CSE 593

Quantitative User Evaluation Methods

Farnaz Jahanbakhsh

Course content adapted with modification from original materials by Nikola Banovic

EECS ELECTRICAL ENGINEERING AND COMPUTER SCIENCE
UNIVERSITY OF MICHIGAN

Logistics

- Assignment 2 is being graded
- Assignment 3 (Group) due tomorrow (Oct 30th) at 5PM.
- Midterm grades and solutions are posted.
- Reminder: You will present your final poster in Tishman Hall on Dec 5th from 3 to 4:30PM.

Goals

Define quantitative evaluation methods

Learn how to conduct a quantitative user study

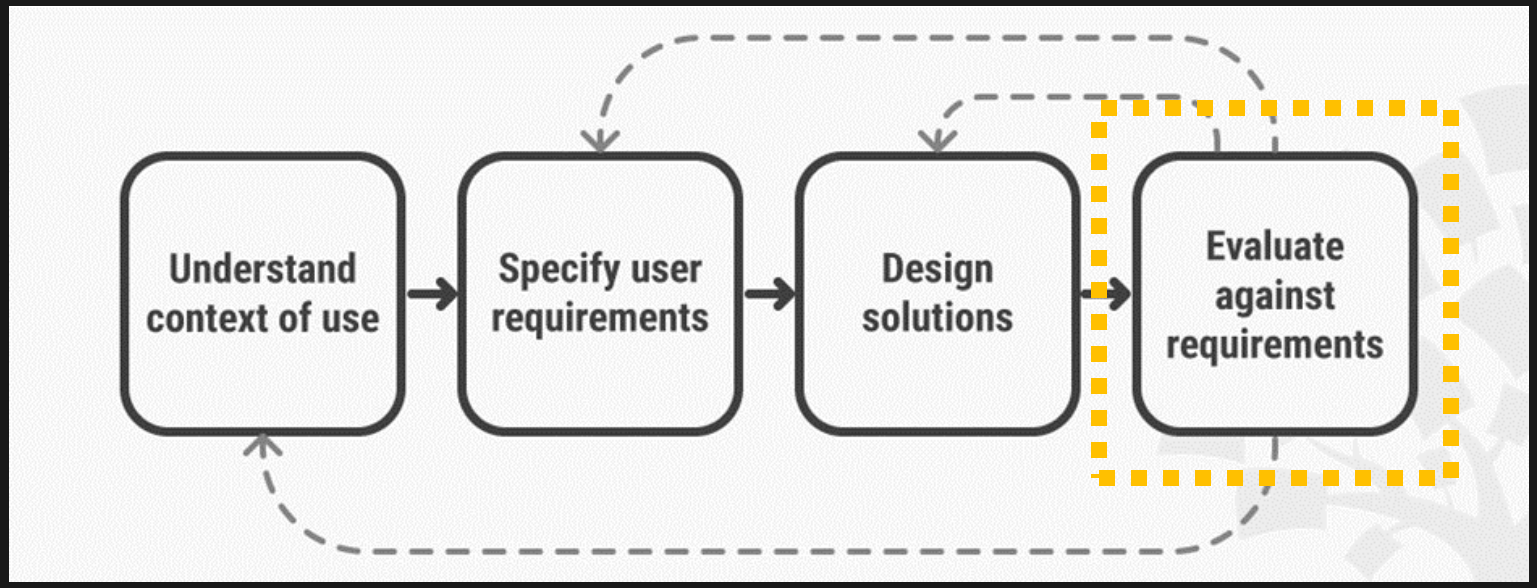
Goals

Define quantitative evaluation methods

Learn how to conduct a quantitative user study

But first, midterm exam review!

User-Centered System Design Process



Interaction Design Foundation (https://www.interaction-design.org/literature/topics/user-centered-design)

The difference between qualitative and quantitative evaluation

Qualitative
(interpretative, descriptive)
vs.
Quantitative
(measuring, statistical analysis)

Often for comparing designs

"Our design is better than the status-quo" (along some dimension)

Objective metrics: speed, accuracy, user engagement, user retention, quality of output, etc.

