ICS 463
Human Computer Interaction

10. Introduction to Evaluation (Part I)

Dan Suthers
Spring 2004

Why you need to evaluate

"Iterative design, with its repeating cycle of design and testing, is the only validated methodology in existence that will consistently produce successful results. If you don’t have user-testing as an integral part of your design process you are going to throw buckets of money down the drain.” -- Bruce Tognazzini, AskTog.com

Framework for Evaluation

• All evaluation requires
  - Something to be measured (a source of data)
    - Design or system itself
    - Expert posing as user (popular for cost-cutting)
    - User in lab
    - User in natural setting (trend towards this)
  - Something to judge the measurements against
    - Model
    - Heuristics
    - Benchmarks
    - Hermeneutics (interpretation)

• Most common combinations:
  - Design evaluated against model (simulate user)
  - Expert applies Heuristics
  - User testing generates data compared to Benchmark
  - Interpretation of user in natural setting
Evaluation Concepts

- **Validity**: does it measure what it is supposed to measure?
- **Ecological Validity**: does it represent what happens in the real world?
- **Reliability**: how replicable are the results in similar contexts?
- **Scope**: how replicable are the results in different contexts?
- **Type I error**: seeing something that is not there
- **Type II error**: failing to see something that is there

Evaluation in Design Lifecycle

**Formative Evaluation**: feedback that informs design
- **Early**
  - Understanding target application
  - checking understanding of requirements
  - quick filtering of ideas
- **Middle**
  - predicting usability
  - comparing alternate designs
  - engineering towards a usability target

**Summative Evaluation**: have we done well?
- **Late**
  - fine tuning of usability
  - verifying conformance to a standard

Preview of Methods of Evaluation

... and what they get you ...
- **Collecting users’ opinions**
  - attitudes
- **Observing and monitoring use**
  - how users interact
- **Experiments**
  - hypothesis testing
- **Interpretive evaluation**
  - how used in natural settings (ecological validity)
- **Predictive evaluation**
  - anticipated usability issues

Here are the major approaches outlined by your text...
“Quick and Dirty”

- Common practice in which designers informally get feedback from users or consultants to confirm that their ideas are in-line with users’ needs and are liked.
- Can be done at any time.
- Emphasis on fast input to the design process rather than carefully documented findings.

Usability Testing

- Users’ performance on typical tasks are recorded in controlled settings.
- Field observations may also be used.
- Performance is watched, recorded on video, and/or logged by software.
- This data is used to calculate performance times, identify errors, and help explain why the users did what they did.
- User satisfaction questionnaires and interviews are used to elicit users’ opinions.

Field Studies

- Undertaken in natural settings
- Aim: understand what users do naturally and how technology impacts them.
- Uses:
  - identify opportunities for new technology
  - determine design requirements
  - decide how best to introduce new technology
  - evaluate technology in use
Predictive Evaluation

- Experts apply their knowledge of typical users, often guided by heuristics, to predict usability problems.
- Another approach involves theoretically based models.
- A key feature of predictive evaluation is that users need not be present.
- Relatively quick and inexpensive.

Things we might measure

- **Analytical**
  - Predicted number of keystrokes needed (GOMS)
  - Number of errors relative to theoretical maximum
  - Number of commands as task that are hidden versus visible
  - Complexity and organization of interface
- **Performance**
  - Time to complete task
  - Number or percent of tasks completed per unit time
  - Number of errors per unit time
  - Time required to reach task criterion or error rate
  - Rate of use of help system
  - Quality of task product
- **Subjective ("Psychometric")**
  - User's attitude towards the system
  - Perception of efficiency
  - Perception of helpfulness
  - Perception of control
  - Perception of learnability

Consider also group level measurements.

Issues to consider

**Evaluation planning should consider**

- Characteristics of users: representative of target population? Gender, experience, culture ...
- Types of activities to be observed or tested
- Environment of use: laboratory or natural setting
- Nature of artifact to be evaluated
Pilot Study

- A small trial run of the main study.
- The aim is to make sure your plan is viable.
- Pilot studies check:
  - that you can conduct the procedure
  - that interview scripts, questionnaires, experiments, etc. work appropriately
- It's worth doing several to iron out problems before doing the main study.
- Ask colleagues if you can't spare real users.

---

How to DECIDE?

_A guide for planning evaluation_

---

**DECIDE:** A framework to guide evaluation

- **Determine** the goals the evaluation addresses.
- **Explore** the specific questions to be answered.
- **Choose** the _evaluation paradigm_ and _techniques_ to answer the questions.
- **Identify** the _practical issues_.
- **Decide** how to deal with the _ethical issues_.
- **Evaluate**, interpret and present the _data_.

---
Determine the Goals

- What are the high-level goals of the evaluation?
- Who wants it and why?
- The goals influence the paradigm for the study
- Some examples of goals:
  - Identify the best metaphor on which to base the design.
  - Check to ensure that the final interface is consistent.
  - Investigate how technology affects working practices.
  - Improve the usability of an existing product.

Explore the Questions

- All evaluations need goals & questions to guide them so time is not wasted on ill-defined studies.
- For example, the goal of finding out why many customers prefer to purchase paper airline tickets rather than e-tickets can be broken down into sub-questions:
  - What are customers' attitudes to these new tickets?
  - Are they concerned about security?
  - Is the interface for obtaining them poor?
- What questions might you ask about the design of a cell phone?

Choose the Paradigm & Techniques

- The evaluation paradigm strongly influences the techniques used, how data is analyzed and presented.
- E.g. field studies do not involve testing or modeling
- (that was lame!)
### Identify Practical Issues

For example, how to:
- select users
- stay on budget
- staying on schedule
- find evaluators
- select equipment

### Decide on Ethical Issues

- Develop an informed consent form
- Participants have a right to:
  - know the goals of the study
  - what will happen to the findings
  - privacy of personal information
  - not to be quoted without their agreement
  - leave when they wish
  - be treated politely

### Evaluate, interpret & present data

- How data is analyzed & presented depends on the paradigm and techniques used.
- The following also need to be considered:
  - Reliability: can the study be replicated?
  - Validity: is it measuring what you thought?
  - Biases: is the process creating biases?
  - Scope: can the findings be generalized?
  - Ecological validity: is the environment of the study influencing it - e.g. Hawthorn effect
Activity

• DECIDE how you would evaluate
  - Your addition to the ICS web site
  - Your project