Evaluation Techniques

• What is evaluation?
• Goals of evaluation
• Styles of evaluation
• Evaluating the design
• Evaluating the implementation

Evaluation
– tests usability and functionality of system
– occurs in laboratory, field and/or in collaboration with users
– evaluates both design and implementation
– should be considered at all stages in the design life cycle
Goals of Evaluation

- assess the extent of system functionality
- assess the effect of interface on user
- identify specific problems

Styles of evaluation

- Laboratory studies
  - Performed under laboratory conditions
- Field studies
  - Conducted in the work environment or in "the field"

Laboratory Studies

- Advantages:
  - specialist equipment available
  - uninterrupted environment
- Disadvantages:
  - lack of context
  - difficult to observe several users cooperating
- Appropriate
  - if system location is dangerous or impractical for constrained single user systems to allow controlled manipulation of use
Field Studies

• Advantages:
  – natural environment
  – context retained (though observation may alter it)
  – longitudinal studies possible

• Disadvantages:
  – Distractions, interruptions
  – noise

• Appropriate
  – where context is crucial for longitudinal studies

Evaluating Designs

• Cognitive Walkthrough
  – evaluates design on how well it supports user in learning task
  – usually performed by expert in cognitive psychology
  – expert ‘walks through’ design to identify potential problems using psychological principles
  – forms used to guide analysis

• Heuristic Evaluation

• Review-based evaluation

• Use of models
Cognitive Walkthrough (ctd)

• For each task, walkthrough considers
  – what impact will interaction have on user?
  – what cognitive processes are required?
  – what learning problems may occur?

• Analysis focuses on goals and knowledge: does the design lead the user to generate the correct goals?

We need

– Description of the prototype of the system
– Description of the task the user is to perform on the system
– A complete written list of the actions needed to complete the task with the given prototype
– Who users are and what kind of experience and knowledge the evaluators can assume about them

• Will the users be trying to produce whatever effect the action has?
• Will users be able to notice that the correct action is available?
• Once users find the correct action at the interface, will they know that it is the right one for the effect they are trying to produce?
• After the action is taken, will users understand the feedback they get?
Heuristic Evaluation

- Heuristic evaluation is the most popular of the usability inspection methods

- Usability criteria (heuristics) are identified (Jakob Nielsen – Usability Engineering http://www.useit.com/jakob/useengbook.html)

- Heuristic evaluation is done as a systematic inspection of a user interface design for usability

Heuristic Evaluation

- The goal of heuristic evaluation is to find the usability problems in the design so that they can be attended to as part of an iterative design process.

- Heuristic evaluation involves having a small set of evaluators examine the interface and judge its compliance with recognized usability principles (the "heuristics").

Heuristics

- Visibility of system status
  - The system should always keep users informed about what is going on, through appropriate feedback within reasonable time

- Match between system and real world
  - The system should speak the users' language, with words, phrases and concepts familiar to the user, rather than system-oriented terms. Follow real-world conventions, making information appear in a natural and logical order
Heuristics

• User control and freedom
  – Users often choose system functions by mistake and will need a clearly marked "emergency exit" to leave the unwanted state without having to go through an extended dialogue. Support undo and redo

• Consistency and standards
  – Users should not have to wonder whether different words, situations, or actions mean the same thing. Follow platform conventions

Heuristics

• Error prevention
  – Even better than good error messages is a careful design which prevents a problem from occurring in the first place. Either eliminate error-prone conditions or check for them and present users with a confirmation option before they commit to the action

• Recognition rather than recall
  – Minimize the user's memory load by making objects, actions, and options visible. The user should not have to remember information from one part of the dialogue to another. Instructions for use of the system should be visible or easily retrievable whenever appropriate

Heuristics

• Flexibility and efficiency of use
  – Accelerators -- unseen by the novice user -- may often speed up the interaction for the expert user such that the system can cater to both inexperienced and experienced users. Allow users to tailor frequent actions

• Aesthetic and minimalist design
  – Dialogues should not contain information which is irrelevant or rarely needed. Every extra unit of information in a dialogue competes with the relevant units of information and diminishes their relative visibility
Heuristics

• Help users recognize, diagnose and recover from errors
  – Error messages should be expressed in plain language (no codes), precisely indicate the problem, and constructively suggest a solution

• Help and documentation
  – Even though it is better if the system can be used without documentation, it may be necessary to provide help and documentation. Any such information should be easy to search, focused on the user's task, list concrete steps to be carried out, and not be too large

Number of evaluators

![Graph showing the number of evaluators and the percentage of evaluators interested in the study](image)

Review-based Evaluation

• Results from the literature used to support or refute parts of design

• Care needed to ensure results are transferable to new design
  – population of users (novice, experts)
  – Assumptions made
  – Analyses performed

• Similarities and differences between experimental context and design considered
Evaluating Implementations

