Evaluation 2

John Stasko

Spring 2007

This material has been developed by Georgia Tech HCI faculty, and continues to evolve. Contributors include Gregory Abowd, Al Badre, Jim Foley, Elizabeth Mynatt, Jeff Pierce, Colin Potts, Chris Shaw, John Stasko, and Bruce Walker. Permission is granted to use with acknowledgement for non-profit purposes. Last revision: January 2007.

Agenda (for 3 evaluation lectures)

- Evaluation overview
- Designing an experiment
  - Hypotheses
  - Variables
  - Designs & paradigms
- Participants, IRB, & ethics
- Gathering data
  - Objective; Subjective data
- Analyzing & interpreting results
- Using the results in your design

Evaluation, Day 2

- Data collection methods, techniques
- Objective data
- Subjective Data
- Quantitative
  - Surveys, Questionnaires
- Qualitative
  - Open-ended questions, interviews

Evaluation is Detective Work

- Goal: gather evidence that can help you determine whether your hypotheses are correct or not.
- Evidence (data) should be:
  - Relevant
  - Diagnostic
  - Credible
  - Corroborated

Data as Evidence

- Relevant
  - Appropriate to address the hypotheses
    - e.g., Does measuring "number of errors" provide insight into how effectively your new air traffic control system supports the users' tasks?
- Diagnostic
  - Data unambiguously provide evidence one way or the other
    - e.g., Does asking the users' preferences clearly tell you if the system performs better? (Maybe)

Data as Evidence

- Credible
  - Are the data trustworthy?
    - Gather data carefully; gather enough data
- Corroborated
  - Do more than one source of evidence support the hypotheses?
    - e.g., Both accuracy and user opinions indicate that the new system is better than the previous system. But what if completion time is slower?

General Recommendations

- Include both objective & subjective data
  - e.g., “completion time” and “preference”
- Use multiple measures, within a type
  - e.g., “reaction time” and “accuracy”
- Use quantitative measures where possible
  - e.g., preference score (on a scale of 1-7)

  Note: Only gather the data required; do so with the min. interruption, hassle, time, etc.

Types of Data to Collect

- “Demographics”
  - Info about the participant, used for grouping or for correlation with other measures
    - e.g., handedness; age; first/best language; SAT score
    - Note: Gather if it is relevant. Does not have to be self-reported: you can use tests (e.g.,Edinburgh Handedness)
- Quantitative data
  - What you measure
    - e.g., reaction time; number of yawns
- Qualitative data
  - Descriptions, observations that are not quantified
    - e.g., different ways of holding the mouse; approaches to solving problem; trouble understanding the instructions

Planning for Data Collection

- What data to gather?
  - Depends on the task and any benchmarks
- How to gather the data?
  - Interpretive, natural, empirical, predictive??
- What criteria are important?
  - Success on the task? Score? Satisfaction?…
- What resources are available?
  - Participants, prototype, evaluators, facilities, team knowledge (programming, stats, etc.)

Collecting Data

- Capturing the Session
  - Observation & Note-taking
  - Audio and video recording
  - Instrumented user interface
  - Software logs
  - Think-aloud protocol - can be very helpful
  - Critical incident logging - positive & negative
  - User journals
- Post-session activities
  - Structured interviews; debriefing
    - "What did you like best/least?"; "How would you change..?"
  - Questionnaires, comments, and rating scales
  - Post-hoc video coding/rating by experimenter

Observing Users

- Not as easy as you think
- One of the best ways to gather feedback about your interface
- Watch, listen and learn as a person interacts with your system
- Preferable to have it done by others than developers
  - Keep developers in background

Observation

- Direct
  - In same room
  - Can be intrusive
  - Users aware of your presence
  - Only see it one time
  - May use 1-way mirror to reduce intrusion
  - Cheap, quicker to set up and to analyze
- Indirect
  - Video recording
  - Reduces intrusion, but doesn't eliminate it
  - Cameras focused on screen, face & keyboard
  - Gives archival record, but can spend a lot of time reviewing it

Location

- Observations may be
  - In lab - Maybe a specially built usability lab
    - Easier to control
    - Can have user complete set of tasks
  - In field
    - Watch their everyday actions
    - More realistic
    - Harder to control other factors

Challenge

- In simple observation, you observe actions but don't know what's going on in their head
- Often utilize some form of *verbal protocol* where users describe their thoughts