Composite metrics: user engagement

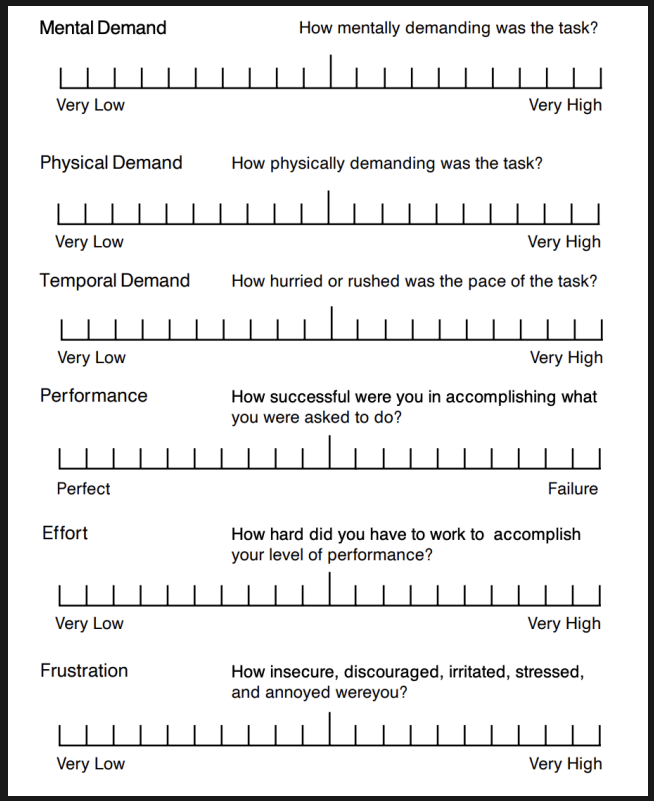
```
fav: 0.5
retweet: 1.0
reply: 13.5
good_profile_click: 12.0
video_playback50: 0.005
reply_engaged_by_author: 75.0
good_click: 11.0
good_click_v2: 10.0
negative_feedback_v2: -74.0
report: -369.0
```

https://github.com/twitter/the-algorithm-ml/tree/main/projects/home/recap

Often for comparing designs

"Our design is better than the status-quo" (along some dimension)

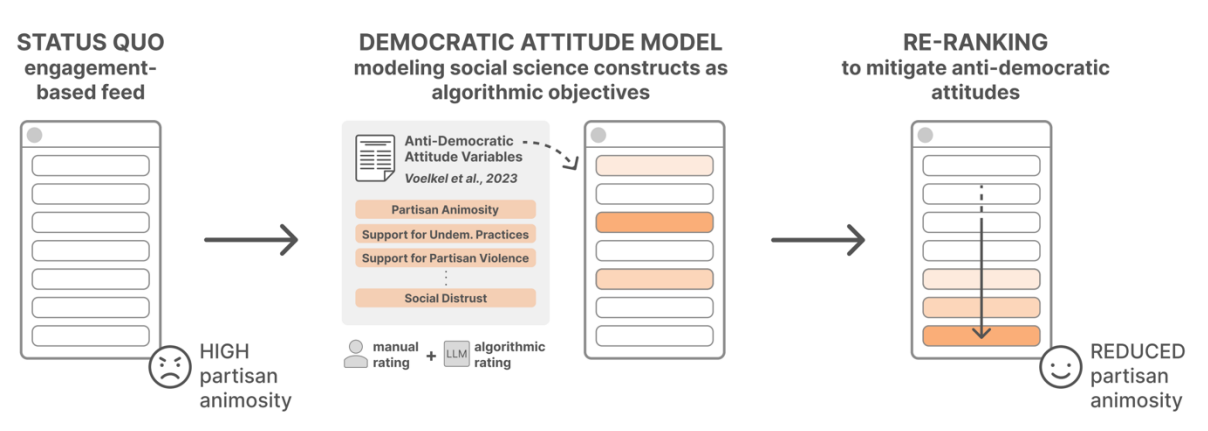
Subjective metrics: satisfaction, task load, perceived control, etc.



https://humansystems.arc.nasa.gov/groups/tlx/downloads/TLXScale.pdf

Often for comparing designs

“Our design is not worse than the status-quo” (along some dimension)



