05. Robust and Ethical Experiments

Blase Ur, April 10\textsuperscript{th}, 2017
CMSC 23210 / 33210
Today’s class

• Recap of (some) HCI methods
• Designing robust & ethical studies
HCI Experimental Methods
Human-Computer Interaction (HCI)

• You are not the user! You know too much!
• Think about the user throughout design
• Involve the user
What is usable?

• Intuitive / obvious
• Efficient
• Learnable
• Memorable
• Few errors
• Not annoying
• Status transparent

Image from http://www.xkcd.com
Determine use cases and goals

• What are the concrete tasks users should be able to accomplish?
  – Based on understanding of users!
• Set realistic metrics
Example: personas

Name: Patricia
Age: 31
Occupation: Sales Manager, IKEA Store
Hobbies: Painting, Fitness/biking, Taking son Devon to the park
Likes: Emailing friends & family, Surprises for her husband, Talking on cell phone with friends, Top 40 radio stations, Eating Thai food, Going to sleep late
Dislikes: Slow service at checkout lines, Smokers
Example: paper prototypes

- Don’t overthink. Just make it.
- Draw a frame on a piece of paper
- Sketch anything that appears on a card
- Make all menus, etc.
- Redesign based on feedback
- “Think aloud”
Iterative prototyping is crucial!

High-fidelity, “Wizard of Oz,” low-fidelity
Example: low-fidelity paper prototype

SCENARIO 1

"I want to listen to alternative music"
Example: paper prototype
Example: think aloud

• Download and install software that lets you encrypt your email
  – “Think aloud” of whatever’s on your mind
  – Give them an example

• Additional things you can ask:
  – What are you thinking now?
  – What do you expect to happen if you do X?
  – How did you decide to do that?
Research Studies and Methods
Research studies: purpose and goals

- What are you hoping to learn?
- What are your hypotheses?
  - Often listed explicitly in a paper
- What are your metrics for success?
  - More secure, quicker to use, more fun, etc.
- What are you comparing to?
- What data might be helpful?
Broad types of studies

- Descriptive study
- Relational study
- Experimental study
- Formative (initial) vs. summative (validate)
Quantitative vs. Qualitative

• Quantitative: you have numbers (timing data, ratings of awesomeness)

• Qualitative: you have non-numerical data (thoughts, opinions, types of errors)
Types of studies (1)

• What people want/think/do overall:
  – Surveys
  – Interviews
  – Focus groups

• What people want/think in context:
  – Contextual inquiry (interviews)
  – Diary study (prompt people)
  – Observations in the field
Types of studies (2)

• Expert evaluation of usability:
  – Cognitive walkthrough
  – Heuristic evaluation

• Usability test:
  – Laboratory (“think aloud”)
  – Online study
  – Log analysis
Types of studies (3)

• Controlled experiments to test causation

• Varying different conditions
  – Full-factorial design or not
  – Independent and dependent variables

• Many methods apply (e.g., surveys can be designed to test causation)
  – Role-playing studies
  – Field studies
Study designs

- **Within subjects**
  - Every participant tests everything
  - Crucial to randomize order! (learning effect)
  - Fewer participants

- **Between subjects**
  - Each participant tests 1 version of the system
  - You compare these groups
  - Groups should be similar (verify!)
  - Still randomize!
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
Hypothesis testing

• **Causation** (X causes Y)
  - vs. **correlation** (X is related to Y)

• Develop a hypothesis
  - Assign to conditions (include a **control**)
  - Terminology: “Condition” = “Treatment”

• H₀ (null hypothesis): there is no effect

• Hₐ or H₁ (alternative hypothesis): there is an effect
Hypothesis testing variables

• Independent variables: the thing(s) you assign / vary

• Dependent variables: the thing(s) you measure for evidence of an effect

• Co-variates: other aspects of a participant that might explain some of the effect (e.g., age, technical expertise, etc.)
P values and statistics

• Much of hypothesis testing involves calculating an appropriate statistic

• p value: probability of observing an effect at least as extreme as observed assuming the null hypothesis is true (i.e., no effect)

• $\alpha$ (alpha): cutoff for rejecting $H_0$
  – Treat this as a binary decision
  – Often $\alpha = .05$ in usable security
Is testing for significance enough?

• No! Consider:
  – Effect size (magnitude of the effect of the manipulation)
  – Power (long-term probability of rejecting $H_0$ if there really is a difference)

• Type 1 error: wrongly reject $H_0$ even if there is no effect ($\alpha$)

• Type 2 error: wrongly fail to reject $H_0$ even if there is an effect ($\beta$)
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
An example study

• 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”
Participants, ethics, and deception
Participants (1)

• Recruit people to do something remotely (e.g., online)
• Recruit people to come to your lab
• Recruit people to let you into their “context”
• Observe people (if possible, get consent! If not possible, consider necessity of design)
Participants (2)

• What recruitment mechanisms?
  – Craigslist, flyers, participant pools, representative sample, standing on street

• How do you compensate them?
  – Ethics of paying $0.00 vs. $10.00 vs. $100,000

• How do you get informed consent?

• What happens to their data?

• Prior knowledge / “what” are they?
Ethics

• How do we protect participants?
  – What risks do we introduce?

• Is there a less invasive method that would give equivalent insight?

• IRB is one arbiter of ethics; experimenters themselves are another crucial arbiter

• How do we make sure participation is voluntary throughout the experiment?
Deception

• Do we mind if participants know precisely what is being studied?
  – Sometimes, it’s crucial that we observe their organic responses in context

• What “deception” or “distraction” task can we introduce?

• How do we **debrief** people at the end?
Institutional Review Board (IRB)

• Is it research? Are there human subjects?
• Full review vs. expedited vs. exempt
• Fill out and submit protocol
  – Include all study materials (e.g., surveys)
  – Include recruitment text and/or poster
  – Leave plenty of time
What to submit to an IRB

- Full consent form (use UChicago model)
- All scripts, survey questions, instructions
- Recruitment plan
- Recruitment materials
  - You can’t emphasize compensation
- Information about how data will be handled
  - Password protection, encryption, etc.
  - Meetings to discuss
Social phishing (Jagatic et al., 2007)

• Use social networking sites to get information for targeted phishing
  – “In the study described here we simply harvested freely available acquaintance data by crawling social network Web sites.”

• “We launched an actual (but harmless) phishing attack targeting college students aged 18–24 years old.”
Social phishing (Jagatic et al., 2007)

- Control group: message from stranger
- Experimental group: message from a friend
- Used university’s sign-on service to verify passwords phished
Ethics (Jagatic et al., 2007)

• How did they obtain consent?

• What ethical concerns are there?
  – What seemed to be done well?
  – What could have been done better?

• Who was potentially affected by the study?

• “The number of complaints made to the campus support center was also small (30 complaints, or 1.7% of the participants).”