Methodology Overview

Why do we evaluate in HCI?
Why should we use different methods?
How can we compare methods?
What methods are there?

Why Do We Evaluate In HCI?

Pre-design
• what do people do?
• evaluation produces
  - key tasks and required functionality
  - work practices
  - organizational practices
  - user type …
• used to understand system functionality
Why Do We Evaluate In HCI?

During initial design ideas and iterative development

_initial stages_
- evaluate choices of initial design ideas and representations
  - is the representation appropriate?
  - does it reflect how people think of their task

_iterative development_
- fine tune the interface, looking for usability bugs
  - can people use this system?

_Evaluation produces:_
- user reaction to design
- validation and list of problem areas (bugs)
- new design ideas

---

Why Do We Evaluate In HCI?

_Post-design_
- _acceptance test:_ did we deliver what we said we would?
  - verify that human/computer system meets expected performance criteria
  - ease of learning, usability, user’s attitude, performance criteria
  - e.g., a first time user will take 1-3 minutes to learn how to withdraw $50. from the automatic teller
- _revisions:_ what do we need to change?
- _effects:_ What did we change in the way people do their tasks?

_evaluation produces_
- testable usability metrics
- actual reactions
- validation and list of problem areas (bugs)
- changes in original work practices/requirements
Why Do We Evaluate In HCI?

Generalized knowledge
• are there general design principles?
• are there theories?
• Can we validate ideas / visions / hypotheses?
• evaluation produces:
  - principles and guidelines
  - evidence supporting/rejection theories

Why Should We Use Different Methods?

Method definition (McGrath)
• tools for gathering and analyzing information

All methods:
• are valuable in certain situations
• have weaknesses and limitations
• can be used to complement each other
Why Should We Use Different Methods?

Information requirements differ
• pre-design, iterative design, post-design, generalizable knowledge…

Information produced differs
• outputs should match the particular problem/needs

Cost/benefit of using method
• cost of method should match the benefit gained from the result,

One method’s strength can complement another’s weakness
• no one method can address all situations

Constraints
• may force you to chose quick and dirty discount usability methods

How Can We Compare Methods?

Relevance
• does the method provide information to our question / problem?

Naturalistic:
• is the method applied in an ecologically valid situation?
  - observations reflect real world settings: real environment, real tasks, real people, real motivation

Generalization
• how well can I generalize the information produced to other situations?

Repeatability
• would the same results be achieved if the test were repeated?
How Can We Compare Methods?

Validity
- *External validity:* can the results be applied to other situations?
- *Internal validity:* do we have confidence in our explanation?

*Does the test measure something of relevance to usability of real products in real use outside of lab?*
- Some typical reliability problems of testing vs real use
  - non-typical users tested
  - tasks are not typical tasks
  - physical environment different
    - quiet lab vs very noisy open offices vs interruptions
  - social influences different
    - motivation towards experimenter vs motivation towards boss

*Partial Solution for external validity*
- use real users
- tasks from task-centered system design
- environment similar to real situation

How Can We Compare Methods?

Quickness
- can I do a good job with this method within my time constraints?

Cost
- Is the cost of using this method reasonable for my question?

Equipment
- What special equipment / resources required?

Personnel, training and expertise
- What people are required to run this method? What expertise must they have?

Subject selection
- how many subjects do I need, who do they have to be, and can I get them?
How Can We Compare Methods?

Type of information (qualitative vs quantitative)
• is the information quantitative and amenable to statistical analysis?

Comparative
• can I use it to compare different things?

Control
• can I control for certain factors, so that I can see what effects they have?

Scope of subjects
• is it good for analyzing individuals? small groups? organizations?

Cross-sectional or Longitudinal
• can it reveal changes over time?

How Can We Compare Methods?

Setting
• field vs laboratory?

Support
• are there tools for supporting the method and analyzing the data?

Routine application
• is there a fairly standard way to apply the method to many situations

Theoretic
• is there a theoretic basis behind the method?

Result type
• does it produce description or explanation
How Can We Compare Methods?

Metrics
• are there useful, observable phenomena that can be measured

Measures
• can I see processes or outcomes

Organizational
• can they be included within an organization as part of a software development process

What methods are there?

Lab tests
• Experimental methodologies
  - highly controlled observations and measurements to answer very specific questions i.e., hypothesis testing
• Usability testing
  - mostly qualitative, less controlled observations of users performing tasks

Interface inspection
• Usability heuristics
  - several experts analyze an interface against a handful of principles
• Walkthroughs
  - experts and others analyze an interface by considering what a user would have to do a step at a time while performing their task
What methods are there?

Field studies
- Ethnography
  - field worker immerses themselves in a culture to understand what that culture is doing
- Contextual inquiry
  - interview methodology that gains knowledge of what people do in their real-world context

Cognitive modelling
- Fitt’s Law
  - mathematical expression that can predict a user’s time to select a target
- Keystroke-level model
  - low-level description of what users would have to do to perform a task that can be used to predict how long it would take them to do it
- Goms
  - structured, multi-level description of what users would have to do to perform a task that can also be used to predict time

Self reporting
- questionnaires
- surveys
Goals of Behavioural Evaluation

Designer:
• user-centered iterative design

Customer
• selecting among systems

Manager
• assisting effectiveness

Marketer
• building a case for the product

Researcher
• developing a knowledge base

(From Finholt & Olsons CSCW 96 Tutorial)