Conducting User Experiments in Recommender Systems

Bart Knijnenburg and Edward Malthouse

POPROX NEWS

POPROX NEWS

Outline

1. What is experimentation and why do it?
2. Steps for running an experiment
3. Fundamental designs
4. Measuring outcomes
5. User-centric measures

POPROX NEWS

1. What is experimentation and why do it?

1. What is experimentation and why do it?

Experimentation—The **manipulation** of one or more **independent variables** (X) by the experimenter in such a way that their effects on one or more **dependent variables** (Y) can be measured.

- RecSys puts experimentation under **user studies**
  - Other types of user studies: observations, interviews, focus groups, surveys
- Why experiment?
  - Long considered the gold standard for establishing causality
  - Allows experimenter to study users when they interact with a system
  - Experimenters can ask users questions, crucial for interpreting results
  - Quantitatively test theories/hypotheses, explore possible causes, optimize processes
- Why not do it?
  - Running experiments can be expensive and difficult, but now you have POPROX!

Bing search example

In 2012 a Bing employee suggested **lengthening the title line** of search ads by combining it with the text from the first line below the title (Kohavi et al. 2020)

Is this a good idea?

Bing experiments showed **12% increase** in clicks, or **$100M increase** annually

POPROX NEWS

2. Steps for running an experiment

Steps for running an experiment

1. Establish hypothesis/hypotheses
2. Select response variable(s), factor(s) and levels
3. Choose an experimental design
4. Determine sample size
5. Select experimental “units” and randomly assign them to treatments
6. Conduct the experiment
7. Analyze the data and draw conclusions

Establish hypothesis (or hypotheses)

Hypothesis: clear, testable statement of how the causal variable affects the specific outcome(s), e.g.,

> Displaying search results with lengthened titles will generate more revenue than the baseline display

Directional statement of cause and effect without reference to research design considerations (e.g., operationalization of constructs, statistical significance)

Classical approach is to have a **rationale/theory** to support it: why should the hypothesis be true?

Factor(s), levels, and response variable(s)

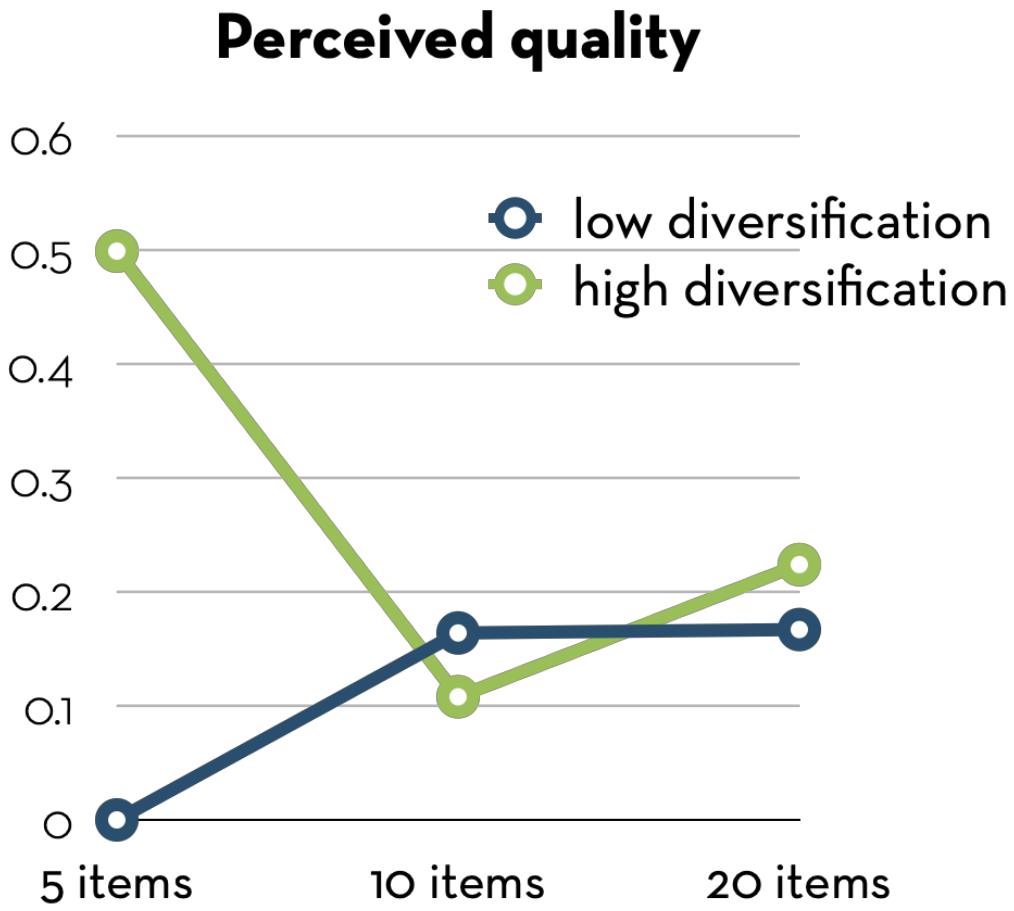
- An independent variable (X) is called a **factor** and its distinct values in the design are called **levels**
  Factor is title display and levels are “lengthened” and “baseline”
- When there’s one factor the level is the **treatment**. When there’s more than one factor then a combination of levels is the treatment.
- Simplest case: one factor with two levels (aka A/B testing)
  - **Control group (CG):** users who receive baseline treatment
  - **Treatment group (TG):** users who receive improved treatment
- **Response variables** (Y) measure the consequence of the treatments (more on this from Bart). For Bing example, revenue

On picking factors

- Factors are aspects of the system that you want to test:
  - The mechanism producing the recommendations (algorithm, objective function, input data, …)
  - The presentation of the recommendations (number, layout, format…)
  - Other system aspects (preference elicitation mechanisms, explanations, visualizations, …)
- TIP: Study one factor at a time, unless you expect factors to interact
  - **Interaction:** the effect of a factor depends on the level of another factor

On picking factors - example

- Two manipulations at the same time:
  - What is the combined effect of list diversity and list length on perceived recommendation quality?
- Test for the interaction effect!