Downrank anti-democratic posts

Measure user dwell time on new feed, compare with the status-quo

Show dwell time is not harmed

and the new feed reduces partisan animosity

Jia, Chenyan, Michelle S. Lam, Minh Chau Mai, Jeffrey T. Hancock, and Michael S. Bernstein. "Embedding democratic values into social media AIs via societal objective functions." *Proceedings of the ACM on Human-Computer Interaction* 8, no. CSCW1 (2024): 1-36

Quantitative evaluation of design

Often takes form of a randomized controlled trial

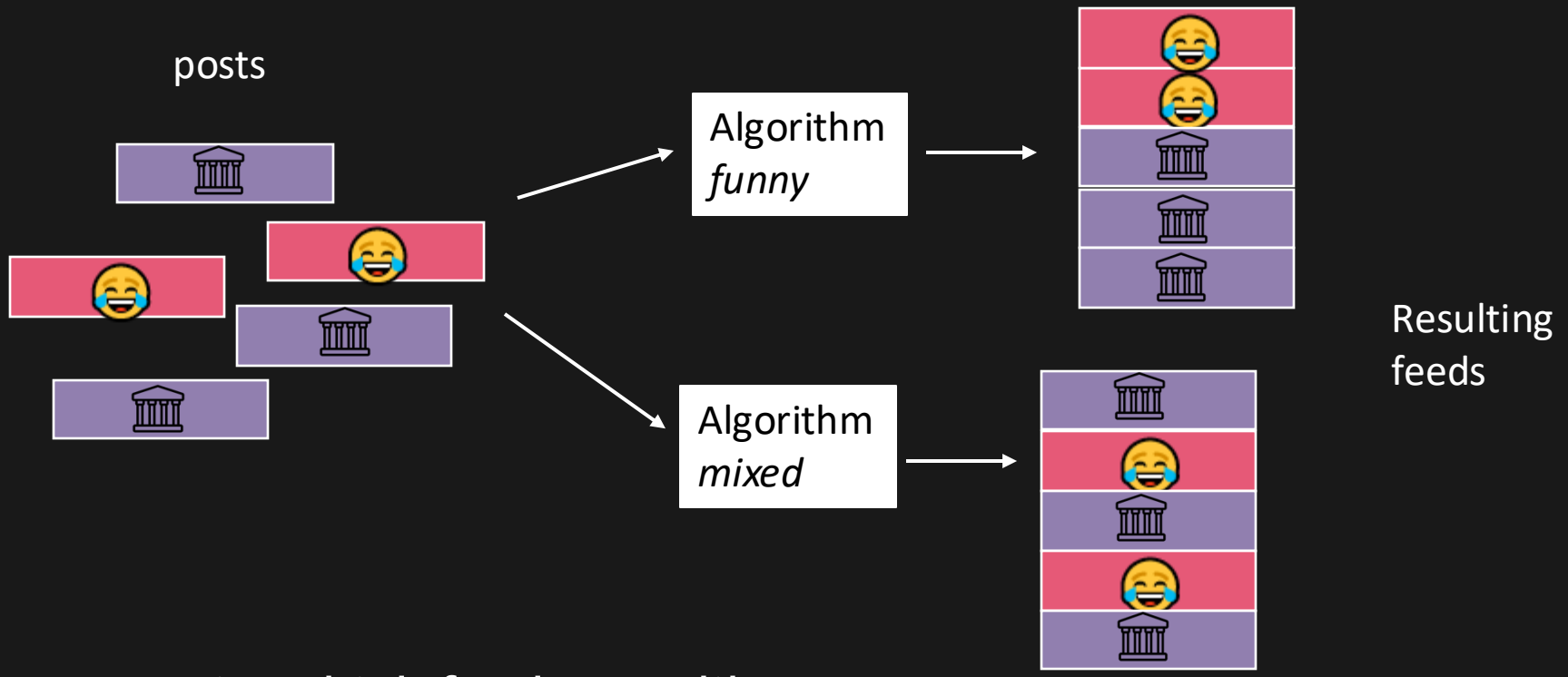
Sometimes performed in highly controlled environments (e.g., lab)

Sometimes via controlled field studies (e.g., A/B testing)

Controlled experiment vs field study

- Controlled environments isolate the effect of design differences
  - Any differences between conditions are likely due to the designs
- Controlled studies have low *ecological validity*
  - Study setup may not reflect real-world user contexts

