Outline

- Why do usability testing?
- Choosing participants
- Ethical considerations
- Designing & conducting the test
- Using the results
- Experimental options & details

Why do Usability Testing?

- Can’t tell how good UI is until?
  - people use it!
- Expert review methods are based on evaluators who may?
  - know too much
  - not know enough (about tasks, etc.)
- Hard to predict what real users will do
Choosing Participants

- Representative of target users. How so?
  - job-specific vocab / knowledge
  - tasks
- Approximate if needed
  - system intended for doctors?
    - get medical students or nurses
  - system intended for engineers?
    - get engineering students
- Use incentives to get participants
  - t-shirt, mug, free coffee/pizza

Ethical Considerations

- Usability tests can be distressing
  - users have left in tears
- Testing/fieldwork can be coercive if there is a power imbalance (e.g., in under resourced communities)

People may feel no option but to speak to you or give you their time even though they may not get anything of value in return.

Ethical Considerations

- You have a responsibility to alleviate these issues
  - make voluntary with informed consent (form)
  - avoid pressure to participate
  - let them know they can stop at any time
  - make collected data as anonymous as possible
- Often must get human subjects approval (IRB)

Usability Test Proposal

- A report that contains
  - objective
  - description of system being testing
  - task environment & materials
  - participants
  - methodology
  - tasks
  - test measures
- Get approved & then reuse for final report
- Seems tedious, but writing this will help "debug" your test

Selecting Tasks

- Tasks from low-fi design can be used
  - may need to shorten if
    - they take too long
    - require background that test user won’t have
- Don’t train unless that will occur in real deployment
- Avoid bending tasks in direction of what your design best supports
- Don’t choose tasks that are too fragmented?
  - fragmented = do not represent a complete goal someone would try to accomplish with your application
  - e.g., phone-in bank test

Two Types of Data to Collect

- Process data
  - observations of what users are doing & thinking
    - qualitative
- Bottom-line data
  - summary of what happened
    - time, errors, success
    - i.e., the dependent variables
    - quantitative
Which Type of Data to Collect?

- Focus on process data first
  - gives good overview of where problems are

- Bottom-line doesn’t tell you where to fix
  - just says: “too slow”, “too many errors”, etc.

- Hard to get reliable bottom-line results
  - need many users for statistical significance

The “Thinking Aloud” Method

- Need to know what users are thinking, not just what they are doing

- Ask users to talk while performing tasks
  - tell us what they are thinking
  - tell us what they are trying to do
  - tell us questions that arise as they work
  - tell us things they read

Thinking Aloud (cont.)

- Prompt the user to keep talking
  - “tell me what you are thinking”

- Only help on things you have pre-decided
  - keep track of anything you do give help on

- Make a recording & take good notes
  - make sure you can tell what they were doing
  - use a digital watch/clock
  - record audio & video
  - or even event logs

Will thinking out loud give the right answers?

- Not always

- If you ask, people will always give an answer, even if it is has nothing to do with facts
  - panty hose example

- Try to avoid specific questions
  (especially that have binary answers)
Using the Test Results

- Summarize the data
  - make a list of all critical incidents (CI)
    - positive & negative
  - include references back to original data
  - try to judge why each difficulty occurred
- What does data tell you?
  - UI work the way you thought it would?
  - users take approaches you expected?
  - something missing?

Using the Results (cont.)

- Update tasks & rethink design
  - rate severity & ease of fixing CIs
  - fix both severe problems & make the easy fixes

Measuring Bottom-Line Usability

- Situations in which numbers are useful
  - time requirements for task completion
  - successful task completion %
  - compare two designs on speed or # of errors
- Ease of measurement
  - time is easy to record
  - error or successful completion is harder
  - define in advance what these mean
- Do not combine with thinking-aloud. Why?
  - talking can affect speed & accuracy

Analyzing the Numbers

- Example: trying to get task time ≤ 30 min.
  - test gives: 40, 5, 20, 90, 10, 15
  - mean (average) = 30
  - median (middle) = 17.5
  - looks good!
- Did we achieve our goal?
- Wrong answer, not certain of anything!
- Factors contributing to our uncertainty?
  - small number of test users (n = 6)
  - results are very variable (standard deviation = 32)
- std. dev. measures dispersal from the mean

Analyzing the Numbers (cont.)

