Experimental Design

Research Methods for Human-Centered Computing

Experimental Design

Today's goal:

Explain how to set up experiments

Outline:

- Choosing manipulations/conditions (the IVs of your study)
- Randomization and between- and within-subjects designs
- Examples from existing work

Proposal presentation

Presenting your proposal to class

Proposal presentation

Take any opportunity to present your work!

- Present finished work at conferences
- Present in-progress work in your lab
- Practice makes perfect!

Present your proposal to the rest of the class

- Carefully explain what you plan to do
- Get audience feedback on all aspects of your proposal

Proposal presentation

Prepare a presentation based on the research question your team chose to investigate

"Conference quality"

20 mins to give a concise overview of the question and the proposed study

10 mins to answer questions from the audience

Proposal presentation

Your presentation should cover:

- Summary of related work (why do we need to know the answer this research question?)
- Research question (What is the question you want to answer)
- Proposed research method (participant recruitment, experimental design, research prototype, etc.)
- Expected results and implications (What do these results mean? How do they relate to existing literature? What are the limitations of your study?)

Proposal presentation

You will be graded on:

- Your knowledge and coverage of the material
- Presentation style
- Clarity and organization of the presentation
- Use of presentation aides (e.g., slides, handouts)
- Ability to engage the class in thoughtful and productive discussion
- Timing (staying within your slot and leaving enough time for questions)

Proposal presentation

During the presentation:

Make sure everyone in your group contributes

Take turn answering and taking notes during the Q&A

After the presentation:

Upload a copy of your presentation and the Q&A notes to Canvas.

Manipulations

Testing A versus B

Manipulations

What should be the manipulations?

Choosing interesting versions (conditions) to test against each other

Remember ceteris paribus!

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

Any difference can be attributed to the manipulation

Manipulations

"Are our users more satisfied if our news recommender shows only recent items?"

Choosing a baseline

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"

Conditions

You can have more than two conditions!

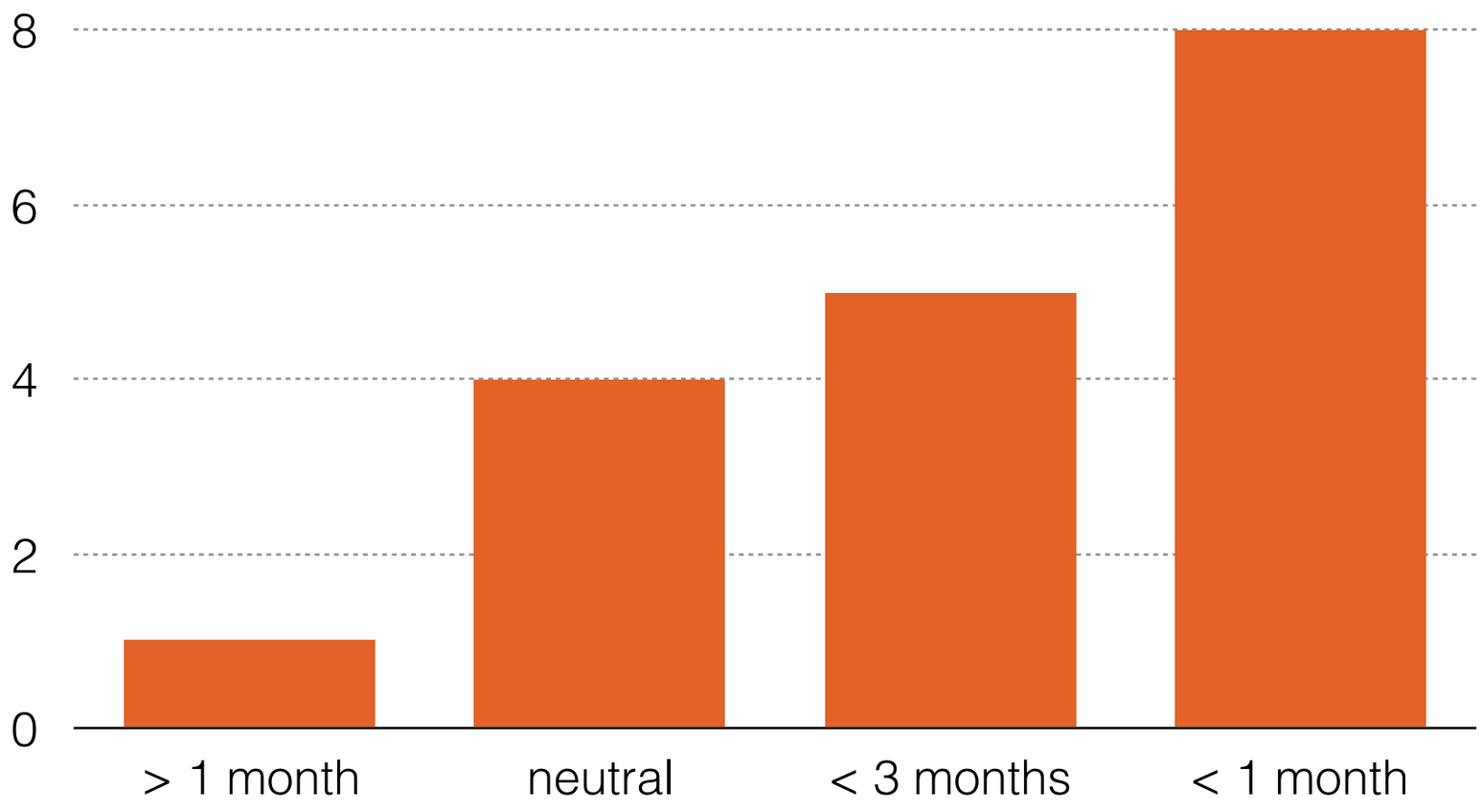
- Multiple baselines, and even multiple treatments
- Beware: the more conditions, the more participants you will need!

News recommender example:

- Only items at least 1 month old (bad baseline)
- No restrictions (neutral baseline)
- Items at most 3 months old (weak manipulation)
- Items at most 1 month old (strong manipulation)

Conditions

Expected effect on perceived novelty:



Factorial designs

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

Why?

