ICS 667
Advanced HCI Design Methods

09. Empirical Evaluation

Dan Suthers
Spring 2005

Methods Continued

• Analytic Evaluation
  - “From expert” or theory driven
  - Models, Metrics, Heuristics
• Empirical Evaluation (Testing)
  - “From user”
  - Subjective: user opinions
  - Performance: observing activity, testing
Some Empirical Methods

- Collecting users’ opinions
  - attitudes
- Interpretive evaluation
  - how used in natural settings (ecological validity)
- Usability Testing
  - how users interact with the system (in detail)
- Experiments
  - hypothesis testing

Subjective Methods (Asking Users)
Subjective Methods

Caveat: "First rule of usability: don't listen to users!" (Watch what they do)

Two major methods

- **Interviews** - qualitative analysis
- **Surveys** - quantitative analysis

**Interviews**

- **Unstructured**
  - No set questions or sequence
  - Rich results
  - May miss information you need; not replicable
- **Structured**
  - Scripted (fixed questions in sequence)
  - Easier to conduct and analyze; replicable
  - May miss opportunistic information
- **Semi-structured**
  - Specific and open ended questions (will discuss two ways to do this)
Basic of Interviewing

• Goals and questions guide all interviews
• Preparation should include
  - Informed consent and procedure for anonymity
  - Checking recording equipment in advance
  - Questions!
• Two types of questions:
  - Closed: predetermined answer format, e.g., ‘yes,’ ‘no,’
    ‘twice a day,’ ‘OS 10’ …
  - Open
  - Closed questions are quicker and easier to analyze
• Avoid
  - Long or complex questions
  - Jargon
  - Leading questions

Organization of an Interview

• Introduction - introduce yourself, explain the
goals of the interview, reassure about the ethical
issues, ask to record, present an informed consent
form.
• Warm-up - make first questions easy & non-
threatening.
• Main body - present questions in a logical order
• A cool-off period - include a few easy questions to
defuse tension at the end
• Closure - thank interviewee, signal the end,
e.g, switch recorder off.
Focus Groups

- Group interviews
- Typically 3-10 participants
- Provide a diverse range of opinions
- Can get synergy between participants
- Need to be managed to:
  - ensure everyone contributes
  - discussion isn’t dominated by one person
  - the agenda of topics is covered

Analyzing interview data

Depends on the type of interview
- Structured interviews can be analyzed like questionnaires (quantitatively)
- Unstructured interviews generate data like that from participant observation
  - Qualitative analysis, data driven generation of categories
- Analyze unstructured interviews as soon as possible
**Questionnaires and Surveys**

- Can reach large populations (paper, email, web)
- Results can go direct to database
- Usually analyzed quantitatively
  - Open questions are hard to analyze
  - Closed questions can be automated but limit responses
- **Design with your analysis in mind**
- Piloting important
- Some types of closed questions and their uses
  - **Checklists**: categorical or background information
  - **Likert scales**: range of agreement or disagreement with a statement
  - **Ranked order**: e.g., rank in order of usefulness
  - **Semantic Differential**: e.g., “Attractive …. Ugly”

---

**Developing a questionnaire**

- Clear statement of purpose & guarantee participants anonymity
- Decide on whether phrases will all be positive, all negative or mixed
- Pilot test questions: are they clear; is there sufficient space for responses
- Decide how data will be analyzed & consult a statistician if necessary
Encouraging responses

- Offer a short version for those who do not have time to complete a long questionnaire
- If mailed, include a self addressed stamped envelope
- Provide an incentive
- 40% response rate is high, 20% is often acceptable
- Follow-up with emails, phone calls, letters
- Ask whether they are willing to be interviewed

Online Questionnaires

- Pros
  - Responses are usually received quickly
  - No copying and postage costs
  - Data can be collected in database for analysis
  - Time required for data analysis is reduced
  - Errors can be corrected easily
- Cons
  - Sampling problematic if population size unknown
  - How to prevent individuals from responding more than once?
  - May change questions in email
Objective Methods: Introduction

Objective Methods

- Observing and monitoring use of artifact
  - in laboratory
  - in natural setting
    - how users interact with system
    - how system interacts with context
    - usability issues
- Useful at any phase of development
**Direct Observation**

- Researcher watches use, takes notes
- **Hawthorne Effect** (users act differently under observation) may contaminate results
- Record may be incomplete
- Only one chance
- Helpful to have *shorthand and/or forms* which you are fluent

**Indirect Observation**

**Video logging**
- User(s) body language, gestures
- Screen activity
- Two uses:
  - *Exploratory evaluation*: review tapes carefully and repeatedly to discover issues
  - *Formal studies*: know what you are looking for!