On picking levels

- Always answer “in comparison to what?” by having a baseline control condition or other realistic alternatives
- Stronger manipulations (i.e., large differences in levels) will tend to show differences in the outcome but the levels should
  - be realistic and practical (explainable RS example)
  - cover a range of interest
- The more levels you have, the more you fragment your sample (or extend study length in a within-subject design)

On picking levels - example

What's the effect of showing "recent news"?

- Proposed system for treatment group:
  - Filter out any items > 1 month old
- What should my control group see?
  - 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"
- Related question: how long should the study run?

POPROX NEWS

3. Fundamental designs

Choose an experimental design

After-only with control:

TG: (R)	X	O1
CG: (R)		O2

- More than two groups possible
- Large between-user variation implies that larger samples will be required to detect difference compared with before-after with control design

Before-after with control:

TG: (R)	O1	X	O2
CG: (R)	O3		O4

- O1 and O3 are **pre-measures**, which estimate user effects
- Simplest example of a **repeated-measures** design

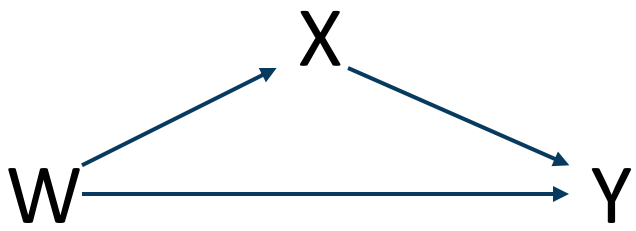
Fundamental designs: Crossover

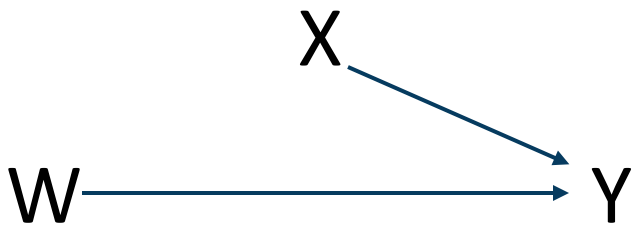
(R)	X1	O11	X2	O12	…
(R)	X2	O21	X1	O22	…

- Also called **within-subject design:** each user receives each treatment
- You can estimate user and treatment effects, reducing sample size requirements (with after-only design user effect is confounded with treatment and we rely on randomization). Users are their own controls
- This design has possible **carryover** effects, where receiving one treatment may affect the user's response to a subsequent treatment.
  - A possible solution is to give respondents the baseline condition for some time before switching treatments
- More than two treatments possible

Confounds and randomization

- **Confound:** Factors other than the experimental variable that affect the dependent variable
- A confounding variable (W) is especially problematic when it causes the dependent variable (Y) and the treatment (X)
- A major contribution (due to Fisher) is to **assign treatments at random**, so that treatments are not correlated with anything
- **Random assignment is key** (W then increases the error term)





Example: unexpected confounds

- Idea: run a news recommendation study, between subjects
  - Start control group on October 28, 2024
  - Start treatment group on November 11, 2024

- What's the problem here?

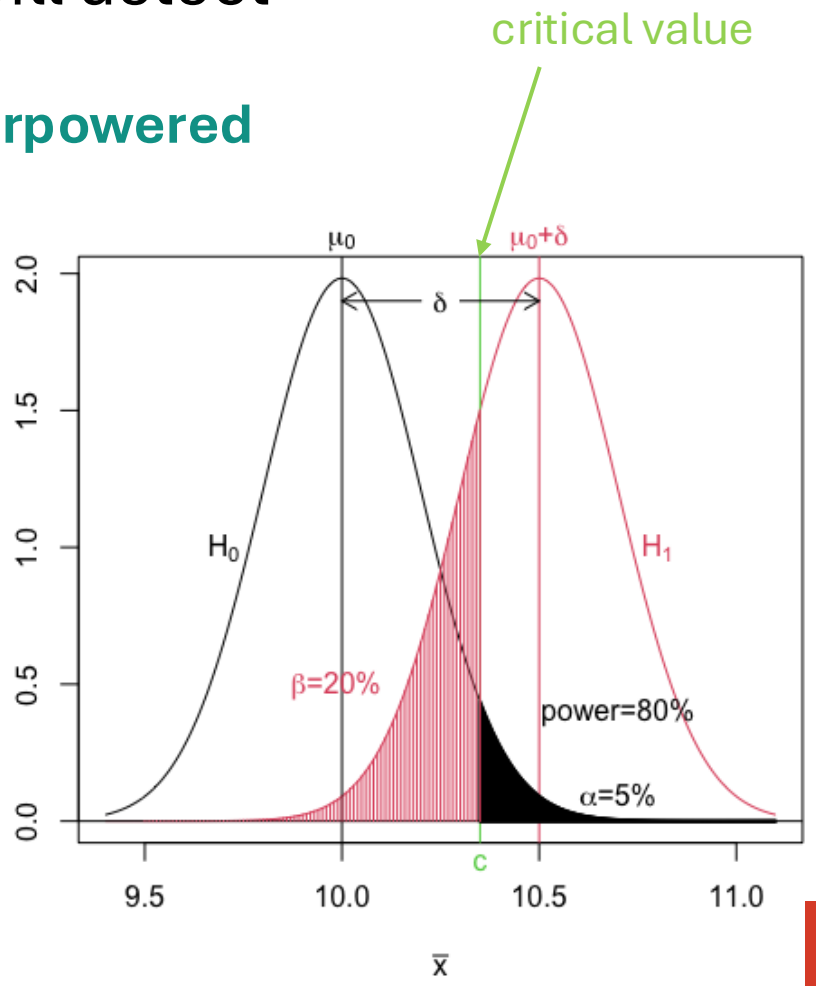
Fundamental designs: randomized block

- Prior to assignment, suppose experimenter has information about users that will correlate with outcome measures
- We can reduce errors, and thus sample size, through **blocking**
  - Partition users in to blocks based on other information
  - Randomly assign users within each block to a treatment
  - Estimate Y = block + treatment + error
- Basic design principle:

