Hall of Shame

- Does not follow OBVIOUS LINKS (K10) pattern
- Navigation separate from content – no links on right
- Why is this about Fry’s ISP? – I’m looking for a store!

Hall of Fame or Hall of Shame?

• frys.com

Hall of Shame

• HFS Husky Card Account Page

Hall of Fame or Hall of Shame?

• HFS Husky Card Account Page
  – violates PREVENTING ERRORS (K12)
Hall of Fame or Shame?

- The page you get if you get it wrong

Hall of Shame

- The page you get if you get it wrong
  - what is Blackboard Academic Suite?
  - where am I?
  - Is this really the UW site?
  - violates SITE BRANDING (E1)
  - what is the error?
  - violates MEANINGFUL ERROR MESSAGES (K13)

Outline

- Review
- Why do user testing?
- Choosing participants
- Designing the test
- Collecting data
- Administrivia
- Analyzing the data
- Course survey

Rapid Prototyping Review

- Informal prototyping tools bridge the gap between paper & high-fi tools
- High-fi UI tools good for testing more developed UI ideas
- Two styles of tools
  - “Prototyping” vs. UI builders
  - what is the difference?
- Both types generally ignore the “insides” of application → this is research

Why do User Testing?

- Can’t tell how good UI is until?
- Other 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
  - system intended for engineers?
    - get engineering students
- Use incentives to get participants

Ethical Considerations

- Sometimes tests can be distressing
  - users have left in tears
- You have a responsibility to alleviate
  - make voluntary with informed consent
  - 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

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

- Should reflect what real tasks will be like
- 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
  - e.g., phone-in bank test

Deciding on Data to Collect

- Two types of data
  - process data
    - observations of what users are doing & thinking
  - bottom-line data
    - summary of what happened (time, errors, success)
    - i.e., the dependent variables

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)

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
• Will thinking aloud give the right answers?
  – not always
  – if you ask a question, people will always give an answer, even if it is has nothing to do with facts
    • panty hose example
  – try to avoid specific questions

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

Video of a Test Session

http://www.maskery.ca/testvideo/webdemo1.html
http://www.maskery.ca/testvideo/webdemo3.html
http://dmc-av.cj1042.umn.edu/usability.ram
Administrivia

- Assignment #6 due Fri in email at 5PM – MUST also be on your web site – needed by others in the class
- Heuristic Evaluation assignment due at start of class on Tue (meet in the lab again)
- Posters due next Wed at 3 PM
- Final presentations next Thursday
- Class presentations will be attended by industry reps
  – I’m catering lunch afterwards if you can stay

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!
- 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.)

<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>10</td>
</tr>
<tr>
<td>4</td>
<td>10</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>6</td>
<td>5</td>
</tr>
</tbody>
</table>

- number of participants: 6
- mean: 30.0
- median: 17.0
- std dev: 31.8
- standard error of the mean = std dev / sqrt (number of samples) = 13.0
- typical values will be mean +/- 2 * standard error = 4 to 56!
- what is plausible? = confidence (alpha=5%, std dev, sample size) = 25.6 -> 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
  – 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

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
- Between groups requires many more participants than within groups
- See if differences are statistically significant
  - assumes normal distribution & same std. dev.

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 2, first with colleagues, then with real users

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

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)

Reporting the Results

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

Summary

- User testing is important, but takes time/effort
- Early testing can be done on mock-ups (low-fi)
- Use ????? tasks & ????? participants
  - real tasks & representative participants
- Be ethical & treat your participants well
- Want to know what people are doing & why, so?
  - collect process data
- Using 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
Next Time

- In lab group heuristic evaluation summary