Chapter 9

Evaluation Techniques

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

- ✓Evaluates level of system functionality
- ✓Evaluates effect of interface on user
- ✓Identifies specific problems

Evaluating Designs

Cognitive Walkthrough
Heuristic Evaluation
Review-based evaluation

Cognitive Walkthrough

Proposed by ***Polson et al.***

- evaluates design on *how well it supports user in learning task*
- usually *performed by expert* in cognitive psychology
- expert 'walks though' design to *identify possible problems* using psychological principles
- forms used to *guide analysis*

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 make the correct goals?

Heuristic Evaluation

- Proposed by **Nielsen and Molich.**
  - *usability criteria* (heuristics) are identified
  - *design examined by experts* to see if these are violated
- Examples of heuristics
  - system behaviour is predictable
  - system behaviour is consistent
  - feedback is provided
- Heuristic evaluation `debugs' design. *Debugging* means identifying errors and fix them

Review-based evaluation

- Results from the written review used to support or refute (disapprove) parts of design.
- Care needed to ensure results are transferable to new design.
- ***Model-based evaluation***

Evaluating through user Participation

Laboratory studies

- Advantages:
  - specialist equipment available
  - uninterrupted environment
- Disadvantages:
  - lack of context(environment)
  - difficult to observe several users cooperating

Field Studies

- Advantages:
  - natural environment
  - context retained (though observation may alter it)
- Disadvantages:
  - distractions
  - noise

Evaluating Implementations

Requires an artefact:

simulation, prototype,

full implementation

Experimental evaluation

- controlled evaluation of interactive behaviour
- evaluator chooses hypothesis(theory) 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

1. independent variable (IV)
   - characteristic changed to produce different conditions
   - e.g. interface style, number of menu items

2. dependent variable (DV)

   characteristics measured in the experiment

   e.g. time taken, number of errors.

Hypothesis (thesis or theory)

- prediction of outcome
  - framed in terms of IV and DV

    e.g. “error rate will increase as font size decreases”

- null hypothesis:
  - states no difference between conditions

    e.g. null hyp. = “no change with font size”

Experimental design

1. Within(inside) groups design
   - each subject performs experiment under each condition.
   - less costly and less likely to suffer from user variation.
2. Between groups design
   - each subject performs under only one condition
   - 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 (fixed) number of values
  - continuous - any value

Analysis - types of test

- Parametric
  - assume normal distribution
  - powerful
- Non-parametric
  - do not assume normal distribution
  - less powerful
  - more reliable
- Likelihood 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?

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

$\Rightarrow$ more expensive

longer time to `settle down'

… even more variation!

difficult to timetable

so … often only three or four groups

The task

1. Difficult task
2. Medium task
3. Easy task

Data gathering

several video cameras
+ direct logging of application

Experimental Laboratory

Field studies

Experiments ruled by group formation

Field studies more realistic:

- work studied in context
- real action is *situated action*
- physical and social environment both crucial

Observational Methods

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

Think Aloud

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

- Advantages
  - simplicity - requires little expertise
  - can provide useful insight
  - can show how system is actually used
- Disadvantages
  - Selective (careful)
  - act of describing may alter task performance

Think Aloud

- I predict that ...
- I can picture ...
- A question I have is ...
- This reminds me of ...
- This is like ...
- I am confused about ...
- The big idea here is ...
- I believe ...

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
  - explanation 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 computer logging – automatic, large amounts of data difficult to analyze
- ✓ user notebooks – coarse and subjective, useful insights, good for longitudinal studies

- o Mixed use in practice.
- o audio/video transcription difficult and requires skill.

automated analysis – EVA

- Workplace project
- Post task walkthrough
  - user reacts on action after the event
- Advantages
  - analyst has time to focus on relevant incidents
  - avoid unnecessary interruption of task
- Disadvantages
  - lack of newness
  - 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 searching

Questionnaires (ctd)

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

Questionnaire

- Styles of question
  1. General - establish background of user
  2. Open-ended
     - 'Can you suggest improvements to interface?'
  3. Scalar
     - It is easy to recover from mistakes.
       Disagree 1 2 3 4 5 Agree
  4. Multi-choice
     - How do you most often get help with the system? Choose one.
       - ❑ online manual
       - ❑ contextual help
       - ❑ command prompt
       - ❑ ask a colleague
  5. Ranked – place a list of items in order

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
  1. fixations: eye maintains stable position. Number and duration indicate level of difficulty with display
  2. saccades: rapid eye movement from one point of interest to another
  3. 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
  - electrical activity in muscle
  - electrical activity in brain
- 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