  “block what you can and randomize what you can’t”

Determine sample size

- **Power:** the probability that a significance test will detect an effect given an effect exists
  - A study with a sample sizes that is too small is **underpowered**
  - Overly large sample sizes waste precious resources
- See G*Power software
- You must specify
  - Minimum difference you want to detect (δ)
  - Whether H_1 is one- or two-sided
  - α, β and the population variance
- See Walpole, et al. *Probability and statistics for engineers and scientists*



Determine sample size - example

- Do married men weigh more than single men?
  - Find 4 married men: Nm = 4, Meanm = 182, SDm = 15
  - Find 4 single men: Ns = 4, Means = 170, SDs = 15
- Effect size: 12 lbs
  - Is this a large effect? —> Need to standardize it!
  - Cohen's d = (Meanm – Means)/pooled SD
  - (182–170)/15 = 0.8… this is indeed a large effect
- Is it significant? No! p = .301

- Small studies (N << 100) may find medium or large effects that are not significant
  - Waste of resources! (unless they are pilot studies)

Determine sample size - example

- Do married men weigh more than single men?
  - Find 4000 married men: Nm = 4000, Mm = 177.5, SDm = 15
  - Find 4000 single men: Ns = 4000, Ms = 176.5, SDs = 15
- Effect size: 1 lb
  - Is this a large effect?
  - (177.5–176.5)/15 = 0.067… this is a very small effect
- Is it significant? Yes! p = .0014

- Large studies (N >> 100) may find very small effects that are significant
  - Also a waste of resources! (could have done with way fewer participants)

Select units

- Inferences from your experiment apply to the **sampled population**—the group from which you have selected your users
  - If you test your algorithm on freshman CS majors from The University of X, the results apply to this group in this context
  - You can offer theoretical arguments for why inferences should apply more widely
- You may need to repeat the experiment on other populations and in other contexts to determine whether results generalize
  - If the results do not hold in some other context then you have discovered boundary conditions for your theory

Analyze results

- See undergraduate textbook on “statistics for scientists and engineers,” e.g.,.
  - Walpole, Myers and Myers
  - Montgomery and Runger
  - Tamhane and Dunlop
- See “Design and Analysis of Experiments” textbook, e.g.,
  - Montgomery
  - Tamhane
  - Box, Hunter, and Hunter (classic)
- Online resources:
  - www.usabart.nl/eval (introductory) and www.usabart.nl/eval2 (advanced)

Internal and external validity

- **Internal validity**—The ability of the experiment to unambiguously show a cause-and-effect relationship, i.e., to what extent can we attribute the effect that was observed to the experimental variable and not other (confounding) factors?
- **External validity**—The extent to which the results of the experiment can be generalized to other populations, contexts, and time periods, i.e., will we get similar results in other settings?
- There is often a **trade-off** between them
  - Highly controlled experiments may have strong IV but lack EV due to artificial ("lab") conditions.
  - Studies in natural settings may have high EV but weaker IV because of the difficulty in controlling all confounds
- See Calder, Phillips and Tybout (1982) for arguments on why IV should be emphasized in theoretical research

Some threats to internal validity

- **Selection:** treatment and control groups equal (WRT outcome variables). Remedy: randomization (and blocking) avoid this threat
- **Hawthorne (placebo) effect:** users respond differently because they know they are being treated. Remedies
  - Users should be **blind**, not knowing their treatment assignment
  - Control users should receive a **placebo** treatment
- **Experimental mortality:** Differential loss of respondents from different groups. Report difference in attrition

Some more threats to internal validity

- **History:** any variables or events, other than the one(s) manipulated by the experimenter, that occur between the pre- and post-measures and affect the dependent variable. Remedy: control group
- **Interactive testing effect:** When a pre-measure changes the user's sensitivity to the independent variable(s). Remedies don't take premeasure or use Solomon four-group design
- **Regression effect:** treatments assigned on (usually large value of) pre-measure, e.g., if you run multiple tests and select largest differences for rollout, then it's likely "you won't do as well in rollout as in test"

Running complex studies (POPROX promo)

The lack of research infrastructure often limits…

…the **complexity** of research studies

- Small samples from a limited population
- Short, ”single shot” studies; carryover and interactive test effects

…the **validity** of research studies

- Invited/compensated participation (no inherent motivation)
- Only immediate outcomes measurable

POPROX allows for longer-term studies with motivated users; with a built-in control group and default pre-/post-measures!

Some additional reading

- Maxwell, S. E., Delaney, H. D., & Kelley, K. (2017). Designing experiments and analyzing data: A model comparison perspective. Routledge. Psychology perspective, comprehensive
- Pearl, J., & Mackenzie, D. (2018). The book of why: the new science of cause and effect. CS, DOE history, causal models, mediation
- Kohavi, R., Tang, D., & Xu, Y. (2020). Trustworthy online controlled experiments: A practical guide to A/B testing. Cambridge University Press. CS, practical and relevant to POPROX
- Knijnenburg, B. P., & Willemsen, M. C. (2015). Evaluating recommender systems with user experiments. Recommender Systems Handbook, tailored to recommender systems
- Owen, A (2020). A First Course in Experimental Design, Stanford. Statistics, modern treatment of 100-year-old topic
- Online resource: www.usabart.nl/methods

POPROX NEWS

4. Measuring outcomes

Measuring outcomes

Ask yourself: What is the goal of your recommender system?

1. Item ranking performance?
2. Conversion / user retention?
3. User experience?
4. Users' goal fulfillment?
5. Positive societal outcomes?

- Offline tests can only do #1 (kind of)
- Online A/B tests add #2
- Most experiments add #3 (and a bit of #4)
- With POPROX we aim to support all 5 goals!