**Interaction logging** (software)
- Often use two or more together
- Must synchronize all data streams
- High volume of data can be overwhelming
Frameworks to guide observation

• The Goetz and LeCompte (1984) framework:
  - Who is present?
  - What is their role?
  - What is happening?
  - When does the activity occur?
  - Where is it happening?
  - Why is it happening?
  - How is the activity organized?

The Robinson (1993) framework

• Space. What is the physical space like?
• Actors. Who is involved?
• Activities. What are they doing?
• Objects. What objects are present?
• Acts. What are individuals doing?
• Events. What kind of event is it?
• Goals. What do they to accomplish?
• Feelings. What is the mood of the group and of individuals?
Planning observations

- Goals & questions
- Which framework & techniques
- How to collect data
- Which equipment to use
- How to gain acceptance
- How to handle sensitive issues
- Whether and how to involve informants
- How to analyze the data
- Whether to triangulate

Data Collection Techniques

- Notes
- Audio
- Still Camera
- Video
- Tracking users:
  - diaries
  - interaction logging
Data Analysis

• Qualitative data - interpreted & used to tell the 'story' about what was observed.
• Qualitative data - categorized using techniques such as content analysis.
• Quantitative data - collected from interaction & video logs. Presented as values, tables, charts, graphs and treated statistically.

Verbal Protocols

• Audio record of spoken language
  - Spontaneous utterances
  - Conversation between multiple users
  - Think-aloud protocol
  - Post-event protocols
• Dangers of introspection, rationalization
• Analyze along with video
**Video/Verbal Analysis**

- Diversity of approaches
- **Task-based**
  - how do users approach the problem
  - difficulties in using the software
  - need not be exhaustive: identify interesting cases
- **Performance-based**
  - frequency and timing of categories of actions, errors, task completion

- Again, time consuming: usability studies often try to do this in real time, use video as backup

---

**More on Analysis of Video/Verbal**

- Requires *classification scheme*, invented or borrowed
- May involve *inter-rater reliability*
- *Often exhaustive and time intensive!*
- **Tools important**
  - Transcribing conversation to text merged with transaction log is tedious
  - Better approach: direct analysis of digital video
  - For basic usability it is best to become skilled at coding "on the fly" using paper forms
Software Instrumentation/Logging

- Time-stamped logs
  - key-presses or higher level actions
  - record what happened but not replayable
- Interaction logging
  - replayable
- Synchronize with video data for rich but overwhelming data
- Analysis issues are similar

Interpretive Evaluation

- Trend away from experiments ...
  - Laboratory too artificial
  - Experimental tasks too artificial
  - Cannot control all variables
  - Not valuing user’s ideas
- ... towards subjective evaluation
  - Researcher immerses in work context
  - Users participate in setting objectives, carrying out and interpreting evaluation
- ... accompanied by shift in world view
  - Reality is subjective
**Interpretive Data Analysis**

- Look for
  - key events that drive the group’s activity
  - patterns of behavior
- Triangulate data sources against each other
- Report findings in a convincing and honest way
- Produce ‘rich’ or ‘thick descriptions’
- Include quotes, pictures, and anecdotes
- Software tools can be useful e.g., Morae, NUDIST, NVivo, Observer ...

**Contextual Inquiry**

- **Evaluate in the user’s normal working environment**
  - Genuine work materials, e.g. documents
  - Realistic time frame and organization of work in time
  - Typical population members
  - Representative tasks
  - Shared control of situation
Participative Evaluation

- A natural extension of participatory design
- Users participate in and guide the evaluation
- Establish groups with representatives from the whole range of users who collaborate on the design (which is viewed as a mutual learning process)
- Provide prototypes that are sufficiently robust for users to evaluate
- Encourage focus on coupling between technical questions and social and political issues in the workplace

Ethnography

- From Anthropology and Sociology
- Researcher immerses in situation
- Role is to learn about participants (users) from their point of view
- Wide range of methods and data sources
- Video plays an important role in HCI ethnography
- Participants may assist in interpretation
Performance Testing

Experiments and Usability Tests: Observations in settings you define

User testing vs. Experiments

**User testing**
- Aim: improve products
- Few participants
- Results inform design
- Not perfectly replicable
- Controlled conditions
- Procedure planned
- Results reported to developers

**Research experiments**
- Aim: discover knowledge
- Many participants
- Results validated statistically
- Replicable
- Strongly controlled conditions
- Experimental design
- Results reported to scientific community
Experiments

• Decide why you are doing it
• Pick a testable hypothesis
• Define a method
  - Subjects
  - Materials
  - Procedures and Instructions
• Choose statistical tests
• Interpret results