Controlled experiment vs field study



Metric: which feed users like more

In a controlled experiment, the pool of tweets given to the algorithms might be set (pre-selected by researchers)

Low ecological validity: posts fed to curation algorithms are highly personalized

Controlled field studies

- A/B testing
  - Assigning different users to different conditions in the wild
  - Assigning the same user to different conditions sequentially over time to observe changes in behavior/outcome
- Assigning different, yet comparable groups, to different conditions: the case of New Zealand
  - Useful for testing social designs

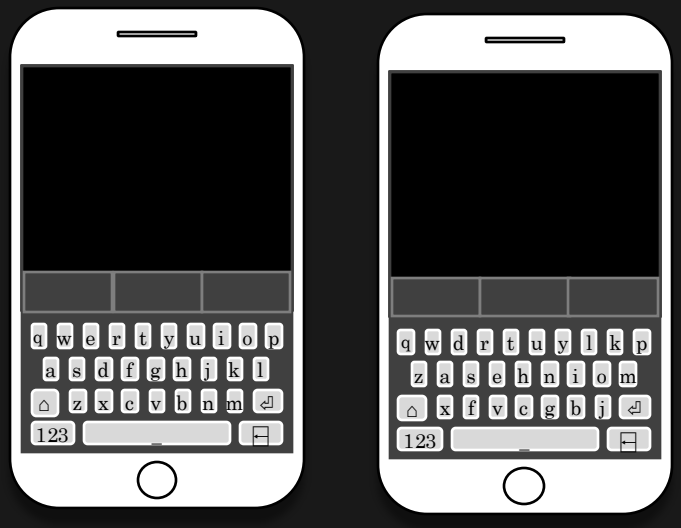
Quantitative evaluation of design

- Useful for testing a hypothesis (e.g., user performance with one design is better than performance with another design)

Quantitative evaluation of design

Pick one user goal, a sub-goal, or a task

- Standard transcription task empirical study
  - For example, a 30 minute typing session (number of phrases depends on typing speed)



Bi et al. 2010. Quasi-qwerty soft keyboard optimization. In Proc. CHI '10.

Revisiting ecological validity

- Standardizing (pre-selecting) the phrases that the users should type
- Constrained task, with potentially low ecological validity depending on phrase sampling procedure
- Dubious generalizability to out of distribution phrases
  - e.g., slang or transliteration:

    Tafa3ol Al-Insan w Al-7asoob ra2e3 gedan

  - Or to post adaptation to the keyboard

Please answer this question in Canvas

Researchers are testing two versions of a virtual assistant: one that offers brief responses and another that provides detailed explanations. Participants are invited to a lab, where they interact with both versions. They are instructed to ask a pre-defined set of questions, such as “What time is it in Tokyo?” and “How do I connect my phone to Wi-Fi?”. The order of the versions is randomized to prevent learning effects. Researchers measure response time through system logs and user satisfaction via post-interaction surveys.

Which of the following aspects of this experiment may reduce its **ecological validity**?

- ❑ The participants ask only pre-selected questions, which may not reflect the full variety of inquiries users typically make in their daily interactions.
- ❑ The experiment takes place in a controlled lab environment, which doesn’t account for interruptions or distractions that occur in real-world usage.
- ❑ Participants interact with both versions of the assistant in a back-to-back setup, which may not reflect how users engage with virtual assistants at different times throughout the day.
- ❑ Participants know they are being observed, which may change how they interact with the virtual assistant compared to unmonitored, everyday use.

Quantitative evaluation of design

Pick one user goal, a sub-goal, or a task

Design study method

Study method design

Pick independent variables – i.e., variables you control. e.g., keyboard

- Keyboard: one variable, 2 values/levels
- One variable gives us 2 experimental conditions:
  - Baseline/control
  - Intervention/treatment
- Experimental conditions: Combinations of independent variables

Study method design

Pick independent variables – i.e., variables you control.