- Efficiency
- Interaction effects

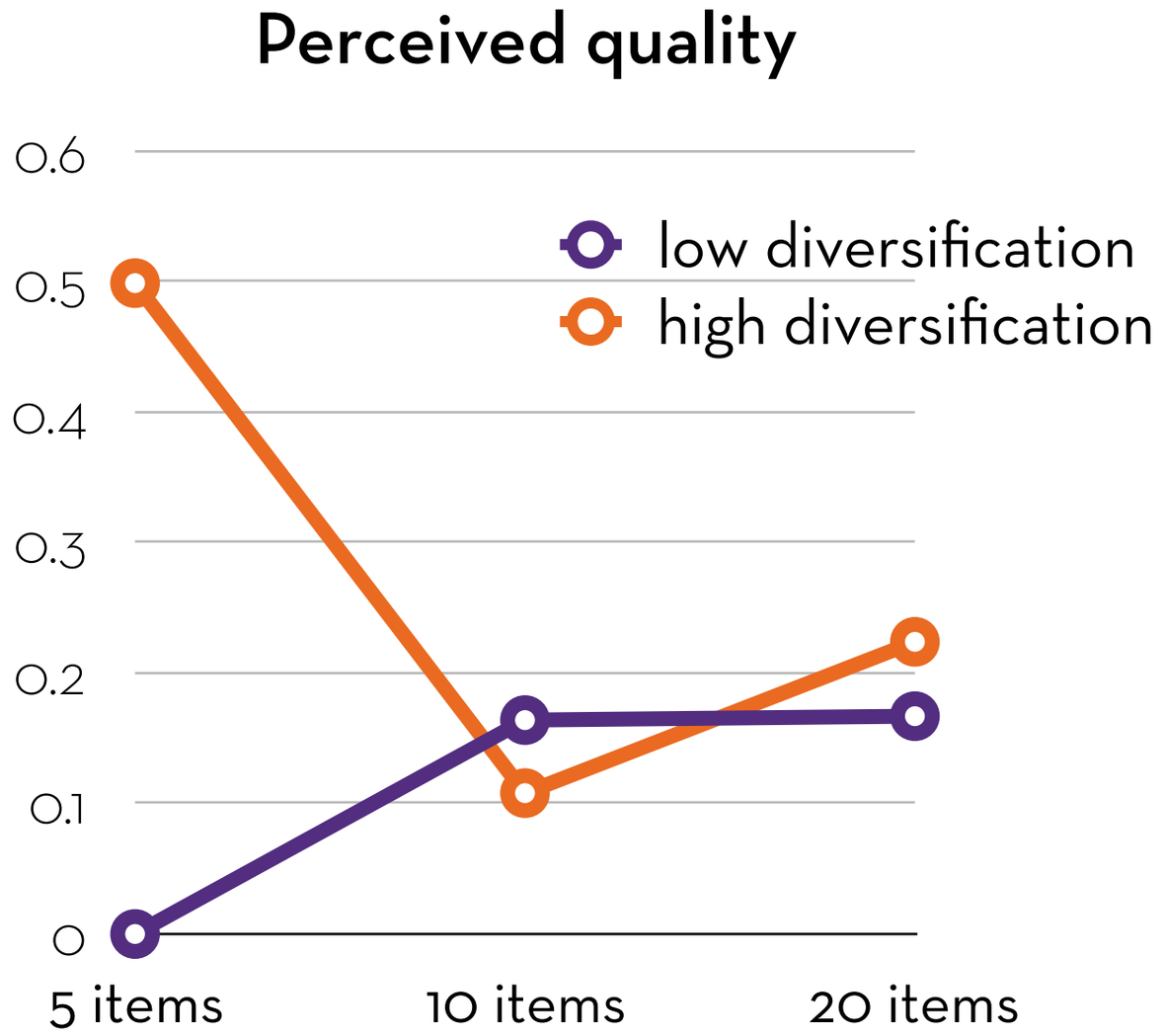
	Low diversity	High diversity
5 items	5+low	5+high
10 items	10+low	10+high
20 items	20+low	20+high

Factorial designs

Allows you to test **interaction effects**

Is the effect of diversification different per list length?

Is the effect of list length different for high and low diversification?

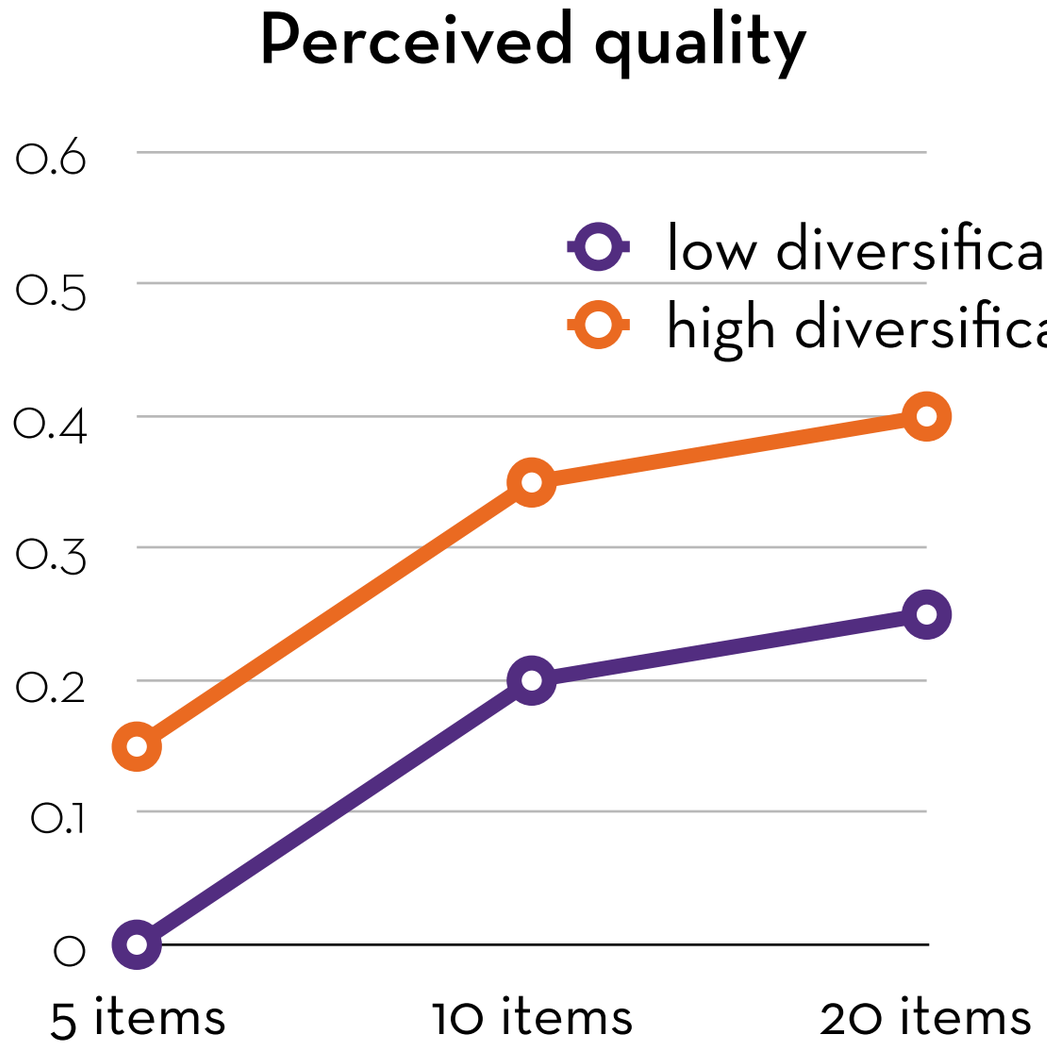


Factorial designs

If there is **no interaction** effect, you still get extra **efficiency**

To test list length, you can collapse across (ignore) diversification

To test diversification, you can collapse across list length



Complex

Let's say you want to test:

