomly assign half the participants to A, half to B

Realistic interaction
Manipulation hidden from user
Many participants needed

100 participants

A

B
AB Manipulations

Within-subjects design:

Give participants A first, then B

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

50 participants
Manipulations

Within-subjects design:

Show participants A and B simultaneously

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

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
**AB Manipulations**

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

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

<table>
<thead>
<tr>
<th>5 items</th>
<th>Low diversity</th>
<th>High diversity</th>
</tr>
</thead>
<tbody>
<tr>
<td>5+low</td>
<td>5+high</td>
<td></td>
</tr>
<tr>
<td>10 items</td>
<td>10+low</td>
<td>10+high</td>
</tr>
<tr>
<td>20 items</td>
<td>20+low</td>
<td>20+high</td>
</tr>
</tbody>
</table>
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)
Manipulations

“We were demonstrating our new recommender to a client. They were amazed by how well it predicted their preferences!”

“Later we found out that we forgot to activate the algorithm: the system was giving completely random recommendations.”

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

Does the question mean the same to everyone?
- John likes the system because it is convenient
- Mary likes the system because it is easy to use
- Dave likes it because the recommendations are good

A single question is not enough to establish content validity

We need a multi-item measurement scale
Perceived system effectiveness:

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

Bad: “The recommendations were not very novel”
Good: “The recommendations felt outdated”
Choose simple over specialized words

Participants may have no idea they are using a “recommender system”

Avoid double-barreled questions

Bad: “The recommendations were relevant and fun”
“We asked users ten 5-point scale questions and summed the answers.”
Measurement

Is the scale really measuring a single thing?
- 5 items measure satisfaction, the other 5 convenience
- The items are not related enough to make a reliable scale

Are two scales really measuring different things?
- They are so closely related that they actually measure the same thing

We need to establish convergent and discriminant validity
This makes sure the scales are unidimensional
Measurement

Solution: *factor analysis*

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

Benefits of factor analysis:

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

Perceived recommendation:
- Variety
- Quality

Choice:
- Difficulty

Movie expertise:
- Exp1
- Exp2
- Exp3

Satisfaction:
- Sat1
- Sat2
- Sat3
- Sat4
- Sat5
- Sat6
- Sat7
Measurement

low communality, high residual with qual2

loads on quality, variety, and satisfaction low communality

low communality

low communality

high residual with var1

var1 var2 var3 var4 var5 var6 qual1 qual2 qual3 qual4 diff1 diff2 diff3 diff4 diff5

perceived recommendation

variety

perceived recommendation

quality

choice
difficulty

movie

expertise

choice

satisfaction

exp1 exp2 exp3 sat1 sat2 sat3 sat4 sat5 sat6 sat7
Measurement

**Perceived Recommendation**
- **Variety**: AVE: 0.622, sqrt(AVE) = 0.789, largest corr.: 0.491

**Perceived Recommendation**
- **Quality**: AVE: 0.756, sqrt(AVE) = 0.870, largest corr.: 0.709

**Choice**
- **Difficulty**: AVE: 0.435 (!), sqrt(AVE) = 0.659, largest corr.: -0.438

**Movie Expertise**
- AVE: 0.793, sqrt(AVE) = 0.891, highest corr.: 0.234

**Choice**
- **Satisfaction**: AVE: 0.655, sqrt(AVE) = 0.809, highest corr.: 0.709
Measurement

“Great! Can I learn how to do this myself?”

Check the video tutorials at www.statmodel.com
Analysis

Statistical evaluation of the results
Manipulation -> perception:
Do these two algorithms lead to a different level of perceived quality?

T-test
Perception -> experience:
Does perceived quality influence system effectiveness?

Linear regression

System effectiveness

Recommendation quality
Analysis

Two manipulations -> perception:

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

Factorial ANOVA

Perceived quality

- low diversification
- high diversification

5 items 10 items 20 items

Willemsen et al.: “Not just more of the same”, submitted to TiiS
Only one method: **structural equation modeling**

The statistical method of the 21st century

Combines **factor analysis** and **path models**

- Turn items into factors
- Test causal relations
Analysis

Very simple reporting

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

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
Analysis

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

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

movie expertise + perceived recommendation

Top-20 vs Top-5 recommendations

Lin-20 vs Top-5 recommendations

perceived recommendation:

variety +

quality +

choice satisfaction +

choice difficulty +

Difficult trade-off (not shown here)

Simple regression

“Full mediation” +

“Inconsistent mediation” +

Measured by var1-var6 Additional effect

(not shown here)
Analysis

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

“Great! Can I learn how to do this myself?”

Check the video tutorials at www.statmodel.com
**Analysis**

**Homoscedasticity** is a prerequisite for linear stats

Not true for count data, time, etc.

Outcomes need to be **unbounded** and **continuous**

Not true for binary answers, counts, etc.
Analysis

Use the correct methods for **non-normal data**

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

**Don’t use “distribution-free” stats**

They were invented when calculations were done by hand

**Do use structural equation modeling**

MPlus can do all these things “automatically”
Analysis

Standard regression requires **uncorrelated errors**

Not true for repeated measures, e.g. “rate these 5 items”

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

Golden rule: data-points should be **independent**
Two ways to account for repeated measures:
- Define a random intercept for each user
- Impose an error covariance structure

Again, in MPlus this is easy
A **manipulation** only causes things

For all other variables:
- Common sense
- Psych literature
- Evaluation frameworks

Example: privacy study

*Knijnenburg & Kobsa.: “Making Decisions about Privacy”*
Analysis

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

George Box
Introduction
User experiments: user-centric evaluation of recommender systems

Evaluation framework
Use our framework as a starting point for user experiments

Hypotheses
Construct a measurable theory behind the expected effect

Participants
Select a large enough sample from your target population

Testing A vs. B
Assign users to different versions of a system aspect, ceteris paribus

Measurement
Use factor analysis and follow the principles for good questionnaires

Analysis
Use structural equation models to test causal models
“It is the mark of a truly intelligent person to be moved by statistics.”

George Bernard Shaw
Resources

User-centric evaluation


Choice overload

Willemsen, M.C., Graus, M.P., Knijnenburg, B.P., Bollen, D.: Not just more of the same: Preventing Choice Overload in Recommender Systems by Offering Small Diversified Sets. Submitted to TiiS.

Resources

Preference elicitation methods


Social recommenders

Resources

User feedback and privacy


Statistics books

Agresti, A.: An Introduction to Categorical Data Analysis. 2nd ed. 2007.

Online tutorial videos at www.statmodel.com
Resources

Questions? Suggestions? Collaboration proposals?
Contact me!

Contact info
E: bart.k@uci.edu
W: www.usabart.nl
T: @usabart