

HCI+D: USER INTERFACE DESIGN + PROTOTYPING + EVALUATION

Usability Testing



Prof. James A. Landay
Computer Science Department
Stanford University

Autumn 2014
November 4, 2014

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

Hall of Fame or Shame?



Apple Remote
courtesy Alexander W.

Hall of Fame!



Sleek, clean & easy to understand
Reduced to 3 key controls

- menu button
- play/pause button
- navigation pad

Buttons easy to press one-handed
Simple color scheme

HCI+D: USER INTERFACE DESIGN + PROTOTYPING + EVALUATION

Usability Testing



Prof. James A. Landay
Computer Science Department
Stanford University

Autumn 2014
November 4, 2014

Outline

- Review Heuristic Evaluation
- Why do user testing?
- Choosing participants
- Designing the test
- Collecting data
- Team Break
- Analyzing the data
- Experimental Details

Autumn 2014

HCI+D: User Interface Design, Prototyping, & Evaluation

6

Review Heuristic Evaluation

- Usability method that relies on who?
 - experts
- Ask evaluators to see if UI complies with heuristics
 - note where it doesn't, say why, & suggest fix
- Combine the findings from 3 to 5 evaluators ?
 - different evaluators find different problems
 - adding more won't be worth the cost
- Cheaper or more expensive than user testing ?
 - cheaper than user testing (time/cost)
- False positives ?
 - HE may find problems that users may never encounter
- Alternate with user testing

Autumn 2014

HCI+D: User Interface Design, Prototyping, & Evaluation

8

Why 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

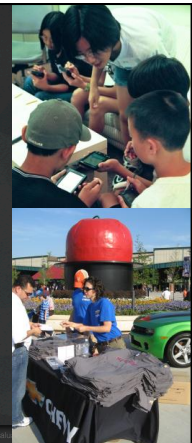


Autumn 2014

HCI+D: User Interface Design, Prototyping, & Evaluation

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

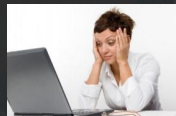


Autumn 2014

HCI+D: User Interface Design, Prototyping, & Evaluation

Ethical Considerations

- Usability tests can be distressing
 - users have left in tears
- You have a responsibility to alleviate
 - 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



Autumn 2014

HCI+D: User Interface Design, Prototyping, & Evaluation

10

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



Autumn 2014

HCI+D: User Interface Design, Prototyping, & Evaluation

12

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 complete with your application
 - e.g., phone-in bank test



Autumn 2014 HCI+D: User Interface Design, Prototyping, & Evaluation 16

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*



Autumn 2014 HCI+D: User Interface Design, Prototyping, & Evaluation 17

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



Autumn 2014 HCI+D: User Interface Design, Prototyping, & Evaluation 18

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



Autumn 2014 HCI+D: User Interface Design, Prototyping, & Evaluation 19

Thinking Aloud (cont.)

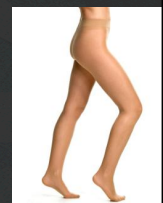
- Prompt the user to keep talking
 - "tell me what you are thinking"
- Only help on things you have pre-decided
 - keep track of anything you do give help on
- Recording
 - use a digital watch/clock
 - take notes, plus if possible
 - record audio & video (or even event logs)



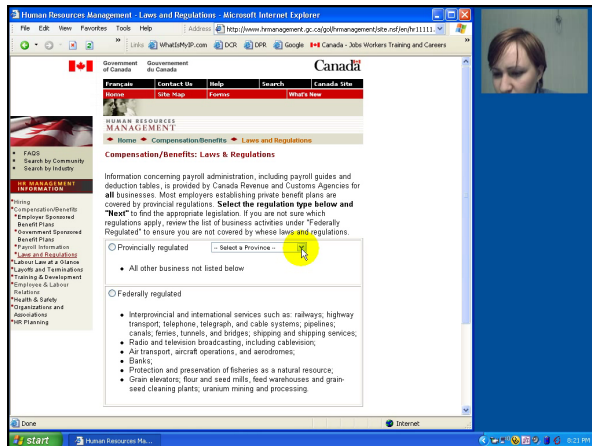
Autumn 2014 HCI+D: User Interface Design, Prototyping, & Evaluation 20

Will thinking out loud give the right answers?