list length (5, 10),
diversification (low, high),
orientation (horiz, vert),
movie poster (no, yes),
predicted rating (no, yes):

	L	D	O	M	P
C1	5	low	horiz	no	no
C2	5	low	horiz	no	yes
C3	5	low	horiz	yes	no
C4	5	low	horiz	yes	yes
C5	5	low	vert	no	no
C6	5	low	vert	no	yes
C7	5	low	vert	yes	no
C8	5	low	vert	yes	yes
C9	5	high	horiz	no	no
C10	5	high	horiz	no	yes
C11	5	high	horiz	yes	no
C12	5	high	horiz	yes	yes
C13	5	high	vert	no	no
C14	5	high	vert	no	yes
C15	5	high	vert	yes	no
C16	5	high	vert	yes	yes
C17	10	low	horiz	no	no
C18	10	low	horiz	no	yes
C19	10	low	horiz	yes	no
C20	10	low	horiz	yes	yes
C21	10	low	vert	no	no
C22	10	low	vert	no	yes
C23	10	low	vert	yes	no
C24	10	low	vert	yes	yes
C25	10	high	horiz	no	no
C26	10	high	horiz	no	yes
C27	10	high	horiz	yes	no
C28	10	high	horiz	yes	yes
C29	10	high	vert	no	no
C30	10	high	vert	no	yes
C31	10	high	vert	yes	no
C32	10	high	vert	yes	yes

Complex

Can I shorten that?

Yes, with a fractional factorial design!

Take P = LDOM

Note: higher-order effects become confounded

	L	D	O	M	P
C1	5	low	horiz	no	yes
C2	5	low	horiz	yes	no
C3	5	low	vert	no	no
C4	5	low	vert	yes	yes
C5	5	high	horiz	no	no
C6	5	high	horiz	yes	yes
C7	5	high	vert	no	yes
C8	5	high	vert	yes	no
C9	10	low	horiz	no	no
C10	10	low	horiz	yes	yes
C11	10	low	vert	no	yes
C12	10	low	vert	yes	no
C13	10	high	horiz	no	yes
C14	10	high	horiz	yes	no
C15	10	high	vert	no	no
C16	10	high	vert	yes	yes

Complex

Even shorter?

Take M = DO, P = LO

Note: even more confounders

	L	D	O	M	P
C1	5	low	horiz	yes	yes
C2	5	low	vert	no	no
C3	5	high	horiz	no	yes
C4	5	high	vert	yes	no
C5	10	low	horiz	yes	no
C6	10	low	vert	no	yes
C7	10	high	horiz	no	no
C8	10	high	vert	yes	yes

Placebo effect

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)

Placebo effect

	“please evaluate these items”	“please evaluate these recommendations”
show users random items	random items presented as “items”	random items presented as “recommendations”
show items computed by an algorithm	recommendations presented as “items”	recommendations presented as “recommendations”

Placebo effect

"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)

Hawthorne effect

Beware of the **Hawthorne** effect

Participants may change their behavior just because they know they are being observed

When in doubt, triangulate!

Do field trial / AB-testing as well

Compare behavior between AB test and experiment

Study designs

Randomization and between- and within-subjects designs

Study designs

"The first 40 participants will get the baseline, the next 40 will get the treatment."

Randomization

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

Randomization

Change conditions each day

People may be happier on e.g., Fridays and Saturdays

Run different conditions in different subject pools

Subject pools might differ in unanticipated ways

Run different conditions in different locations

Same thing

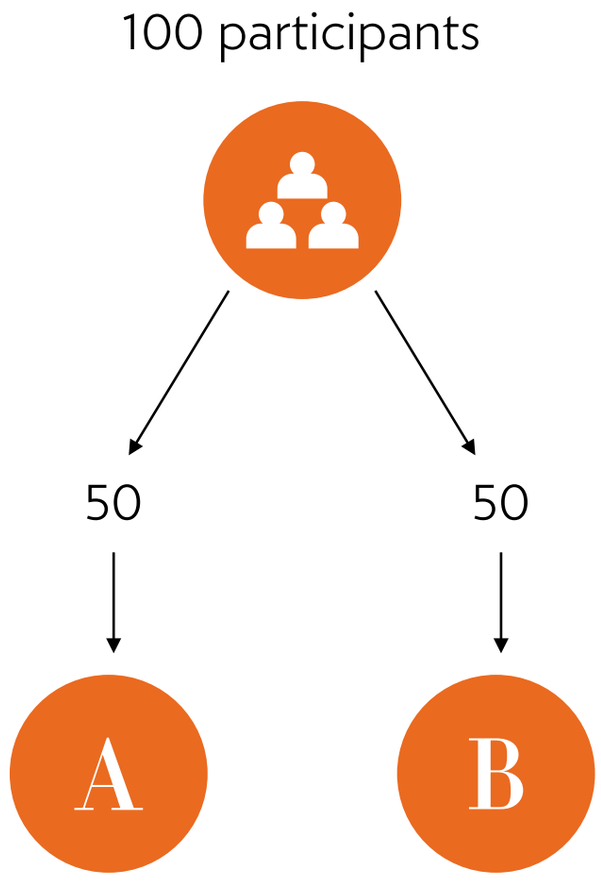
Unless you randomize days/pools/locations

Study becomes a “nested” design

Between-subjects

Randomly assign half the participants to A, half to B

- Realistic interaction
- Manipulation hidden from user
- Many participants needed