Traditional metrics

Justification: used extensively in the RecSys community, but confusion exists about the best way to calculate them

- Construct: Ranking
  - Normalized discounted cumulative gain (NDCG)
  - Mean reciprocal rank (MRR)
  - Other measures (ERR, RBP)
- Construct: Success and Conversion
  - Click-through rate (CTR)
  - Hit rate (fraction of recommended items with a click)

See [cite] for help in calculating them

Equity, fairness, and diversity

Justification: impact to society; help exposure users to a wider range of content

- Construct: Distribution Equity
  - Gini index
  - Expected exposure
  - User space coverage vs item coverage (related to fairness due to exposure measure)
- Construct: Diversity
  - Intra-list diversity (average pairwise distance of recommended items)
  - Inter-recommendation diversity (similar to item coverage, it is an overall system measure)
    - Aggregate intra-list diversity (Gini coefficient, Gini-Simpson's index, entropy)
- Construct: Fairness across items
  - Equity of amortized attention

Online behavior metrics

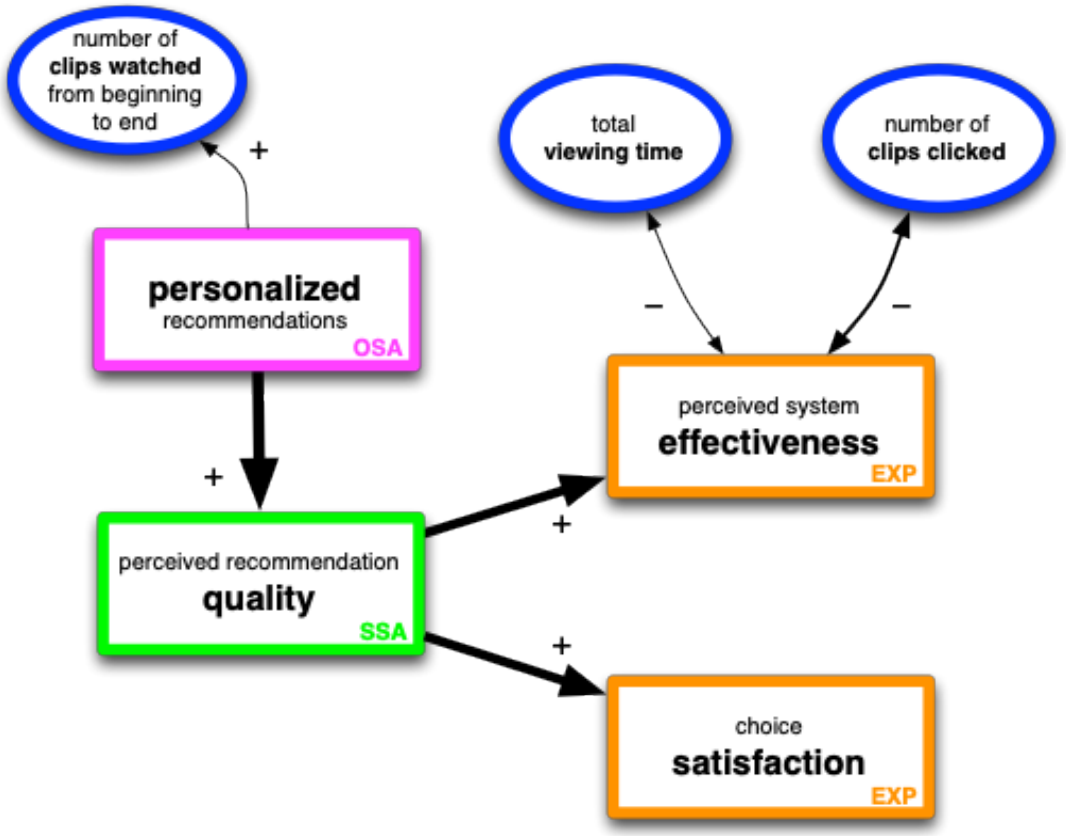
Justification: used extensively in (industry) field trials; “ground truth” user behaviors

- Construct: Engagement
  - Implicit item-based metrics: clicks, engagement time, social sharing
  - “System” metrics: dwell time, use frequency / consistency, bounce rate
- Construct: Interest
  - Explicit item-based feedback metrics, e.g. rating, thumbs up/down, like button
  - Referral / forwarding
- Construct: Retention
  - Time since last use

Why go subjective?