- Not always
 - If you ask, people will always give an answer, even if it has nothing to do with facts
 - panty hose example
- Try to avoid specific questions



Autumn 2014 HCI+D: User Interface Design, Prototyping, & Evaluation 21



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 task analysis & 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/>

Administrivia

- Poll on Piazza for potential class times for CS 147B (course number to change)
 - final approval is imminent, but not done
 - will be able to count as Senior Project course (right now by petition – clarifying)
- Questions on HE assignment?
- CAs sending you email with HE team assignments
 - complete reports individually
 - bring your report on paper & computer/USB to class
 - in class you will combine reports using template
- Web sites
 - all teams received email on AFS space
 - put simple page up with all your assignments
 - recommend put video in prominent position
 - see examples for how to do it well (not hard)
 - try to get something up (at least basic) this week

**TEAM BREAK
 (20 MINUTES)**
DISCUSS WEB SITES

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



Autumn 2014

HCI+D: User Interface Design, Prototyping, & Evaluation

25

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

Autumn 2014

HCI+D: User Interface Design, Prototyping, & Evaluation

26

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

Autumn 2014

HCI+D: User Interface Design, Prototyping, & Evaluation

27

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
 - 4 times 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

Autumn 2014

HCI+D: User Interface Design, Prototyping, & Evaluation

28

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



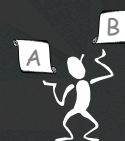
Autumn 2014

HCI+D: User Interface Design, Prototyping, & Evaluation

29

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, ...



Autumn 2014

HCI+D: User Interface Design, Prototyping, & Evaluation

30

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?)

Autumn 2014

HCI+D: User Interface Design, Prototyping, & Evaluation

31

Experimental Details

- Order of tasks
 - choose one simple order (simple → complex)
 - unless doing within groups experiment
- Training
 - depends on how real system will be used
- What if someone doesn't finish
 - assign very large time & large # of errors or remove & note
- Pilot study
 - helps you fix problems with the study
 - do two, first with colleagues, then with real users

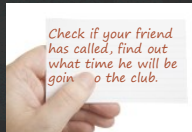
Autumn 2014

HCI+D: User Interface Design, Prototyping, & Evaluation

32

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



Autumn 2014

HCI+D: User Interface Design, Prototyping, & Evaluation

33

Details (cont.)

- Keeping variability down
 - recruit test users with similar background
 - brief users to bring them to common level
 - perform the test the same way every time
 - don't help some more than others (plan in advance)
 - make instructions clear
- Debriefing test users
 - often don't remember, so demonstrate or show video segments
 - ask for comments on specific features
 - show them screen (online or on paper)

Autumn 2014

HCI+D: User Interface Design, Prototyping, & Evaluation

34

Reporting the Results

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



Autumn 2014

HCI+D: User Interface Design, Prototyping, & Evaluation

35

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

Autumn 2014

HCI+D: User Interface Design, Prototyping, & Evaluation

36

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

Autumn 2014

HCI+D: User Interface Design, Prototyping, & Evaluation

37

Next Time

- Design Patterns
- Read
 - The Design of Sites*, by van Duyne, Hong, & Landay
 - 1) "Making the Most of Web Design Patterns" (Ch 2)
 - 2) "Up-Front Value Proposition" (Pattern C2)
 - 3) "Process Funnel" (Pattern H1)
 - 4) "Meaningful Error Messages" (Pattern K13)

Autumn 2014

HCI+D: User Interface Design, Prototyping, & Evaluation

38