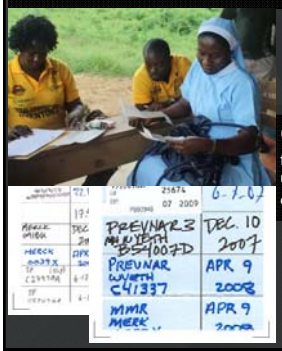
dt + UX DESIGN THINKING FOR USER EXPERIENCE DESIGN + PROTOTYPING + EVALUATION

Usability Testing

Prof. James A. Landay 刘哲明
Computer Science Department
Stanford University

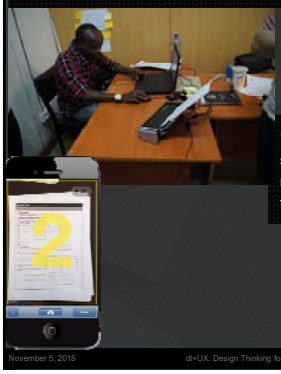
CS 147
Autumn 2015
November 5, 2015

Hall of Fame or Shame?



Captricity
forms used for public health/education data in under resourced regions
often slow to get data & data entry error prone

Hall of Fame!



Captricity
scan-in form data
machine learning to parse & humans to verify
faster & less error prone than existing methods

Hall of Fame or Shame?

Apple One Button Mouse

Hall of Shame!

How to hold this?
- No tactile clue that you were holding the mouse in the correct orientation
- Later designs added a dimple in the button yet remained ergonomically difficult to use

dt + UX DESIGN THINKING FOR USER EXPERIENCE DESIGN + PROTOTYPING + EVALUATION

Usability Testing

Prof. James A. Landay 刘哲明
Computer Science Department
Stanford University

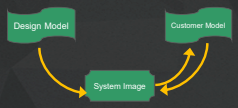
CS 147
Autumn 2015
November 5, 2015

Outline

- Review Conceptual Models
- Why do user testing?
- Choosing participants
- Ethical considerations
- Designing & conducting the test
- Using the results
- Team Break
- Experimental options & details

Review Conceptual Models

- Conceptual models ?
  - mental representation of how the object works & how interface controls effect it
- Design model should equal customer's model ?
  - mismatches lead to errors
  - use customer's likely conceptual model to design
- Design guides ?
  - make things visible
  - map interface controls to customer's model
  - provide feedback



Why do User Testing?

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

Choosing Participants

- Representative of target users ?
  - 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

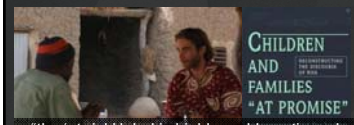
Ethical Considerations

- 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

- 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)



"the 'at-risk' label is highly problematic and often implicitly racist and classist… [it] locates problems in individuals, families, and communities, rather than in institutional structures that create and maintain inequality."

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)

"When the interviewer is a foreign researcher requiring a translator, the bias towards the interviewer's artifact increases to 5x."

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
  - stress that you are testing the system, not them
  - make collected data as anonymous as possible
- Often must get human subjects approval (IRB)

http://research.uthscsa.edu/irb/images/IRBmeeting4.jpg

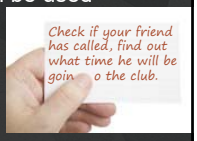
User Test Proposal

- 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 analysis & design can be used
  - may need to shorten if
    - they take too long
    - require background that test user won't have



- Try not to train unless that will happen 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*

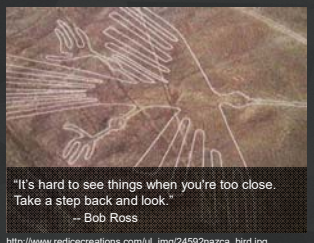
http://allazollers.com/discovery-research.php

- Bottom-line data
  - summary of what happened
    - time, errors, success
  - i.e., the dependent variables
  - *quantitative*

http://www.fusionfarm.com/content/uploads/2012/10/analyzing-data.jpg

Which Type of Data to Collect?