- 2 independent variables:
  - Keyboard: new or baseline
  - Task complexity: simple or complex
- 4 conditions:

keyboard

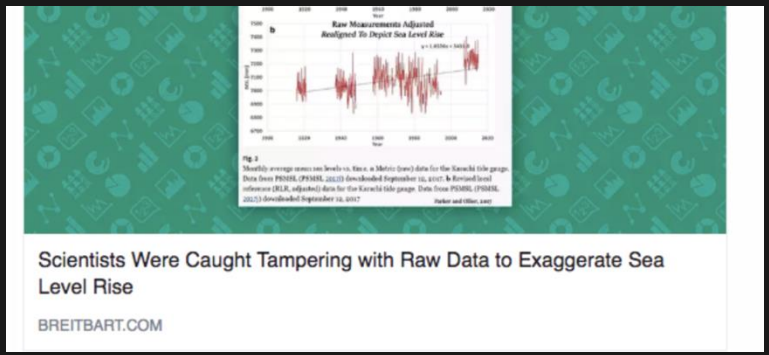
Task complexity	Baseline	New
Simple		
complex		

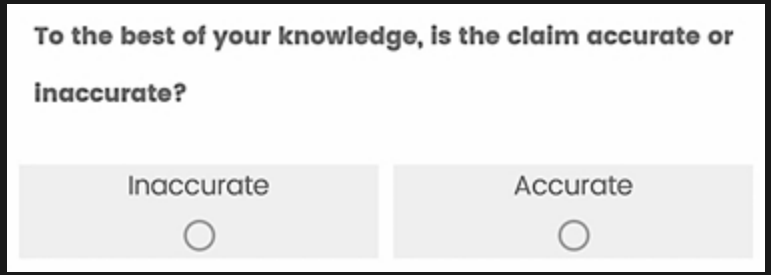
Full factorial design

Partial factorial design

Goal: Redesigning the share functionality to reduce misinformation

How: by asking users to reflect on content accuracy

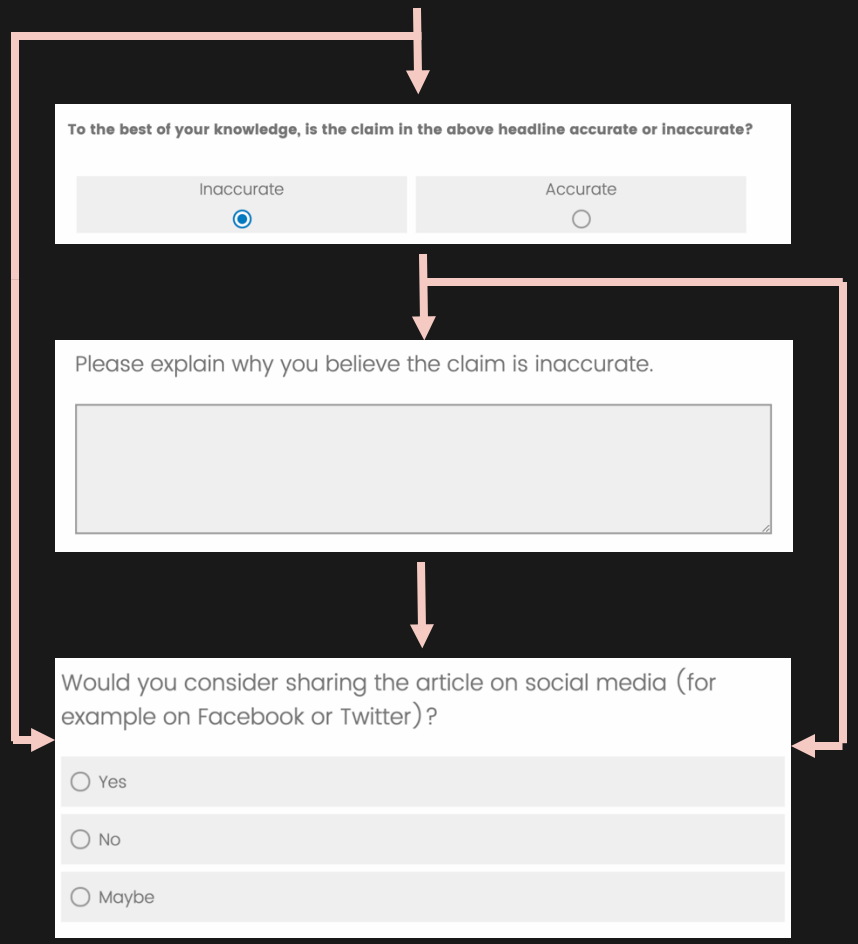




Jahanbakhsh, Farnaz, Amy X. Zhang, Adam J. Berinsky, Gordon Pennycook, David G. Rand, and David R. Karger. "Exploring lightweight interventions at posting time to reduce the sharing of misinformation on social media." *Proceedings of the ACM on human-computer interaction* 5, no. CSCW1 (2021): 1-42.