Verbal Protocol

- One technique: *Think-aloud*
  - – User describes verbally what s/he is thinking while performing the tasks
    - What they believe is happening
    - Why they take an action
    - What they are trying to do

Think Aloud

- Very widely used, useful technique
- Allows you to understand user's thought processes better
- Potential problems:
  - – Can be awkward for participant
  - – Thinking aloud can modify way user performs task

Teams

- Another technique: *Co-discovery learning* (Constructive interaction)
  - – Join pairs of participants to work together
  - – Use think aloud
  - – Perhaps have one person be semi-expert (coach) and one be novice
  - – More natural (like conversation) so removes some awkwardness of individual think aloud

Alternative

- What if thinking aloud during session will be too disruptive?
- Can use *post-event protocol*
  - – User performs session, then watches video and describes what s/he was thinking
  - – Sometimes difficult to recall
  - – Opens up door of interpretation

Historical Record

- In observing users, how do you capture events in the session for later analysis?
  - ?

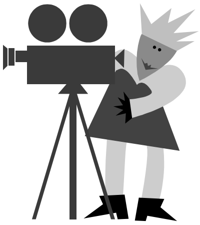
Capturing a Session

1. Paper & pencil
   - Can be slow
   - May miss things
   - Is definitely cheap and easy

		Task 1	Task 2	Task 3	…
Time	10:00		S		
	10:03		e	S	
	10:08			e	
	10:22				

Capturing a Session

2. Recording (audio and/or video)
   - Good for talk-aloud
   - Hard to tie to interface
   - Multiple cameras probably needed
   - Good, rich record of session
   - Can be intrusive
   - Can be painful to transcribe and analyze



Sun Microsystems Usability Lab

Observation Room

Large viewing area in this one-way mirror which includes an angled sheet of glass the improves light capture and prevents sound transmission between rooms.

Doors for participant room and observation rooms are located such that participants are unaware of observers movements in and out of the observation room.

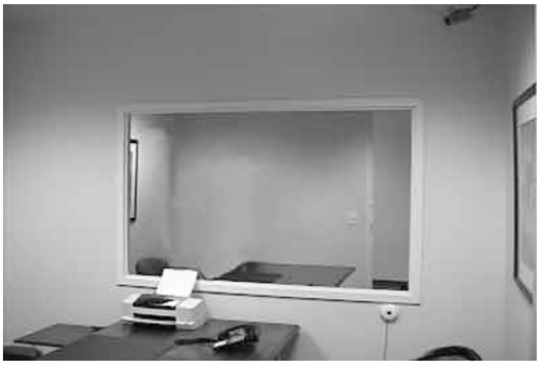
http://www.surgeworks.com/services/observation_room2.htm

Usability Lab - Observation Room

- State-of-the-art observation room equipped with three monitors to view participant, participant's monitor, and composite picture in picture.
- One-way mirror plus angled glass captures light and isolates sound between rooms.
- Comfortable and spacious for three people, but room enough for six seated observers.
- Digital mixer for unlimited mixing of input images and recording to VHS, SVHS, or MiniDV recorders.

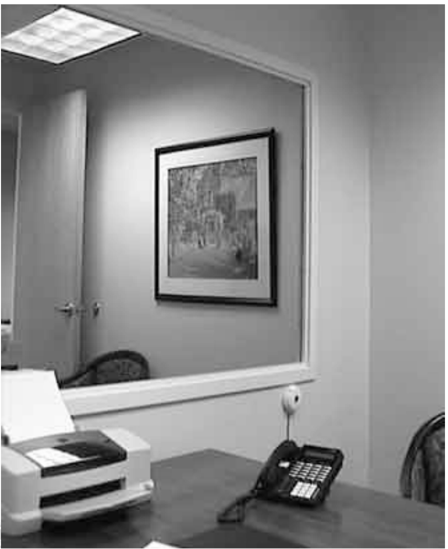
Usability Lab - Participant Room

- Sound proof participant room with a feel similar to a standard office environment.
- Pan-tilt-zoom high resolution digital camera (visible in upper right corner).
- Microphone pickup can be moved near participant or left in location, which is just below right side of observation window.
- Observation room door not visible by participants from reception/waiting area. Participants unaware of people entering or leaving observation room.



