HUMAN-COMPUTER INTERACTION THIRD EDITION

DIX FINLAY ABOWD BEALE

chapter 9

evaluation techniques

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

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

Evaluating Designs (expert based)

Cognitive Walkthrough
Heuristic Evaluation
Review-based evaluation

Cognitive Walkthrough

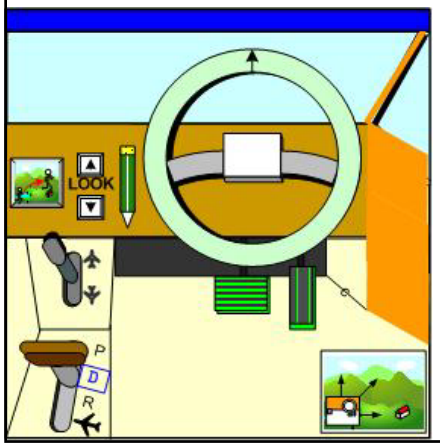
Proposed by Polson *et al. 1992*

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

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?

Pen-based interface for LIDS



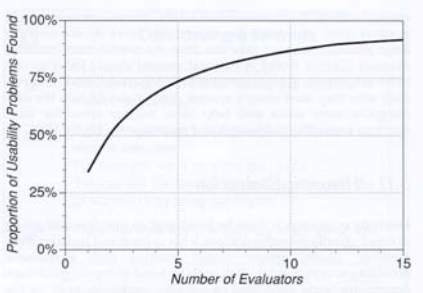
- UA: Press look up button
- SD: Scroll viewpoint up
- UA: Press steering wheel to drive forwards
- SD: Move viewpoint forwards
- UA: Press look down button
- SD: Scroll viewpoint down
-
-
-

Pen interface walkthrough

- UA 1: Press look up button
  1. Is the effect of the action the same as the user's goal at this point?
     Up button scrolls viewpoint upwards.
  2. Will users see that the action is available?
     The up button is visible in the UI panel.
  3. Once users have found the correct action, will they know it is the one they need?
     There is a lever with up/down looking symbols as well as the shape above and below the word look. The user will probably select the right action.
  4. After the action is taken, will users understand the feedback they get?
     The scrolled viewpoint mimics the effect of looking up inside the game environment.

Heuristic Evaluation

- Proposed by Nielsen and Molich.
- usability criteria (heuristics) are identified
- design examined by experts to see if these are violated



Heuristic Evaluation

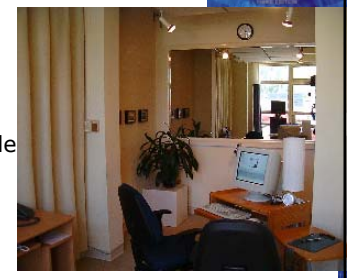
- Rank by severity
  - 0=no usability problem
  - 4=usability catastrophe
- Heuristics such as 10 from Nielsen
  - Visibility of system status
  - Match between system and real world
  - User control and freedom, etc.
- Heuristic evaluation `debugs' design.

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.
- Model-based evaluation (e.g., GOMS, keystroke)
- Cognitive models used to filter design options
  e.g. GOMS prediction of user performance.
- Design rationale can also provide useful evaluation information

Evaluating through user Participation

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
  - noise
- Appropriate
  - where context is crucial for longitudinal studies

Evaluating Implementations

Requires an artefact:
simulation, prototype,
full implementation

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
    - not the programmer friend, boss, etc.
    - huge variability in effectiveness (e.g., programmers)
- 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
- dependent variable (DV)
  characteristics measured in the experiment
  e.g. time taken, number of errors.

Hypothesis

- 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
  - aim is to disprove this

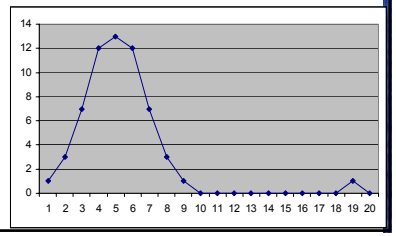
    e.g. null hyp. = "no change with font size"

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 (e.g. average=5.25 – but 4.9 without outlier)
  - 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

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