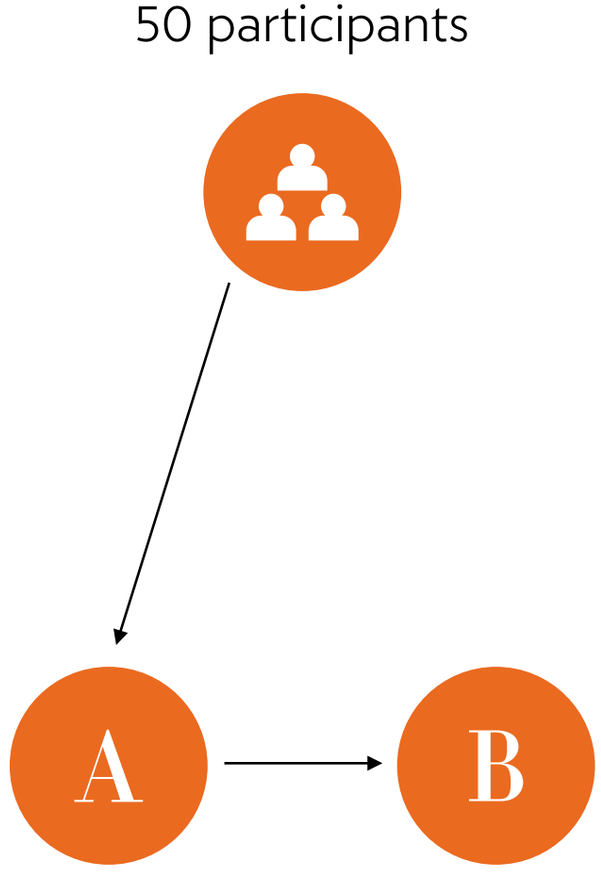


Within-subjects

Give participants A first, then B

- Remove subject variability
- Participant may see the manipulation (induces demand characteristics)
- Spill-over effect

Order should be counter-balanced!



Counter-balancing

	Trial 1	Trial 2
half of pps	A	B
other half	B	A

	T1	T2	T3	T4
1 of 4	A	B	C	D
2 of 4	B	C	D	A
3 of 4	C	D	A	B
4 of 4	D	A	B	C

Demand characteristics

For optimal experimental control, participants should be "blind" to your manipulation(s)

If they are not, this may introduce **demand characteristics**

Example: ask for a product rating before and after an AR experience

- It will be clear to participants that you are testing the AR experience
- They will know that you want the AR experience to work
- They will (unknowingly) want to please you

Demand characteristics

Solutions:

Make your test between-subjects

Half the people get the AR experience, the other half not

Use a “placebo” baseline

Test one AR experience against another (arguably less effective) experience (e.g. a TV ad)

Disassociate from the manipulation

Say you’re testing someone else’s solution

Spill-over effects

Spill-over effects in within-subjects studies

When the experience in T1 affect the experience in T2

Examples:

- Learning (positive spill-over)
- Novelty/boredom (negative spill-over)

In most cases, counter-balancing neutralizes the effect

However, it doesn't **remove** the effect (shows up as noise)

Spill-over effects

Counter-balancing does **not** work when there is **asymmetric transfer**

going from A to B has a different spill-over than going from B to A

Examples:

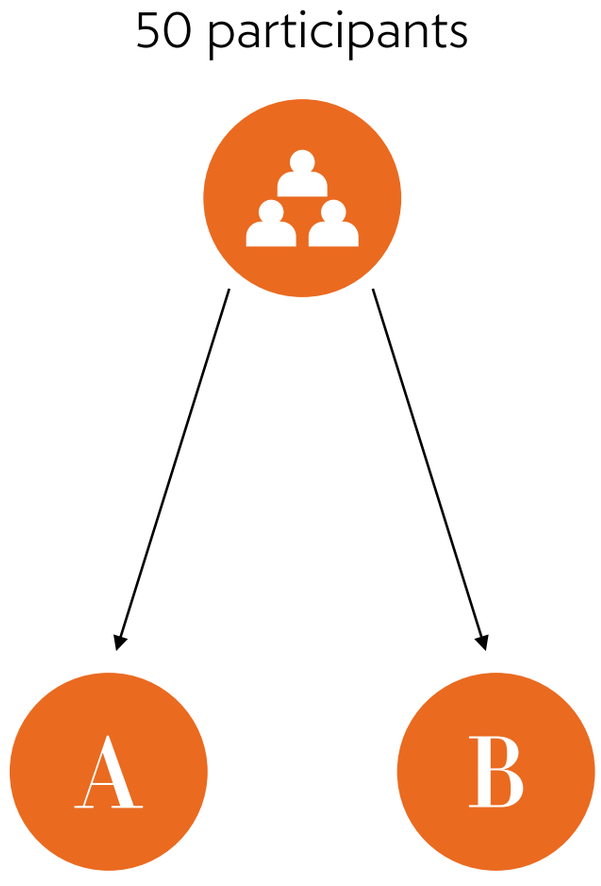
- Comparison effect
- Anchoring effect (subconscious comparison)

May reduce or exacerbate the difference between A and B!

