06. Robust and Ethical Experiments (Part II); Survey Design; Privacy on Social Media

Blase Ur
April 18th, 2019
CMSC 23210 / 33210
An example study

• (Bad) research question: “Is UChicago the place where fun comes to die?”

• Recruiting participants: what can go wrong?

• Independent variable: assign a university

• Dependent variable: some proxy for fun
  – Hours not studying?
  – Hours not in the Reg?
  – Agreement with statement “We are having fun”
Research question

• Succinct, precisely stated, falsifiable statement or question
  – Usually, but not always, encodes some sort of hypothesis

• Goals of the research can be broad, whereas RQs are usually more narrow
Validity

- To what degree are we confident that X causes Y *(internally valid)*?
- To what degree can we generalize about our results *(externally valid)*?
  - What biases does our sample introduce?
- Is this study *(ecologically valid)*?
  - Does it mirror real-life conditions and context?
- Balancing all of these is hard!
What we conclude from studies

• It’s very rare that we conclude something like “for all humans there is an X% effect of Y” or “Z% of people care about privacy”
  – Be clear what population you have sampled

• We often use proxies in measurement
What we conclude long-term

• **Repeatability**: findings consistent with same researchers and same infrastructure

• **Reproducibility**: findings consistent with different researchers and different (comparable) infrastructure

• Sadly, few studies are replicated
  – Bias against successful replication in peer review
  – (Also) bias against publishing negative results
Some potential confounds (1/3)

• Measurement accuracy / resolution
• Differences caused by different experimental platforms and conditions
• Order of recruiting matters
  – Round-robin (123123123, etc.), Latin squares
• Time of day for recruiting matters
• Failing to account for study dropout or non-participation (very subtle!)
Some potential confounds (2/3)

• Learning effect
  – Randomize order of tasks
  – Consider learning effect as a covariate

• Different instructions for different participants

• Biases of recruitment / representativeness

• Self-report biases
  – Don’t ask people to rate expertise
Some potential confounds (3/3)

- Different demographics in conditions
- Placebo effect
  - Why you need a control condition
- Hawthorne effect (changing behavior in response to being observed)
- Participants try to please experimenter
  - I like yours better!
  - Minimize knowledge of what’s being tested
Methodology sections

• Be clear and honest about what you did
  – Be honest about limitations

• Give enough detail for someone to replicate
  – Study materials as appendix if possible
  – Correctly report stats (e.g., APA guidelines)

• Release code if possible

• Release data if possible
  – Requires approval from IRB and participants
Pilot studies

- Conduct pilot studies!!!
- Check wording
- Encourage pilot participants to tell you when there is ambiguity or uncertainty
- Verify that you’re getting the measurements you thought and that your software works
- Have people talk through even protocols that will be conducted remotely
Data to collect during experiments

- Actions and decisions
- Performance (time, success rate, errors)
- Opinions and attitudes (self-reported)
- Audio recording, screen capture, video, mouse movements, keystrokes
Even more data to collect

• Demographics
  – Age, gender, technical background, income, education, occupation, location, ability, first language, privacy attitudes, etc.

• Open-ended questions

• Preferences and attitudes (Likert scale)

Please respond to the following statements:
*This user interface was difficult to understand
1- Strongly disagree  2- Disagree  3- Neutral  4- Agree  5- Strongly agree
*This tool was fun to use
1- Strongly disagree  2- Disagree  3- Neutral  4- Agree  5- Strongly agree
Logistics for a study

• How many participants?
  – Statistical power
  – Time, budget, participants’ time

• What kind of participants?
  – Skills, background, interests
  – Their motivations
  – Often not a representative sample

• What do you need to build, if anything?
  – Prototype fidelity