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

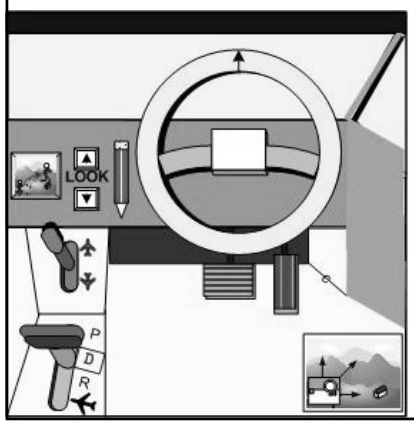
Proposed by Polson *et al. 1992*

- evaluates design on how well it supports user in learning task
- usually performed by expert in cognitive psychology
- expert 'walks through' design to identify potential problems using psychological principles
  - Based on the idea of a code walkthrough in conventional code testing
- 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



- UA1: Press look up button
- SD1: Scroll viewpoint up
- UA2: Press steering wheel to drive forwards
- SD2: Move viewpoint forwards
- UA3: Press look down button
- SD3: Scroll viewpoint down
- .
- .
- .

Pen interface walkthrough

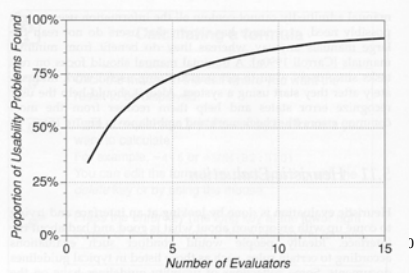
- UA 1: Press look up button
  1. Is the effect of the action the same as the user's goal at this point?
     Up button scrolls viewpoint upwards.
  2. Will users see that the action is available?
     The up button is visible in the UI panel.
  3. Once users have found the correct action, will they 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.

Cognitive walkthrough results

- Fill out a form
  - Track time/date of walkthrough, who the evaluators were
  - For each Action, answer the four proforma questions (as per prev slide, text pp. 321-322)
  - Any negative answer to any question should be documented on a separate Problem Sheet, indicating how severe the evaluators think the problem is, and whether they think it'll occur often

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
  - 1=cosmetic – fix if have extra time
  - 2=minor – fixing is low priority
  - 3=major – important to fix
  - 4=usability catastrophe – imperative to fix
- Heuristics such as 10 from Nielsen
  - Visibility of system status
  - Match between system and real world
  - User control and freedom, etc.

  (p. 325-326) [remember – these will be used to assess your assignment 1 prototype!]
- Heuristic evaluation `debugs' design

Review-based evaluation

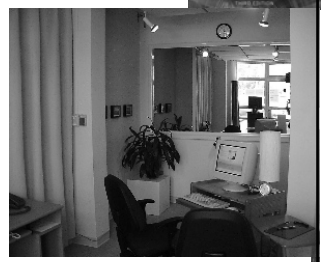
- Results from the literature used to support or refute parts of design
  - Care needed to ensure results are transferable to new design
- Model-based evaluation (e.g., GOMS, keystroke)
  - Cognitive models used to filter design options

    e.g. GOMS prediction of user performance

  (we look at these later in the semester)
- Design rationale can also provide useful evaluation information

Evaluating through user Participation

Laboratory studies



- User taken out of their normal work environment for a controlled test
- 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 – *see next slide*)
  - longitudinal studies possible
- Disadvantages:
  - distractions
  - noise
- Appropriate
  - where context is crucial for longitudinal studies

Unwanted biases in studies

- You can't always take a study result at face value… must be attentive to what subjects are feeling
- Hawthorne effect
  - Worker is more productive when observed
- John Henry effect
  - Worker is [stubbornly] more productive when using his old tools (see http://www.ibiblio.org/john_henry/)
- Placebo effect
  - [Patient usually] gets some benefit just because they expect a benefit
- Pygmalion effect
  - Student performs better simply because they are expected to do so

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 (i.e., the users)
  - who – representative, sufficient sample
    - not the programmer's friend, boss, etc.
    - huge variability in performance of individuals
- Variables
  - things to modify and measure
- Hypothesis
  - what you'd like to show
- Experimental design
  - how you are going to show it
  - Includes 'Protocol' – what the subjects do

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

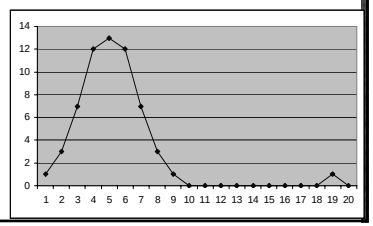
    e.g. null hyp. = “no change with font size”

Experimental design

- “within groups” design (also called “repeated measures”)
  - each subject performs experiment under each condition
  - transfer of learning possible (practice makes performance better; or alternatively fatigue or boredom makes it worse)
  - less costly and less likely to suffer from user variation (each user is compared to themselves)
- between groups design
  - each subject performs under only one condition
  - no transfer of learning
  - more users required

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

ANOVA – analysis of variance

- Quite easy to test whether there's a significant difference between groups in Excel
  - Need to invoke Tools/Add-ins/Analysis Tookpack to enable
  - Then just apply Tools/Data Analysis/ANOVA: Single Factor to the data

ANOVA from Excel

Say we have three columns of numbers representing the time to complete a task for 5, 5 and 7 users using three variations of an interface

If P-value < 0.05 then we usually say the result is 'significant' (result is more than expected chance variation)

Anova: Single Factor

SUMMARY

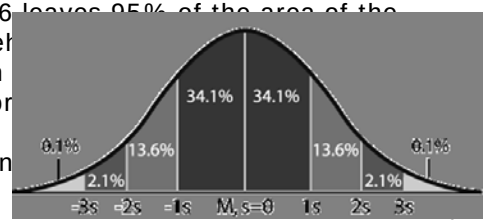
Groups	*Count*	*Sum*	*Average*	*Variance*
Group 1	5	82	16.4	9.3
Group 2	5	67	13.4	10.8
Group 3	7	79	11.28571	3.904762

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	76.28908	2	38.14454	5.244339	0.019959	3.738892
Within Groups	101.8286	14	7.273469			
Total	178.1176	16				

When is a difference a difference?

- In the world of parametric stats, we look for a statistic to be large enough to be 'significant'
  - On the Gaussian ('normal') curve a $|Z|=1.96$ leaves 95% of the area of the curve beyond it. This is a common 'critical value' for claiming significance



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

About statistics

- It's an amazingly complex field
  - A lot of hidden complexities in running experiments and saying that the observed differences really make a difference ('threats to validity')
  - When the outcomes are serious (e.g., medical trials) professional statisticians are always used in design of the experiment as well as analysis and reporting of the findings
  - Plenty of texts and courses on stats available (the Wikipedia is pretty good on this topics, too – e.g., for ANOVA)