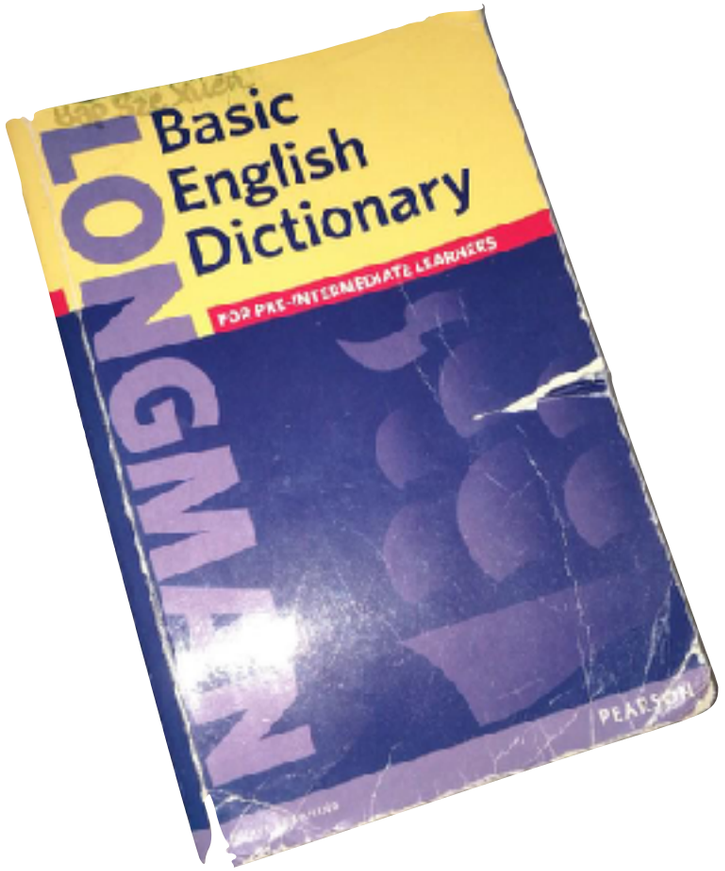
Within-subjects

Show participants A and B simultaneously

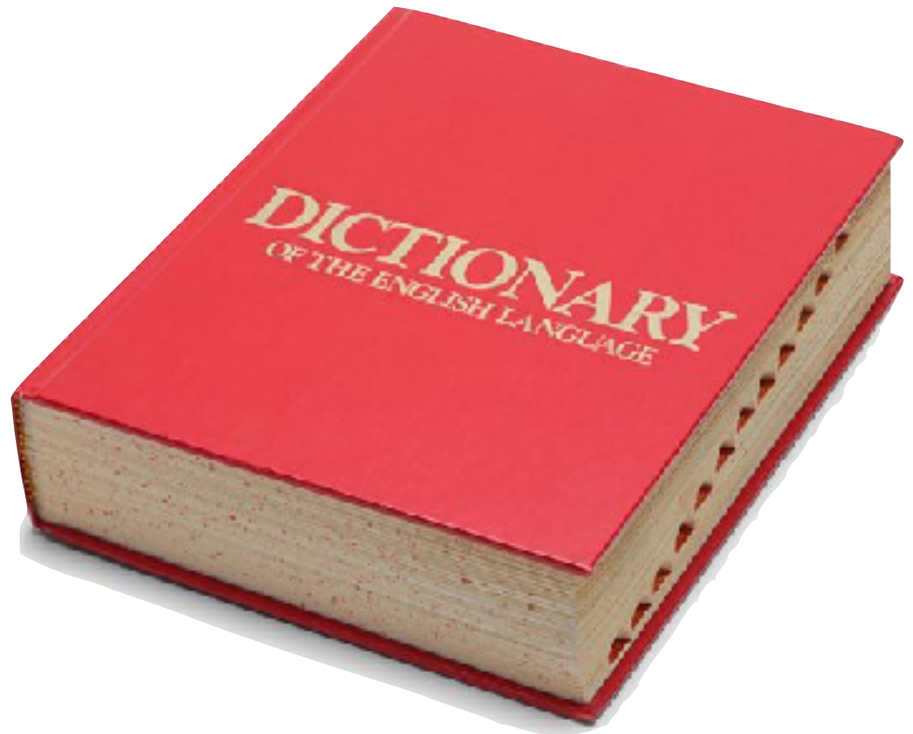
- Remove subject variability
- Participants can compare conditions
- Not a realistic interaction



Comparison?



20,000 words,
used, torn cover



10,000 words,
new condition

Which one?

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

Note: factorial designs can be within, between, or mixed

Multiple levels

Let's say I want to test the effect of gender on performance in this class...

Ankur (M)

Kevin (M)

Matias (M)

Paritosh (M)

Yifang (F)

Multiple levels

In **two** classes...

Treat class as a covariate

2016:

Ankur (M)

Kevin (M)

Matias (M)

Paritosh (M)

Yifang (F)

2017:

Adam (M)

Brian (M)

Chen (M)

Daphne (F)

Elisa (F)

Fiona (F)

Grant (M)

Multiple levels

In **many** classes...

repeated measures!

2016:

Ankur (M)

Kevin (M)

Matias (M)

Paritosh (M)

Yifang (F)

2017:

Adam (M)

Brian (M)

Chen (M)

Daphne (F)

Elisa (F)

Fiona (F)

Grant (M)

2018:

Hosub (M)

Izak (M)

James (M)

Kathy (F)

Lydia (F)

Moury (F)

Noopur (F)

Olga (F)

2019:

Praneet (M)

Quincy (M)

Rohit (M)

Sonya (F)

Thomas (M)

Multiple levels

In **many** classes + multiple assignments

...three-level model

2016:

Ankur (M) a1...a7

Kevin (M) a1...a7

Matias (M) a1...a7

Paritosh (M) a1...a7

Yifang (F) a1...a7

2017:

Adam (M) a1...a7

Brian (M) a1...a7

Chen (M) a1...a7

Daphne (F) a1...a7

Elisa (F) a1...a7

Fiona (F) a1...a7

Grant (M) a1...a7

2018:

Hosub (M) a1...a7

Izak (M) a1...a7

James (M) a1...a7

Kathy (F) a1...a7

Lydia (F) a1...a7

Moury (F) a1...a7

Noopur (F) a1...a7

Olga (F) a1...a7

2019:

Praneet (M) a1...a7

Quincy (M) a1...a7

Rohit (M) a1...a7

Sonya (F) a1...a7

Thomas (M) a1...a7

Multiple levels

Variables exist on multiple levels:

- Assignment: difficulty, time given, etc.
- Person: ability, gender, etc.
- Year: I dunno... whether Clemson won the championship the year before?

You can use any of these as covariates in your model

Outcome variables should ideally be at the lowest level

- Otherwise you'll have to "pool" variables at lower levels

Signal + noise

Signal: true difference between A and B

Noise: random variation

- Environment
- Participants
- Measurements

Within-subjects experiments: get rid of participant noise

Repeated measures: reduce measurement noise

This is why this class has multiple assignments/tests!