Usability Lab - Participant Room

- Note the half-silvered mirror



Capturing a Session

3. Software logging
   - Modify software to log user actions
   - Can give time-stamped keypress or mouse event
     - Synch with video
   - Commercial software available
   - Two problems:
     - Too low-level, want higher level events
     - Massive amount of data, need analysis tools

Issues

- What if user gets stuck on a task?
- You can ask
  - "What are you trying to do..?"
  - "What made you think..?"
  - "How would you like to perform..?"
  - "What would make this easier to accomplish..?"
  - Maybe offer hints
- Can provide design ideas

Subjective Data

- Satisfaction is an important factor in performance over time
- Learning what people prefer is valuable data to gather

Methods

- Ways of gathering subjective data
  - Questionnaires
  - Interviews
  - Booths (e.g., trade show)
  - Call-in product hot-line
  - Field support workers
- (Focus on first two)

Questionnaires

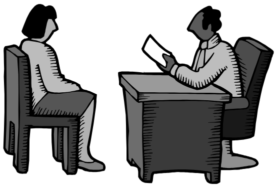
- Preparation is expensive, but administration is cheap
- Oral vs. written
  - Oral advs: Can ask follow-up questions
  - Oral disadvs: Costly, time-consuming
- Forms can provide more quantitative data

Questionnaires

- Issues
  - Only as good as questions you ask
  - Establish purpose of questionnaire
  - Don’t ask things that you will not use
  - Who is your audience?
  - How do you deliver and collect questionnaire?

Questionnaire Topic

- Can gather demographic data and data about the interface being studied
- Demographic data:
  - Age, gender
  - Task expertise
  - Motivation
  - Frequency of use
  - Education/literacy



Interface Data

- Can gather data about
  - screen
  - graphic design
  - terminology
  - capabilities
  - learning
  - overall impression
  - ...

Question Format

- Closed format
  - – Answer restricted to a set of choices
  - – Typically very quantifiable
  - – Variety of styles

Closed Format

- Likert Scale
  - – Typical scale uses 5, 7 or 9 choices
  - – Above that is hard to discern
  - – Doing an odd number gives the neutral choice in the middle
  - – You may not want to give a neutral option

Characters on screen were:

hard to read easy to read
1 2 3 4 5 6 7

Other Styles

Which word processing systems do you use?

- □ LaTeX
- □ Word
- □ FrameMaker
- □ WordPerfect

Rank from
1 - Very helpful
2 - Ambivalent
3 - Not helpful
0 - Unused

___ Tutorial
___ On-line help
___ Documentation

Closed Format

- Advantages
  - Clarify alternatives
  - Easily quantifiable
  - Eliminate useless answer
- Disadvantages
  - Must cover whole range
  - All should be equally likely
  - Don't get interesting, "different" reactions

Open Format

- Asks for unprompted opinions
- Good for general, subjective information, but difficult to analyze rigorously
- May help with design ideas
  - “Can you suggest improvements to this interface?”

Questionnaire Issues

- Question specificity
  - “Do you have a computer?”
- Language
  - Beware terminology, jargon
- Clarity
  - “How effective was the system?” (ambiguous)
- Leading questions
  - Can be phrased either positive or negative

Questionnaire Issues

- Prestige bias - (British sex survey)
  - People answer a certain way because they want you to think that way about them
- Embarrassing questions
  - “What did you have the most problem with?”
- Hypothetical questions
- “Halo effect”
  - When estimate of one feature affects estimate of another (eg, intelligence/looks)
  - Aesthetics & usability, one example in HCI

Deployment

- Steps
  - Discuss questions among team
  - Administer verbally/written to a few people (pilot). Verbally query about thoughts on questions
  - Administer final test
  - Use computer-based input if possible
  - Have data pre-processed, sorted, set up for later analysis at the time it is collected

Interviews

- Get user's viewpoint directly, but certainly a subjective view
- Advantages:
  - Can vary level of detail as issue arises
  - Good for more exploratory type questions which may lead to helpful, constructive suggestions

Interviews

- Disadvantages
  - Subjective view
  - Interviewer(s) can bias the interview
  - Problem of inter-rater or inter-experimenter reliability (a stats term meaning agreement)
  - User may not appropriately characterize usage
  - Time-consuming
  - Hard to quantify

Interview Process

- How to be effective
  - Plan a set of questions (provides for some consistency)
  - Don’t ask leading questions
    - “Did you think the use of an icon there was really good?”
- Can be done in groups
  - Get consensus, get lively discussion going

HW 3

- Airline speech interfaces
- Some observations…

Upcoming

- Data inspection & analysis
  - – Feedback into the design
- Universal Design
  - – P3 due on Tuesday