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

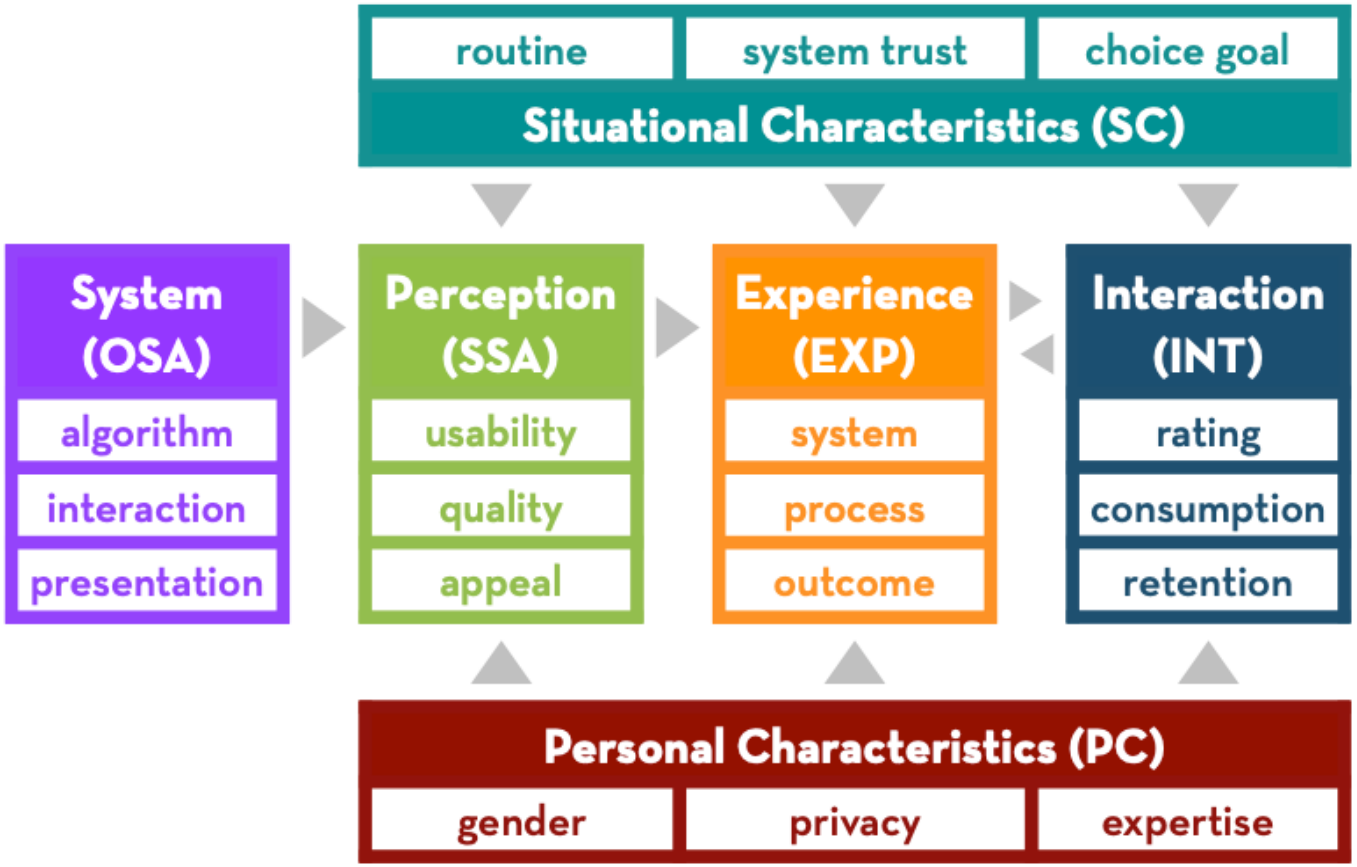
Asking for subjective evaluations explains why this happened



POPROX NEWS

5. User-centric metrics

User-centric metrics



Justification: widespread use in RecSys user experiments

- Measuring user perceptions (SSA) and experiences (EXP) to explain the effect of system manipulations (OSA) on user behavior (INT)
- Helps to produce robust, generalizable study results

See Knijnenburg, B. P., & Willemsen, M. C. (2015). Evaluating Recommender Systems with User Experiments

User-centric metrics

- Subjective System Aspects (SSA)
  - Constructs: Perceived recommendation quality, perceived recommendation diversity, understandability, perceived control, perceived use effort / ease of use
- User Experience (EXP)
  - Constructs: System satisfaction, perceived system effectiveness/usefulness, choice difficulty (usage satisfaction), choice satisfaction
- Metrics:
  - Each construct is a multi-item scale (usually 4-7 items) of statements rated on a 7-point agreement scale.
  - Best practice in psychometrics: creates a shared conceptual understanding between the participants and the researchers (and others!)
  - Multi-item scales allow for an evaluation of their validity and robustness (note: the ones listed above have been validated extensively in prior work)

Developing your own scales? See DeVellis (2011). Scale Development: Theory and Applications

Uses and gratifications metrics

Justification: used extensively in communication research; ideal for measuring the real-world impact of our system. Highly predictive of future use.

- Originally developed to explain media choice; measure how the recommendations contribute to a user's life goals:
  - Utilitarian tips and advice ("recommendations that you can use")
  - Curation and learning ("makes me smarter")
  - Social facilitation--promotes positive social contacts
  - Feel good, inspiration--being uplifted and motivated to be a better person
  - Surprise and serendipity--encountering something surprising or out of the ordinary
  - Expose me to different perspectives

Self-actualization metrics

Justification: based on a proposal to push the boundaries of the RecSys field

Use recommendation technology to support users in developing, exploring, and understanding their preferences based on their long-term goals and ambitions, using scales such as:

- Interest coverage (do the recommendations cover all my interests?)
- Interest clarification potential (does the system help me explore and understand my interests?)
- Interest development potential (does the system help me move beyond / develop my interests?)
- Perceived self-actualization (does the system help me meaningfully improve my life?)

See Knijnenburg et al. (2016). Recommender Systems for Self-Actualization

Covariates

Justification: measure things about the user to make sure that innovations have an equitable impact (or: show how certain manipulations affect only a subset of users)

- Demographics (mostly stable)
  - Age, gender, location, education, SES, political leaning, domain interests, etc.
- Personal characteristics (mostly stable)
  - Constructs: domain knowledge, choice maximization, need for cognition, privacy concerns, familiarity with recommenders
- Situational characteristics (may vary over time)
  - Constructs: trust (multi-faceted), mood, current goals (see uses and gratifications)

Another POPROX promo slide

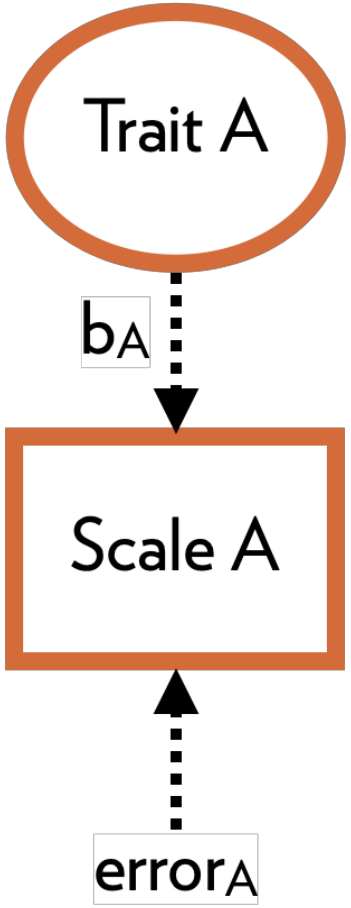
- POPROX measures all mentioned constructs with at least one item each, by default!
  - Added bonus: We have a history of baseline values for each metric for each user
- You can select scales to **expand** to multiple items
  - Can be helpful for robustness sake
- You can also **add your own** scales
  - Tailored to your study
- Above all: because these are **real users**, the goal fulfillment metrics have external validity