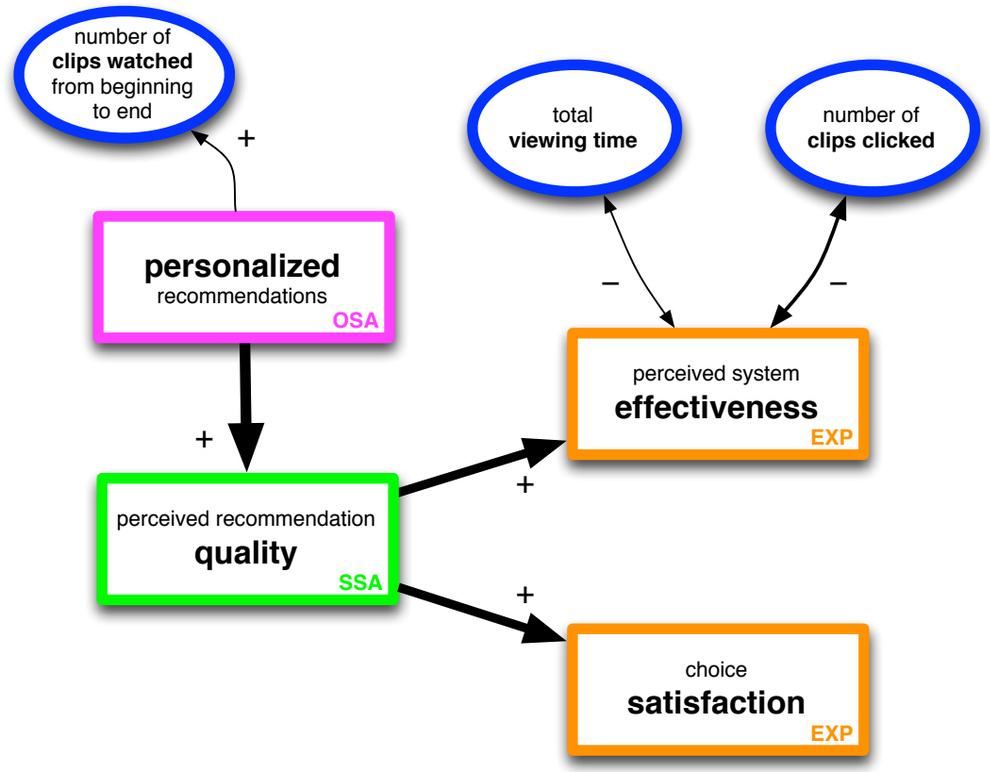
Examples

Experimental designs

Examples

The effect of recommendations on viewing clips

- Half of the participants saw random items, the other half saw personalized items



Examples

What is the effect of the number and composition of recommendations on choice overload?

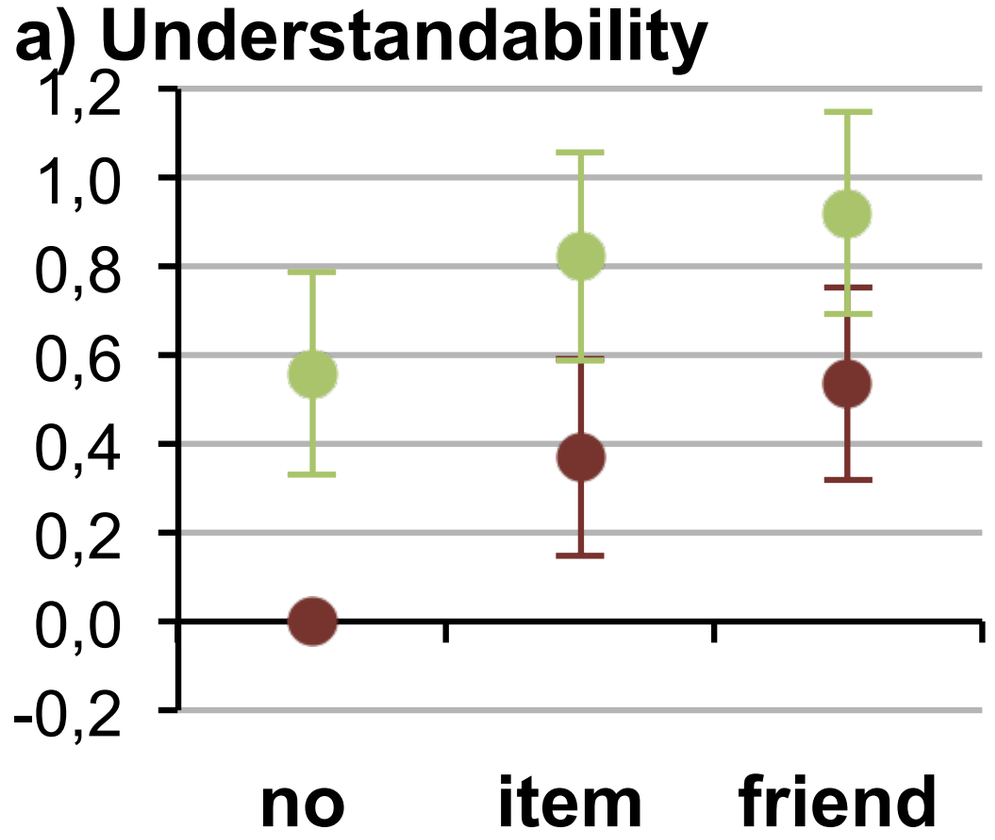
Participants randomly assigned to 1 of 3 conditions:

- 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

Examples

The effect of inspectability and control in social recommender systems:

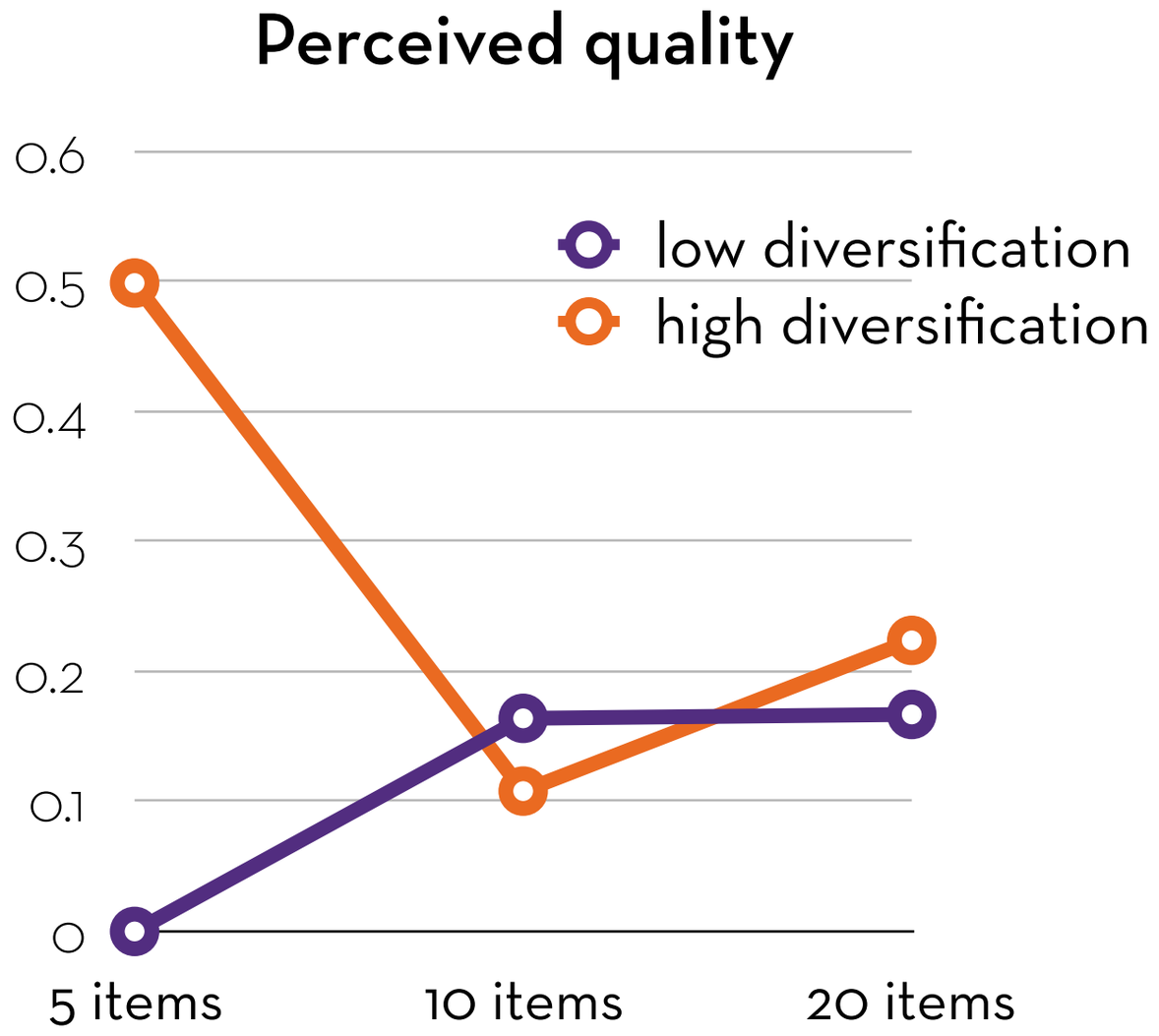
Participants randomly assigned to 1 of 2 (list view vs. graph view) x 3 (no control, item control, friend control) = 6 conditions