Requires an artefact:

- simulation
- prototype
- full implementation

Types of evaluation

- Empirical or experimental methods
- Observational methods
- Query techniques

Experimental Evaluation

- controlled evaluation of specific aspects of interactive behaviour
- evaluator chooses hypothesis to be tested
- a number of experimental conditions are considered which differ only in the value of some controlled variable
- changes in behavioural measure are attributed to different conditions
### Experimental Factors

- **Subjects**
  - who – representative, sufficient sample
- **Variables**
  - things to modify and measure
- **Hypothesis**
  - what you’d like to show
- **Experimental design**
  - how you are going to do it

### Variables

- **Independent variable (IV)**
  - characteristic changed to produce different conditions
  - e.g. interface style, number of menu items, level of help, number of menu items, icon design
- **Dependent variable (DV)**
  - characteristics measured in the experiment
  - e.g. time taken, number of errors, user preference, user performance

### Hypothesis

- **Prediction of outcome of an experiment**
  - framed in terms of IV and DV
  - e.g. “error rate will increase as font size decreases”
- **null hypothesis**:  
  - states no difference between conditions  
  - aim is to disprove this  
  - e.g. null hyp. = “no change with font size”
Experimental Design

- within groups design
  - each subject performs experiment under each condition
  - transfer of learning possible
  - less costly and less likely to suffer from user variation
- between groups design
  - each subject performs under only one condition
  - no transfer of learning
  - more users required
  - variation can bias results

Analysis of Data

- Before you start to do any statistics:
  - look at data
  - save original data
- Choice of statistical technique depends on
  - type of data
  - information required
- Type of data
  - discrete - finite number of values
  - continuous - any value

Analysis - Types of Test

- parametric
  - assume normal distribution
  - robust
  - powerful
- non-parametric
  - do not assume normal distribution
  - less powerful
  - more reliable
- contingency table
  - classify data by discrete attributes
  - count number of data items in each group
Analysis of Data (Cont.)

• What information is required?
  – is there a difference?
  – how big is the difference?
  – how accurate is the estimate?

• Parametric and non-parametric tests
  mainly address first of these

Statistical techniques

<table>
<thead>
<tr>
<th>Independent variable</th>
<th>Dependent variable</th>
<th>Parametric test</th>
<th>Non-parametric test</th>
<th>Contingency test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nominal and Ordinal</td>
<td>Nominal and Ordinal</td>
<td>Student's t-test</td>
<td>Wilcoxon or Mann-Whitney</td>
<td>Chi-square test</td>
</tr>
<tr>
<td>Continuous</td>
<td>Continuous</td>
<td>Analysis of variance</td>
<td>Sign or Rank-Sum test</td>
<td>&quot;H&quot; test</td>
</tr>
<tr>
<td>Continuous</td>
<td>Continuous</td>
<td>Linear regression analysis</td>
<td>&quot;H&quot; test</td>
<td>&quot;H&quot; test</td>
</tr>
</tbody>
</table>

Experimental Studies on Groups

More difficult than single-user experiments

Problems with:
– subject groups
– choice of task
– data gathering
– analysis
Subject Groups

larger number of subjects ⇒ more expensive

longer time to ‘settle down’ ... even more variation!
difficult to timetable
so ... often only three or four groups

The Task

must encourage cooperation

perhaps involve multiple channels

options:
- creative task e.g. ‘write a short report on …’
- decision games e.g. desert survival task
- control task e.g. ARKola bottling plant

Data Gathering

several video cameras
  + direct logging of application

problems:
- synchronisation
- sheer volume!

one solution:
- record from each perspective
Analysis

N.B. vast variation between groups
solutions:
- within groups experiments
- micro-analysis (e.g., gaps in speech)
- anecdotal and qualitative analysis
look at interactions between group and media
controlled experiments may ‘waste’ resources!

Field Studies

Experiments dominated by group formation
Field studies more realistic:
distributed cognition ⇒ work studied in context
real action is situated action
physical and social environment both crucial
Contrast:
- psychology – controlled experiment
- sociology and anthropology – open study and rich data

Observational Methods

• Think Aloud
• Cooperative evaluation
• Protocol analysis
• Automated analysis
• Post-task walkthroughs
Think Aloud

- user observed performing task
- user asked to describe what he is doing and why, what he thinks is happening, etc

Advantages
- simplicity - requires little expertise
- can provide useful insight
- can show how system is actually use

Disadvantages
- subjective
- selective
- act of describing may alter task performance

Cooperative Evaluation

- variation on think aloud
- user collaborates in evaluation
- both user and evaluator can ask each other questions throughout

Additional advantages
- less constrained and easier to use
- user is encouraged to criticize system
- clarification possible