Multi-item measurement

- For subjective traits, single-item measurements tend to lack **content validity**
  - Each participant may interpret the item differently
  - This reduces precision and conceptual clarity
- Accurate measurement requires a **shared conceptual understanding** between all participants and researcher
- Solution: use a multi-item **scale**

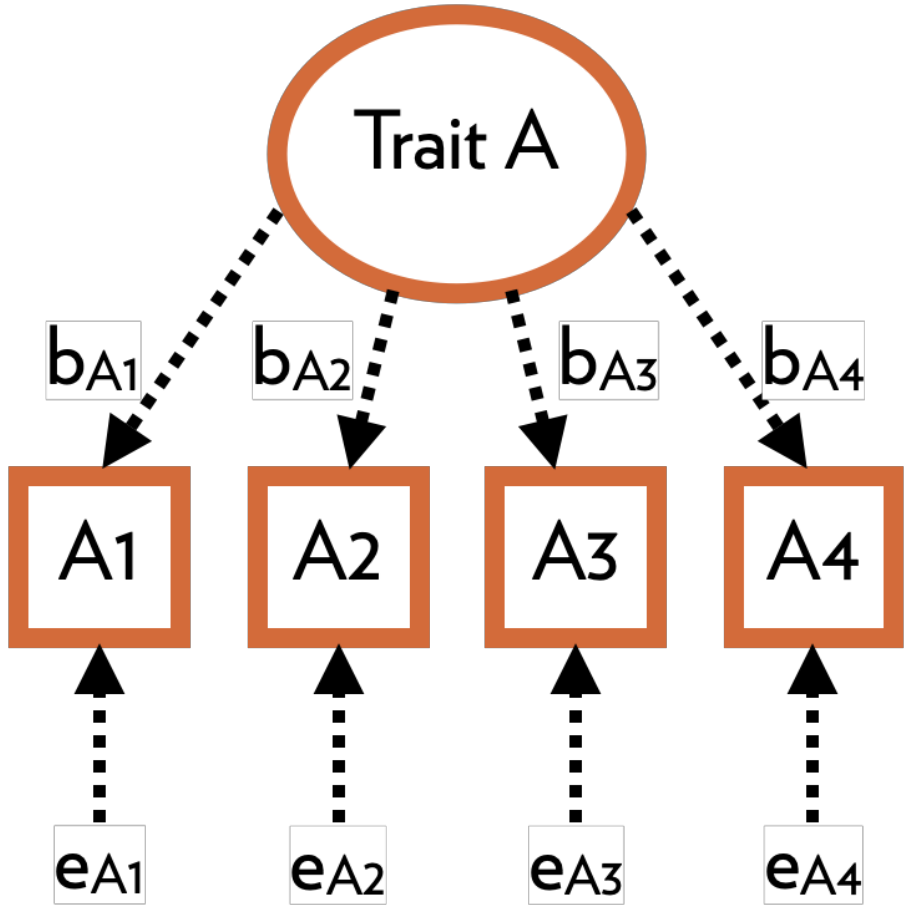
Multi-item measurement

- Even then, a scale is an **imperfect** way of measuring a subjective trait
  - Our real goal is to measure the trait, not the scale
- We can think of the traits as latent variables and the scales as observed variables
  - The trait causes my answers on the scale
- Like a regression with an unobserved X
  - Scale A = a + bATrait A + errorA
  - The R^2 of this regression determines how well we are measuring Trait A



Multi-item measurement

- How do we get this R2?
  - Trick: if you have multiple items, you can derive b's from the correlation between the items
- See example:
  - The b's are "loadings"
  - The e's are "uniqueness"
  - R^2 = 1-e is called "communality"
- Each item uses the others as a yardstick
  - Once a scale is validated, a single item may suffice*

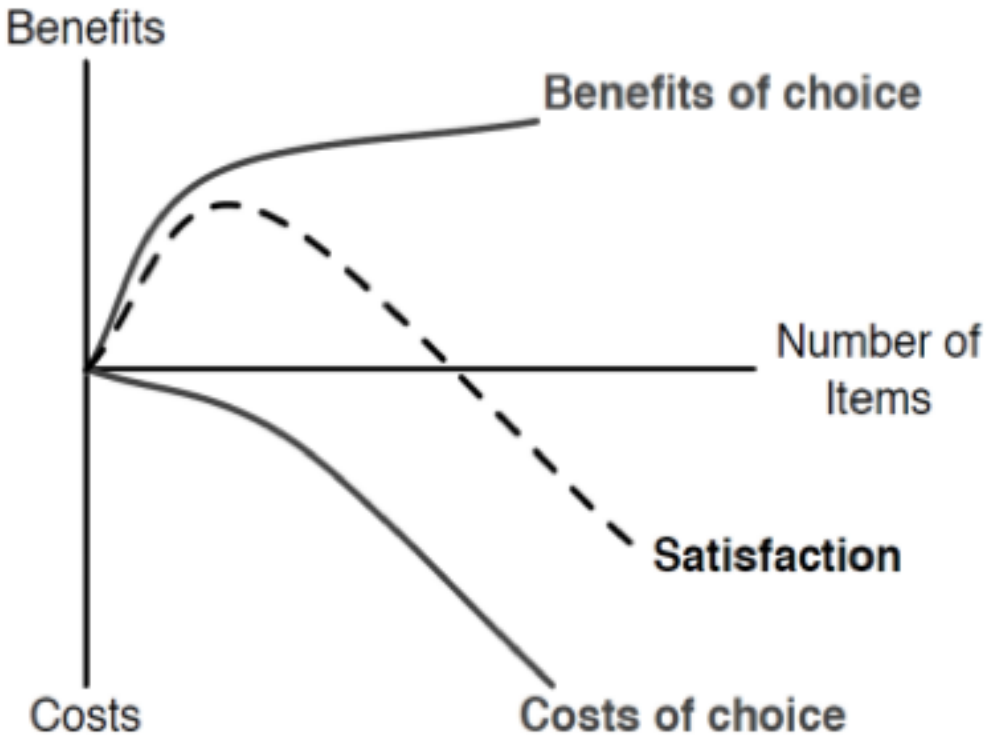


Applying theory

- To learn something from a study, we need a theory behind the effect
  - This makes the work generalizable
  - This may suggest future work
- Measure **mediating variables**
  - Find out how they mediate the effect on usability
- Evaluate the data using path modeling or structural equation modeling