Examples

The effect of diversification and list length on choice overload

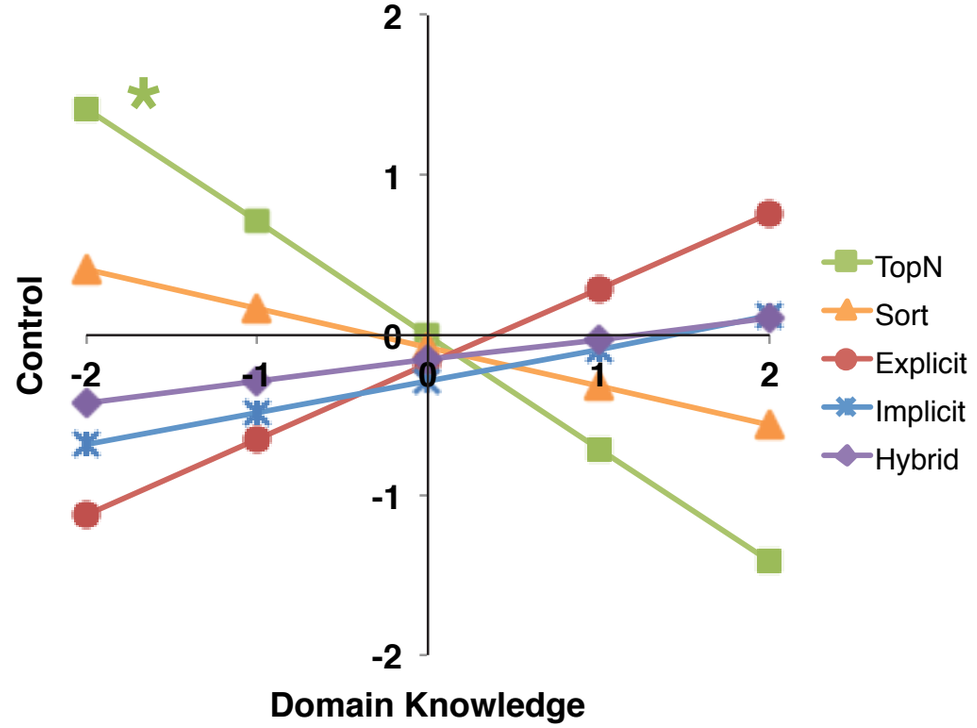
Participants randomly assigned to 1 of 2 (low vs. high diversification) x 3 (5 items, 10 items, 20 items) = 6 conditions



Examples

Domain knowledge and preference elicitation

- Participants randomly assigned to 1 of 5 conditions (Top-N, sort, explicit feedback, implicit feedback, hybrid)
- Domain knowledge as a covariate



Examples

Another one on list length and diversification…

In this case each participants saw three lists of recommendations (low, medium, high diversification)

Participants were randomly assigned to 1 of 5 list length conditions (5, 10, 15, 20, 25 items)

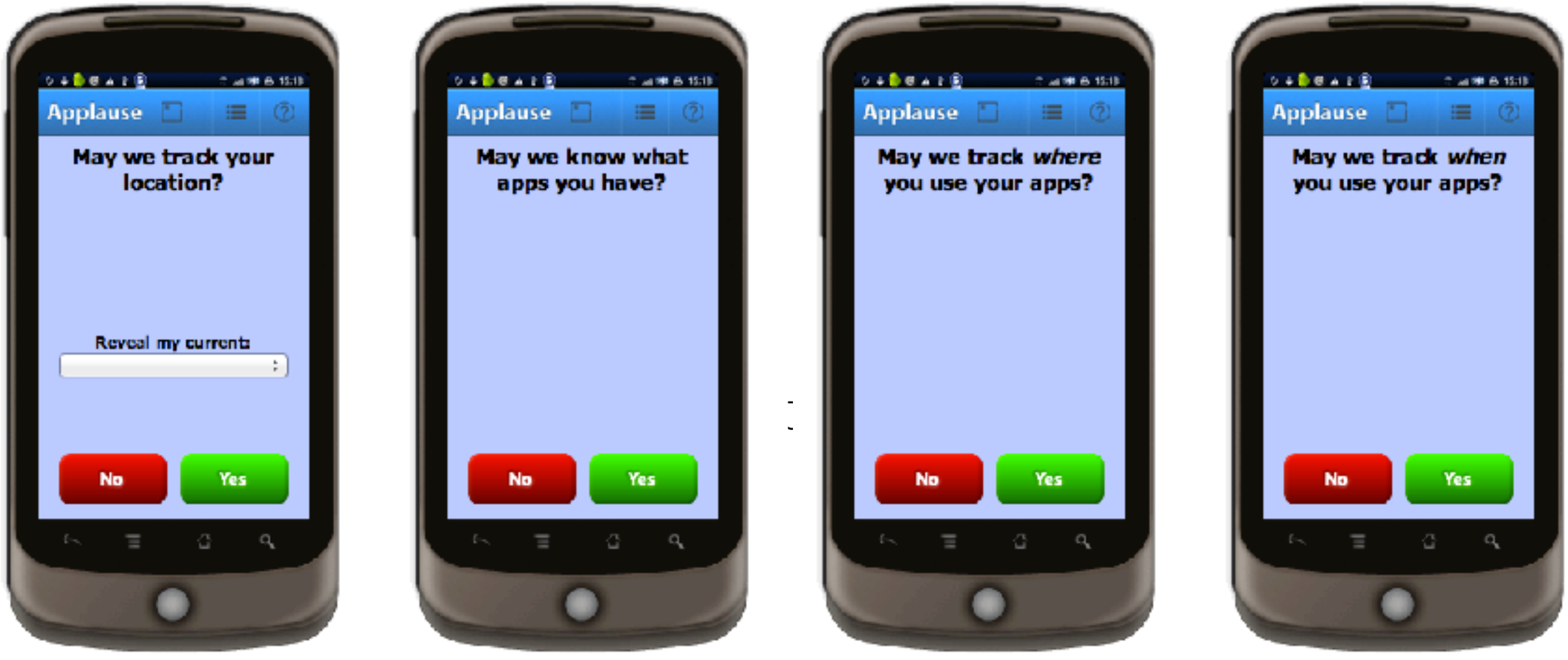
Length turned out to have no effect here