Partial factorial design

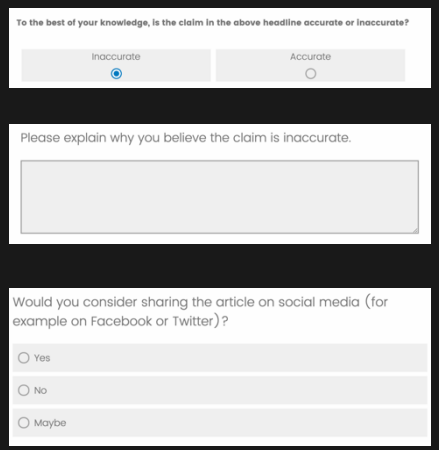
- Independent variables:
  - User is (not) asked about accuracy
  - User is (not) asked about their reasoning



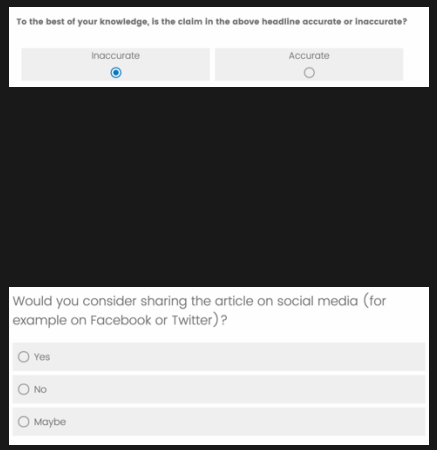
Jahanbakhsh, Farnaz, Amy X. Zhang, Adam J. Berinsky, Gordon Pennycook, David G. Rand, and David R. Karger. "Exploring lightweight interventions at posting time to reduce the sharing of misinformation on social media." *Proceedings of the ACM on human-computer interaction* 5, no. CSCW1 (2021): 1-42.

Partial factorial design

Condition 1



Condition 2



Condition 3

Condition 4

- Conditions:
  1. is asked about accuracy, is asked about reasoning
  2. is asked about accuracy, is not asked about reasoning
  3. is not asked about accuracy, is not asked about reasoning
  4. ~~is not asked about accuracy, is asked about reasoning???~~

Jahanbakhsh, Farnaz, Amy X. Zhang, Adam J. Berinsky, Gordon Pennycook, David G. Rand, and David R. Karger. "Exploring lightweight interventions at posting time to reduce the sharing of misinformation on social media." *Proceedings of the ACM on human-computer interaction* 5, no. CSCW1 (2021): 1-42.

Study method design

Pick independent variables – i.e., conditions (e.g., baseline, intervention)

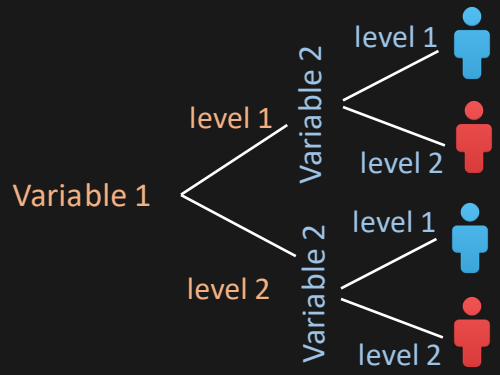
Decide on type of study (e.g., between subjects, within subjects, mixed design)

Study setup

Between subjects

Within subjects

Mixed design



Make sure users across conditions are comparable

Study method design

Pick independent variables – i.e., conditions (e.g., baseline, intervention)

Decide on type of study (e.g., between subjects, within subjects, mixed design)

Pick dependent variables (e.g., performance)