Applying theory - example

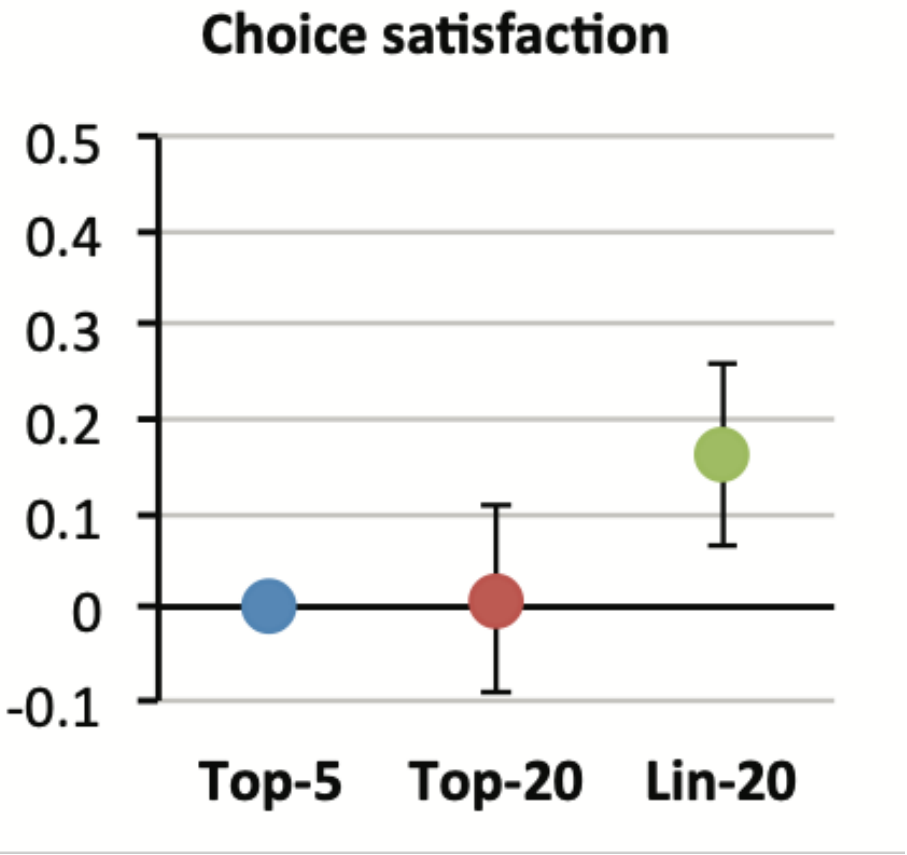
- **Choice overload:**
  Satisfaction = benefit – cost
  - Benefit of more options: easier to find the right option
  - Cost of more options: more comparisons, higher potential regret
- Is this also true for recommendations?