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



"It's hard to see things when you're too close. Take a step back and look."
-- Bob Ross

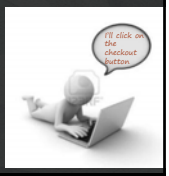
http://www.redicecreations.com/ul_img/24592nazca_bird.jpg

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
- Make a recording or take good notes
  - make sure you can tell what they were doing



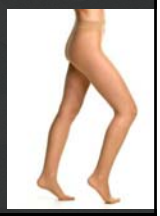
Thinking Aloud (cont.)

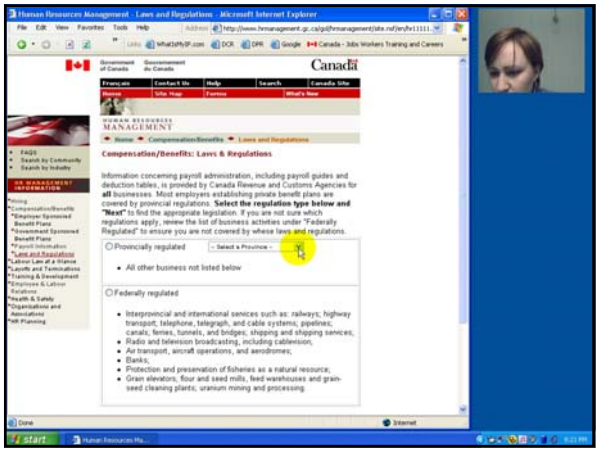
- Prompt the user to keep talking
  - "tell me what you are thinking"
- Only help on things you have pre-decided
  - keep track of anything you do give help on
- Recording
  - use a digital watch/clock
  - take notes, plus if possible
    - record audio & video (or even event logs)

Will thinking out loud give the right answers?

- Not always
- If you ask, people will always give an answer, even it is has nothing to do with facts
  - panty hose example

→Try to avoid specific questions





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

http://www.thetomorrowplan.com/exchange/policies-prairie-chickens-and-parking/

Team Break

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: 20, 15, 40, 90, 10, 5
  - 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 statistics is for
- Crank through the procedures and you find
  - 95% certain that typical value is between 5 & 55

Analyzing the Numbers (cont.)

Web Usability Test Results

Participant #	Time (minutes)
1	20
2	15
3	40
4	90
5	10
6	5
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)	25.4 --> 95% confident between 5 & 56

Analyzing the Numbers (cont.)

- This is what statistics is for
- Crank through the procedures and you find
  - 95% certain that typical value is between 5 & 55
- Usability test data is quite variable
  - need lots to get good estimates of typical values
  - 4x as many tests will only narrow range by 2x
    - breadth of range depends on *sqrt of # of test users*
  - this is when online methods become useful
    - easy to test w/ large numbers of users

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, feelings, not realistic 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

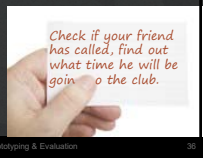
- *Between groups* experiment
  - two groups of test users
  - each group uses only 1 of the systems
- *Within groups* experiment
  - one group of test users
    - each person uses both systems
    - can't use the same tasks or order (learning)
  - best for low-level interaction techniques
    - e.g., new mouse, new swipe interaction, …

Comparing Two Alternatives

- Between groups requires many more participants than within groups
- See if differences are statistically significant
  - assumes normal distribution & same std. dev.
- Online companies can do large AB tests
  - look at resulting behavior (e.g., buy?)

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



Reporting the Results

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

HE 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

- Children and Families "At Promise, Beth B. Swadener, Sally Lubeck, editors, SUNY Press, 1995, http://www.sunypress.edu/p-2029-children-and-families-at-promis.aspx
- "Yours is better!" Participant Response Bias in HCI, Proceedings of CHI 2012, by Nicola Dell, et al., http://research.microsoft.com/pubs/163718/CHI2012-Dell-ResponseBias-proc.pdf
- "Strangers at the Gate: Gaining Access, Building Rapport, and Co-Constructing Community-Based Research", Proceedings of CSCW 2015, by Christopher A. Le Dantec & Srah Fox, http://dl.acm.org/citation.cfm?id=2675133.2675147&coll=DL&dl=ACM
- "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 review
- Project
  - Work on it in class