Protocol Analysis

- paper and pencil — cheap, limited to writing speed
- audio — good for think aloud, difficult to match with other protocols
- video — accurate and realistic, needs special equipment, obtrusive
- computer logging — automatic and unobtrusive, large amounts of data difficult to analyze
- user notebooks — coarse and subjective, useful insights, good for longitudinal studies

- Mixed use in practice
- audio/video transcription difficult and requires skill
- Some automatic support tools available
Automated Analysis – EVA

- Workplace project
- Post task walkthrough
  - user reacts on action after the event
  - used to fill in intention
- Advantages
  - analyst has time to focus on relevant incidents
  - avoid excessive interruption of task
- Disadvantages
  - lack of freshness
  - may be post-hoc interpretation of events

Post-task Walkthroughs

- transcript played back to participant for comment
  - immediately → fresh in mind
  - delayed → evaluator has time to identify questions
- useful to identify reasons for actions and alternatives considered
- necessary in cases where think aloud is not possible

Query Techniques

- Interviews
- Questionnaires
Interviews

• analyst questions user on one-to-one basis
  usually based on prepared questions
• informal, subjective and relatively cheap

• Advantages
  – can be varied to suit context
  – issues can be explored more fully
  – can elicit user views and identify unanticipated problems

• Disadvantages
  – very subjective
  – time consuming

Questionnaires

• Set of fixed questions given to users

• Advantages
  – quick and reaches large user group
  – can be analyzed more rigorously

• Disadvantages
  – less flexible
  – less probing

Questionnaires (ctd)

• Need careful design
  – what information is required?
  – how are answers to be analyzed?

• Styles of question
  – general
  – open-ended
  – scalar
  – multi-choice
  – ranked
Physiological Methods

• Eye tracking
• Physiological measurement

Eye Tracking

• head or desk mounted equipment tracks the position of the eye
• eye movement reflects the amount of cognitive processing a display requires
• measurements include
  • fixations: eye maintains stable position. Number and duration indicate level of difficulty with display
  • saccades: rapid eye movement from one point of interest to another
  • scan paths: moving straight to a target with a short fixation at the target is optimal

Physiological Measurements

• emotional response linked to physical changes
• these may help determine a user’s reaction to an interface
• measurements include:
  • heart activity, including blood pressure, volume and pulse
  • activity of sweat glands: Galvanic Skin Response (GSR)
  • electrical activity in muscle: electromyogram (EMG)
  • electrical activity in brain: electroencephalogram (EEG)
• some difficulty in interpreting these physiological responses - more research needed
Choosing an Evaluation Method

- when in process: design vs. implementation
- style of evaluation: laboratory vs. field
- how objective: subjective vs. objective
- type of measures: qualitative vs. quantitative
- level of information: high level vs. low level
- level of interference: obtrusive vs. unobtrusive
- resources available: time, subjects, equipment, expertise

Classification of techniques

<table>
<thead>
<tr>
<th>Cognitive walkthrough</th>
<th>Heuristic evaluation</th>
<th>Review based</th>
<th>Model based</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stage</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Style</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Objective?</td>
<td>No</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Measure</td>
<td>Qualitative</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Information</td>
<td>Low level</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Immediacy</td>
<td>N/A</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intrusive?</td>
<td>No</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Time</td>
<td>Medium</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Equipment</td>
<td>Low</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Expertise</td>
<td>High</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Classification of techniques

<table>
<thead>
<tr>
<th>Experiment</th>
<th>Interviews</th>
<th>Questionnaire</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stage</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Style</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Objective?</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>Measure</td>
<td>Quantitative</td>
<td></td>
</tr>
<tr>
<td>Information</td>
<td>Low/High level</td>
<td>Qualitative/quantitative</td>
</tr>
<tr>
<td>Immediacy</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>Intrusive?</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>Time</td>
<td>High</td>
<td></td>
</tr>
<tr>
<td>Equipment</td>
<td>Medium</td>
<td></td>
</tr>
<tr>
<td>Expertise</td>
<td>Low</td>
<td></td>
</tr>
</tbody>
</table>
## Classification of techniques

<table>
<thead>
<tr>
<th>Think aloud†</th>
<th>Protocol analysis‡</th>
<th>Post-task MT†</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stage</td>
<td>Implementation</td>
<td>Implementation</td>
</tr>
<tr>
<td>Sub-task</td>
<td>LabField</td>
<td>LabField</td>
</tr>
<tr>
<td>Objective</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Instrument</td>
<td>HighFlow level</td>
<td>HighFlow level</td>
</tr>
<tr>
<td>Necessity</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Time</td>
<td>High</td>
<td>High</td>
</tr>
<tr>
<td>Equipment</td>
<td>High</td>
<td>Low</td>
</tr>
<tr>
<td>Expertise</td>
<td>Medium</td>
<td>Low</td>
</tr>
</tbody>
</table>

*Assuming a simple paper and pencil record.
† Including video, audio and system logging.
‡ "MT" = walk-through.