More complex...

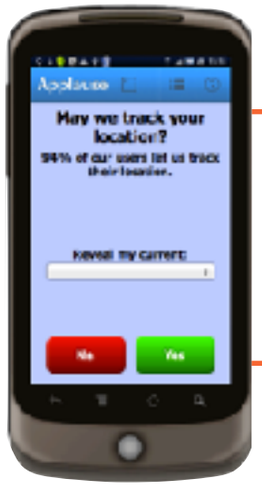
Effect of request order and justifications on privacy decision-making

- Each participant makes 31 decisions
- With one of 5 justifications
- In one of two overall request orders

More complex…

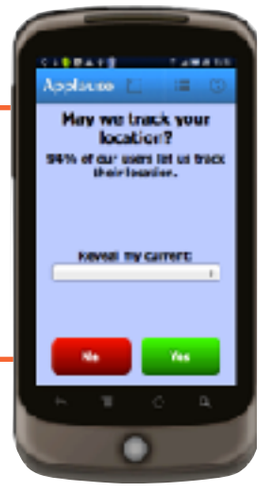


More complex...

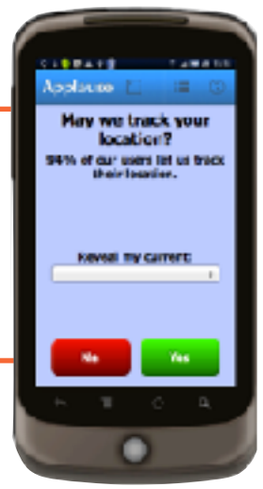


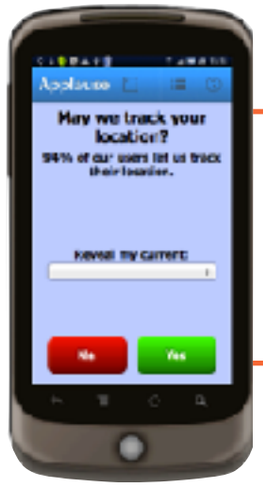
Location, etc.

Gender, etc.



Gender, etc.





Location, etc.

More complex…

5 justification types

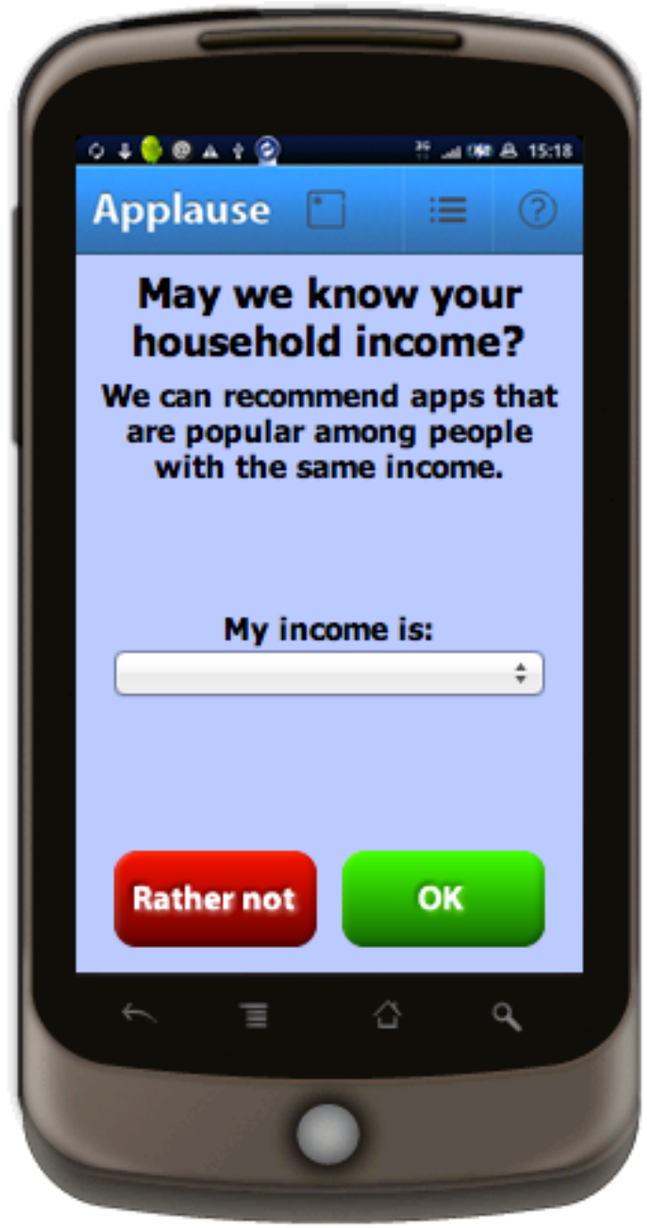
None

Useful for you

Number of others

Useful for others

Explanation



Even more complex

Effect of context variables on smart home privacy decisions

- 12 scenarios per participant
- Scenarios are manipulated along who (8 levels), what (12 levels), purpose (4 levels), storage (3 levels), action (4 levels) = 4,608 experimental conditions!
- Plus 3 levels of defaults and 3 levels of framing

Example scenario: “Your smart TV (Who) uses a camera (What) to give you timely alerts (Purpose), the data is stored locally (Storage) and used to optimize the service (Action).”