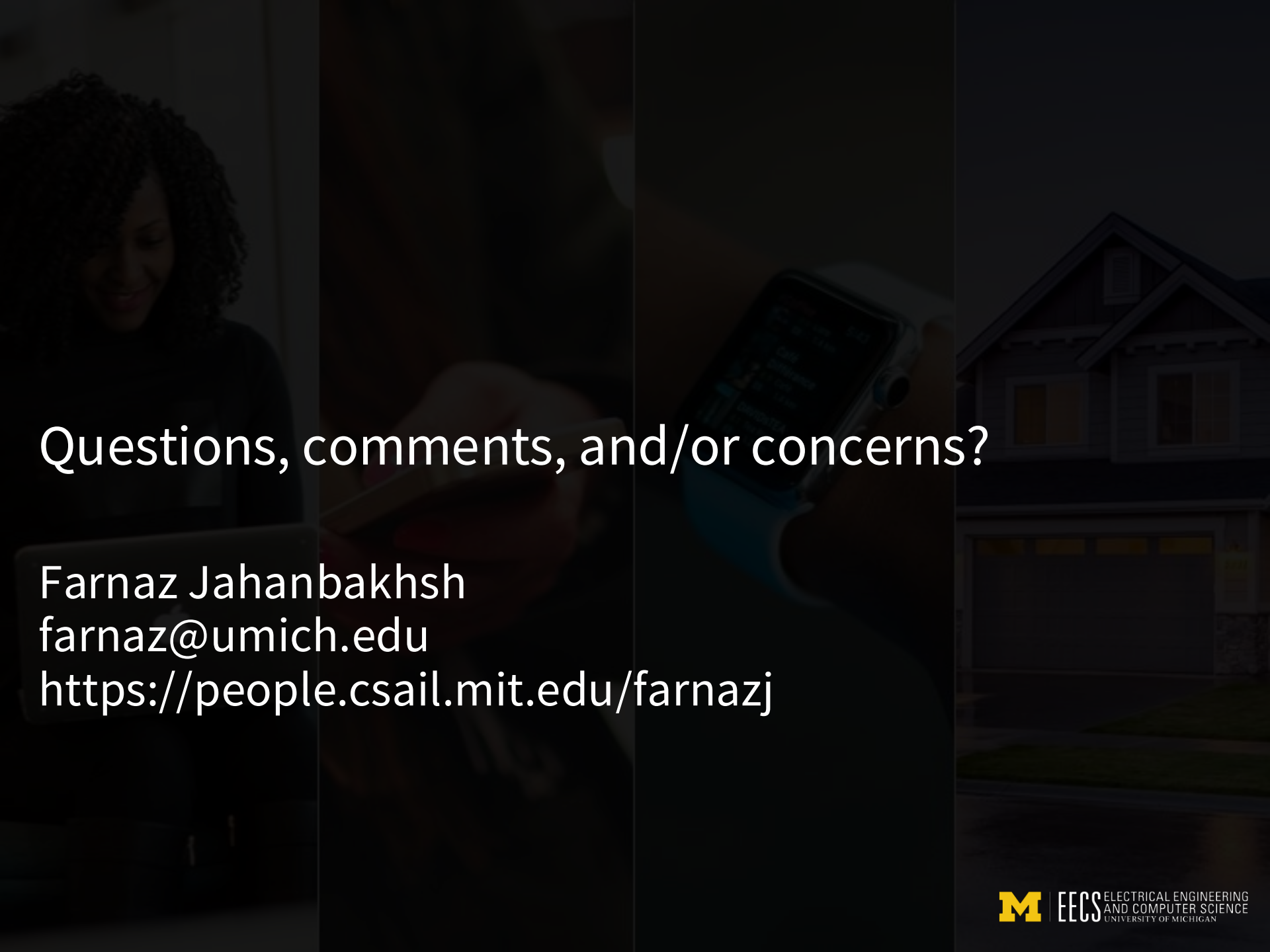
Study method design

Pick independent variables – i.e., conditions (e.g., baseline, intervention)

Decide on type of study (e.g., between subjects, within subjects, mixed design)

Pick dependent variables (e.g., performance)

Pick a statistical test



Questions, comments, and/or concerns?

Farnaz Jahanbakhsh
farnaz@umich.edu
https://people.csail.mit.edu/farnazj