- This is what basic statistics can be used for
- Crank through the procedures and you find
  - 95% certain that typical value is between 5 & 55
Analyzing the Numbers (cont.)

<table>
<thead>
<tr>
<th>Participant</th>
<th>Time (minutes)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>20</td>
</tr>
<tr>
<td>2</td>
<td>15</td>
</tr>
<tr>
<td>3</td>
<td>40</td>
</tr>
<tr>
<td>4</td>
<td>90</td>
</tr>
<tr>
<td>5</td>
<td>10</td>
</tr>
<tr>
<td>6</td>
<td>5</td>
</tr>
</tbody>
</table>

- number of participants: 6
- mean: 30.0
- median: 17.5
- std dev: 31.8
- standard error of the mean: stddev / sqrt(#samples) = 13.0
- typical values will be mean ± 2*standard error → 4 to 56!
- what is plausible? = confidence (alpha=5%, stddev, sample size) = 95% confident between 4.6 & 55.4

Web Usability Test Results

2019/11/13
dt+UX: Design Thinking for User Experience Design, Prototyping & Evaluation

Measuring User Preference

- How much users like or dislike the system
  - can ask them to rate on a scale of 1 to 10
  - or have them choose among statements
  - “best UI I’ve ever…”, “better than average”…
  - hard to be sure what data will mean
    - novelty of UI, unrealistic setting…
  - If many give you low ratings → trouble

- Can get some useful data by asking
  - what they liked, disliked, where they had trouble, best part, worst part, etc.
  - redundant questions are OK

Comparing Two Alternatives

- **Between groups** experiment
  - two groups of test users
  - each group uses only 1 of the systems

- **Within groups** experiment
  - one group of test users
    - each person uses both systems (cheaper)
    - can’t use the same tasks or order (learning)
    - best for low-level interaction techniques
    - e.g., new mouse, new swipe interaction, …

Instructions to Participants

- Describe the purpose of the evaluation
  - “I’m testing the product; I’m not testing you”
- Tell them they can quit at any time
- Demonstrate the equipment
- Explain how to think aloud
- Explain that you will not provide help
- Describe the task
  - give written instructions
  - one task at a time
- Check if your friend has called, find out what time he will be going to the club.

25
26
27
28
29
30
31
Reporting the Results

• Report what you did & what happened
• Images & graphs help people get it!
• Video clips can be quite convincing

Heuristic Evaluation vs. User Testing

• HE is much faster
  - 1-2 hours each evaluator vs. days-weeks
• HE doesn’t require interpreting user’s actions
• User testing is far more accurate (by def.)
  - takes into account actual users and tasks
• HE may miss problems & find “false positives”
• Good to alternate between HE & user testing
  - find different problems
  - don’t waste participants

Summary

• User testing is important, but takes time/effort
• Use ????? tasks & ????? participants
  - real tasks & representative participants
• Be ethical & treat your participants well
• Want to know what people are doing & why? collect
  - process data
• Bottom line data requires ???? to get statistically reliable results
  - more participants
• Difference between between & within groups?
  - between groups: everyone participates in one condition
  - within groups: everyone participates in multiple conditions

Further Reading on Ethical Issues With Community-based Research

• “Imperialist Tendencies” blog post by Jan Chipchase, http://janchipchase.com/content/essays/imperialist-tendencies/
• “To Hell with Good Intentions” by Ivan Illich, speech to the Conference on InterAmerican Student Projects (CIASP), April 20, 1968, http://www.swaraj.org/illich_hell.htm

Next Time

• Lecture
  - Midterm (“closed-book”)
• Studio
  - Hi-fi prototype planning session