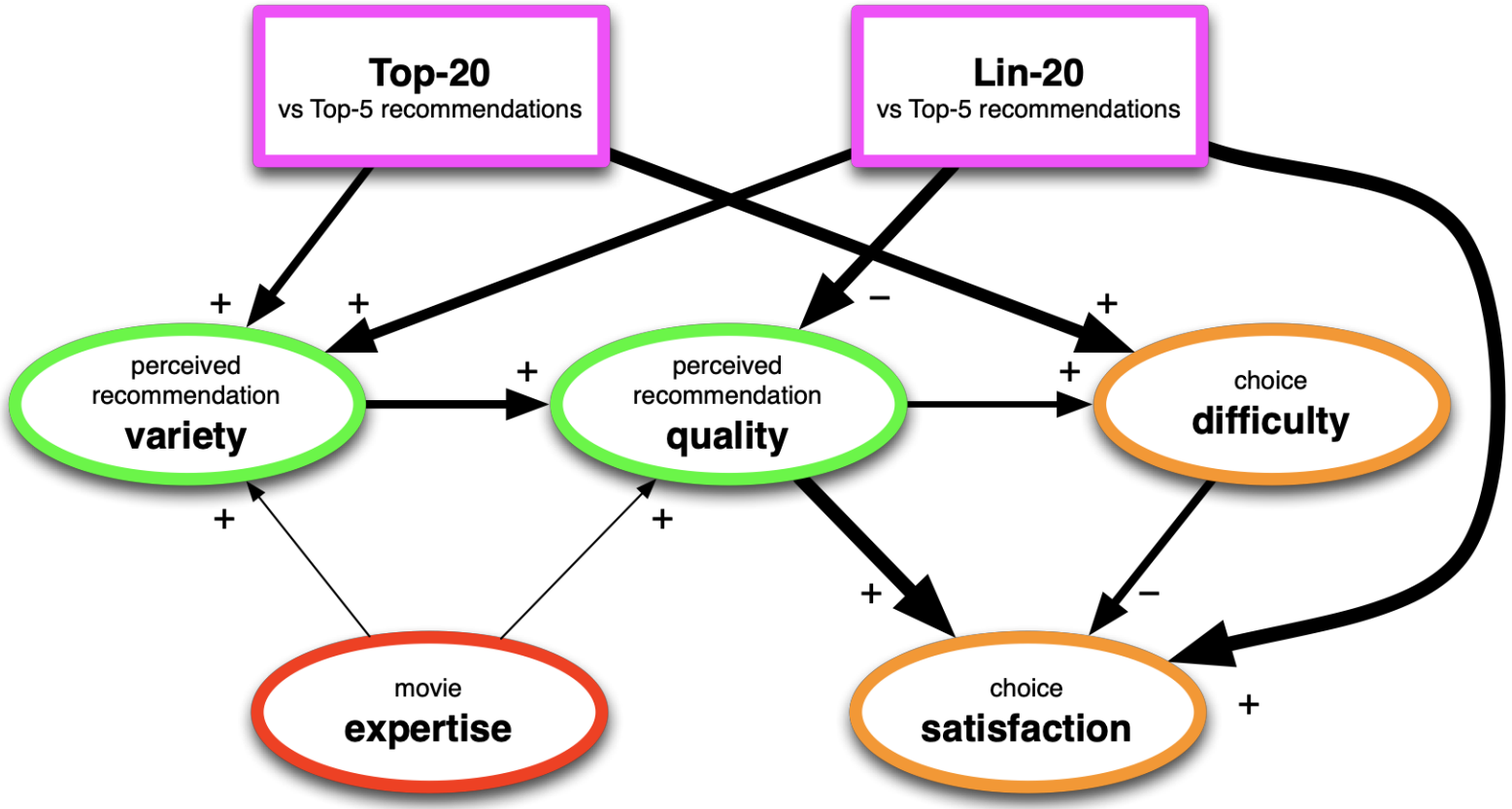
Applying theory - example

- Example from Bollen et al.: “Choice Overload”
  - What is the effect of the number of recommendations?
  - What about the composition of the recommendation list?
- Tested with 3 conditions:
  - Top 5:
    - recs: 1 2 3 4 5
  - Top 20:
    - recs: 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
  - Lin 20:
    - recs: 1 2 3 4 5 99 199 299 399 499 599 699 799 899 999 1099 1199 1299 1399 1499

Applying theory - example

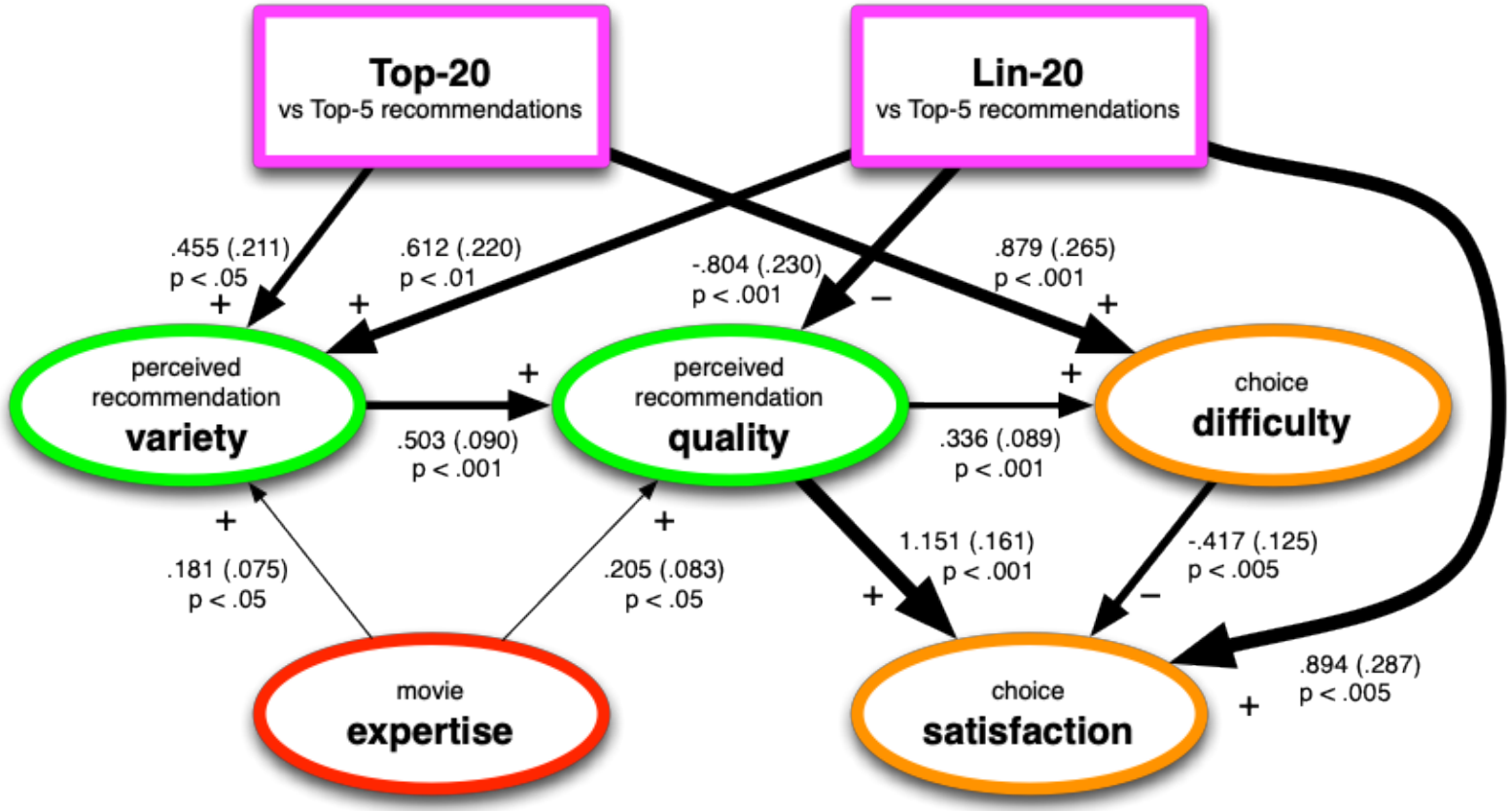


Applying theory - example



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

Applying theory - example



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