Human-Computer Interaction Design Studio

February 21 – Testing

Terry Winograd and Bill Verplank
TAs: Kevin Collins & Marcello Bastea-Forte

08 January 2008

http://cs247.stanford.edu

When do you evaluate?

- Summative Evaluation – After design is deployed
  - Measure effectiveness
  - Check standards
  - Guide adoption decisions
- Formative Evaluation – During design and development process

Formative Evaluation

- Exploration
  - Assess what it would take for a design to fulfills users' needs and likes
- Development
  - Evaluate alternatives
  - Anticipate breakdowns
- Deployment
  - Upgrade subsequent versions
  - Continual improvement
    - 'Web-based software and services (perennial beta)

Evaluation Paradigms

- Quick and Dirty Evaluation (Discount usability)
  - With users
  - With experts
- Usability Testing
- Field Studies
  - Site-based
  - Web-based

Expert Evaluation

- Usability Inspection
  - Heuristic Evaluation
  - Cognitive Walkthrough
  - Feature Inspection
  - Consistency Inspection
  - Standards Inspection

Nielsen and Mack, *Usability Inspection Methods*

Log-based evaluation

- Seeing what users actually do
  - Eudora, Microsoft,...
- A/B tests for testing alternatives
  - Google, Amazon, Microsoft..

Controlled Usability Tests

- How do you set them up and run them?
- What can you learn?

How do you do a User Study?

- Setting
- Artifacts
- Goals
- Logistics
- Framing
- Tasks
- Capture
- Analysis

Setting – Where do you do it?

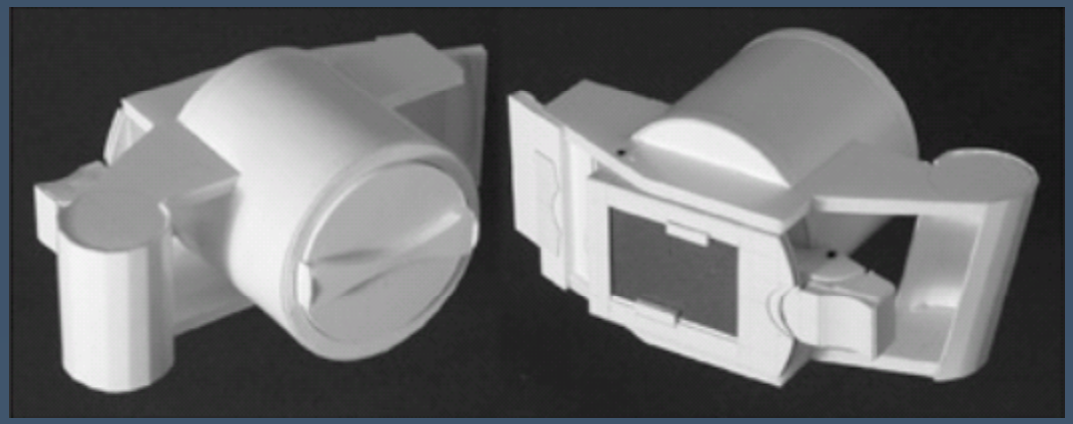
- Usability laboratory
- Neutral setting
- User's setting
- Actual use setting

Artifacts – What will they be working with?

- Representations (e.g. sketches)
- Mockups at various levels of detail
- Working prototypes
  - Physical prototype
  - Interaction prototype

Before you start, run through the full test yourselves to be sure all the relevant pieces are there and working

Physical Mockups and Pr[illegible]es



Goals – What are you trying to learn from the test?

- Goals and questions should guide all evaluation studies
  - Problem spotting
  - Comparison of alternatives
  - General assessments
- What's important for this project at this time?

Logistics – How do you treat the user?

- Mechanics of the test setting, enrollment, welcoming, etc.
  - Laboratory vs. informal
- Permissions
  - Privacy – Use of captured data
  - Identity - Photos, etc.
  - Human Subjects approval
- Quid pro quo
  - Payments, friendship, ...

Framing - What does the experience mean to the user?

- Why you are doing this
- Who/what is being tested
  - NOT the intelligence of the user
  - NOT their response to you
- How will data be used
  - The feeling of being watched and assessed

To minimize distortions, try to think of the situation from the user's point of view

Tasks – What is the user asked to do?

- Scripted tasks
  - Needed for incomplete prototypes
  - Valuable for cross-comparison
    - *"Add CS160 to the list of courses and see if it conflicts with anything"*
- Open-ended tasks
  - Can be in speculative or operational mode
    - *"What would you expect this screen to let you do?"*
    - *"Try browsing some pages about ..."*
- Naturalistic
  - Requires thorough prototype
    - *"Try doing your regular email"*

Script Advice

- Have a clear script for all the tasks before each test.
- Choose script tasks that cover the functionality you are interested and the questions you want the test to aswer
- Run through the script yourself before the test.
- You can revise between tests if you aren't doing quantitative comparisons.

Capture – What can you observe?

- User actions
  - In the system (loggable)
  - Physical (video)
- Overt comments
  - Spontaneous
  - In response to questions
- Think-aloud (individual or pair)

Use capture technology appropriately – decide what you will learn from audio or video recordings, system logs, notes, etc. and whether they are justified for this test.

Analysis – When do you do what?

- In-session notes
- Debrief and interpretive notes
- Review of capture (e.g., video)
- Formal (quantitative) analysis

Always do an interpretive debrief as soon after the session as possible, with more than one person. You won't remember as much as you think you will.

Analysis - What can you learn from the results?

- Quantitative
  - Measurable items
    - Usage, Efficiency, Subjective satisfaction ...
  - External vs. internal validity
  - Statistical significance
- Qualitative
  - Identify problem areas and priorities
  - Suggest possibilities
  - Material for presentations

Analysis - What is hard to learn

- How will people really use it in the long run
  - How do users learn and adapt?
  - What is the actual utility
- What will happen with technologies with network effects
  - Until a lot of people use it it doesn't pay off
- How does your test generalize
  - Real user group may not be amenable to testing
  - High variability in tasks, users, interactions,...

When have you tested enough?

- HCI research vs. product development
  - Generalizability vs. specficity
- Resource limitations
  - Testing costs
  - Time to market
    - (assignment deadlines!)
- Diminishing returns
  - Pareto Principle