Variables

• Independent variable
  - Hypothesized causal factor
  - What you modify
    - “Input”
    - Vary one at a time for interpretable results*
• Dependent variable
  - Hypothesized effect
  - What you measure
    - “Output”

*more complex designs can vary more than one at a time
Subjects or Participants

- Must balance for ...
  - Age
  - Gender
  - Prior experience
  - Aptitude
- Consider incentive to participate
- Obtain informed consent
  - Aware of risks and benefits
  - Option to quit at any time

Experimental Designs at a glance

- Between Subjects: each experimental condition has different subjects
- Within Subjects: each condition has the same subjects
- One-way, 2x2, etc.
Between Subjects Designs

• Good for ensuring no cross-treatment effects
• Require more subjects
• Issues of whether the groups are the same.
• Two ways to make the groups the same:
  • Independent subject design: randomly assign to experimental conditions
  • Matched subject design: design the groups to be similar by matching subjects

Within subjects designs

• Pre-post test: measure before & after treatment
  - Problem: confounds with time on task
• Repeated Measures: each subject tested in both experimental conditions
  - Problem: TOT and cross-condition effects
  - Solution: counterbalance order in which conditions given
    • Half the subjects: Condition A, Condition B
    • Other half: Condition B, Condition A
  - Can get complicated with more conditions
2x2 designs

- Condition A, Condition 1
- Condition A, Condition 2
- Condition B, Condition 1
- Condition B, Condition 2

- Good for finding interactions between two variables (e.g., using ANOVA)
- Balancing order of treatment gets complex

Critiquing Results

Not only statistical significance, but also:
- Is size of effect meaningful?
- Are there alternative explanations?
- Are the results consistent? Compare dependent variables
- How general are the results?
Usability Testing

- Can be done in laboratory or field
- Purpose is to identify as many usability defects as possible, maximizing information gained from each participant
- In some aspects, preparation is similar to experiments (but you only run a few and they are not controlled)
- Pilot tests are critical
### Selecting Participants

- Representative of user groups
- Volunteers: Don’t coerce
- May need proxy in some situations (real users too busy, inaccessible)
- Also include usability expert
- Three to five is adequate per iteration
- Partial overlap of participants between iterations of evaluation (about half new, half repeats)

### Selecting Tasks

- **UCD**: Include focal use cases
- **SBD**: Include subtasks that relate to critical claims
- **Mediated evaluation**: include tasks for which analytic evaluation suggest problems
- These should be tasks for which there are usability specifications
- Also include other representative tasks and use of documentation if of interest
Task Script

- State what the user does, but not how
- Best to provide written instructions but have users read them out loud
  - This helps verify the participant understood it
  - Tells you when to start the timer
- Write two scripts: one for the evaluator (annotated), and one for the user

Procedures

- If laboratory, show participant around
- Introduction to study
  - Introduce the system being evaluated
  - Purpose is to evaluate the system, not the participant
- Informed Consent form
- Run the session
  - Prompt participant to think aloud
  - Give only minimal hints or help
- Debriefing
- (We’ll have an example)
Common Measures (Wilson & Wixon, '97)

- Time to complete a task
- Time to complete a task after a specified time away from the product
- Number and type of errors per task
- Number of errors per unit of time
- Number of navigations to online help or manuals
- Number of users making a particular error
- Number of users completing task successfully
- Reduction in errors over multiple trials

A modestly priced usability lab

- Video over participant’s shoulders inset in live video of screen capture
Let's Try an Evaluation

- Choose 4 “participants”
- Evaluation teams (3 people minimum)
  - Lead evaluator directs and gives hints
  - Time recorder records time for each task
  - Error recorder records attempts to do task that fail, other slips
  - (normally 1 skilled person could do all of this)
- Participants leave room while I explain

Handouts and forms

- Instructions - participant reads
- Consent form - participant reads and signs
- Participant’s Script - participant reads each line aloud and does the task
- Evaluator’s Script - as above but has hints in case participant is stuck
- Sample Diagram - for giving help (show it if needed)
- Data Forms - two, for timer and error counter
- Benchmark form - to combine your data at the end
  ... Try the software now ...
Running the session

- Participant reads instructions
- Participant reads consent form
- Any questions?
- Sign
- Run the session
  - Participant reads each line of script aloud and does the task
  - Timer times from end of reading to completion of task
  - Error counter watches for false starts and slips
  - Lead evaluator coordinates and observes, gives hints or shows sample if needed
- At end
  - Debrief participant
  - Use benchmark form to combine data

Discussion

- What potential usability defects did you uncover?
- How could our procedure be improved?
- What did you learn about usability testing?
- How does it compare to heuristic evaluation? collaborative walkthroughs?