HUMAN-COMPUTER INTERACTION THIRD EDITION DIX FINLAY ABOWD BEALE

Chapter 9

Evaluation techniques

1

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
  - can be performed either by experts who analyze the system or through user participation

2

Goals of evaluation

- Assess extent of system functionality (i.e. the current design allows users to perform their tasks easily)
- Assess effect of interface on user (e.g. how easy is it to use, whether it is pleasant to use, etc.)
- Identify specific problems with the design (e.g. the current design causes unexpected problems or confusion to users)

3

Evaluating designs through expert analysis

Cognitive walkthrough
Heuristic evaluation
Review-based evaluation

4

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 potential problems using psychological principles
- forms used to guide analysis

5

Cognitive walkthrough (ctd)

- To do a walkthrough, the following are needed:
  - a specification or prototype of the system, not complete but fairly detailed
  - a description of the task the user wants to perform on the system; it should be a typical task that most users will want to do
  - a complete list of the actions needed to complete the task
  - an indication of the profile of a typical user (experience, knowledge, etc.)

6

Cognitive walkthrough (ctd)

- For each action in the list of item 3, answer the following questions:
  - is the effect of the action the same as the user's goal at that point (e.g. if the effect is to save a document, is this the intention of the user?)?
  - will users see that the action is available?
  - once users have found the correct action, will they know it is the one they need?
  - after the action is taken, will users understand the feedback they get?

7

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?

8

Heuristic evaluation

- Proposed by Nielsen and Molich
- Usability criteria (heuristics) are identified
- Design examined by experts to see if these heuristics are violated
- A group of experts is needed, about three to five, experience showing that five will uncover about 75% of usability problems
- Heuristic evaluation 'debugs' design

9

Heuristic evaluation (ctd)

Nielsen's ten heuristics are:

1. *Visibility of system status* (always keep user informed about what is going on)
2. *Match between system and the real world* (system messages should use concepts familiar to the user)
3. *User control and freedom* (e.g. support do/undo button)
4. *Consistency and standards* (a term has the same meaning in all contexts)
5. *Error prevention* (not making errors in the first place is better than helpful error messages)

10

Heuristic evaluation (ctd)

Nielsen's ten heuristics (ctd):

6. *Recognition rather than recall* (relevant information should be visible rather than needed to be remembered)
7. *Flexibility and efficiency of use* (support shortcuts for expert users)
8. *Aesthetic and minimalistic design* (extra not needed information diminishes the relative value of relevant information)
9. *Help users recognize, diagnose and recover from errors* (error messages should be clear and helpful)
10. *Help and documentation* (good on-line help facilities)

11

Heuristic evaluation (ctd)

- Each expert assesses the system and notes violations of these heuristics, that could indicate a potential usability problem
- This assessment is based on four factors:
  - how common is the problem
  - how easy is it for the user to overcome it
  - will it be a one-off problem or a persistent one
  - how seriously will the problem be perceived
- These can be combined into an overall rating:
  0. I don't agree this is a problem at all
  1. Cosmetic problem to be fixed only if time permits
  2. Minor problem, fixing it should be of low priority
  3. Major problem, important to fix, high priority
  4. Usability catastrophe, imperative to fix

12

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 (e.g. interfaces for novice vs expert users)
- Model-based evaluation (e.g. dialog models)
- Cognitive models used to filter design options
  e.g. GOMS prediction of user performance
- Design rationale can also provide useful evaluation information

13

Evaluating through user participation

14

Laboratory studies

- Advantages:
  - specialist equipment available (e.g. two way mirrors, audio/visual recording, etc.)
  - uninterrupted environment
- Disadvantages:
  - lack of context (e.g. filing cabinets, wall calendars, interruptions, noise, create a "real" environment)
  - 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 (e.g. space stations)

15

Field studies

- Advantages:
  - natural environment
  - context retained (though observation may alter it)
  - longitudinal studies (taking days or weeks) possible
- Disadvantages:
  - distractions
  - noise
- Appropriate
  - where context is crucial for longitudinal studies

16

Evaluating implementations

Requires an artefact:
simulation, prototype,
full implementation

17

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

18

Experimental factors

- Subjects
  - who – representative, sufficient sample (minimum 5-10)
- Variables
  - things to modify and measure
- Hypothesis
  - what you'd like to show
- Experimental design
  - how you are going to do it

19

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 to complete a task, number of errors made

20

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"

21

Experimental design

- *Between groups (or randomized)* design
  - each subject performs under only one condition
  - there are at least two conditions:
    - the experimental (in which the variable has been manipulated)
    - the control (that ensures it is the manipulation that is responsible for any differences that are measured)
  - the advantage is that any learning effect resulting from the user performing in one condition and then the other is controlled: users perform under only one condition, so transfer of learning between one condition to the next, which could affect the result, is not happening
  - but more users required, and
  - significant variation between groups can bias results

22

Experimental design (ctd)

- *Within groups (or repeated measures)* design
  - each subject performs experiment under each condition
  - transfer of learning possible
    - negative effect can be reduced if order of tackling conditions varies between groups (group A does first condition followed by the second and group B does them in reverse order)
  - less costly (fewer users required)
  - less likely to suffer from user variation
- Choice of method depends on resources available, to what extent learning transfer can be controlled, and how representative is the group
- If more than one independent variable is involved, a mixed approach can be used

23

Analysis of data

- Before you start to do any statistics:
  - look at data; try to spot outliers (single data items that are very different from the rest)
  - save original data (to be available for different analysis methods, if such a need arises)
- Choice of statistical technique depends on
  - type of data
  - information required / questions need to be answered
- Type of data
  - discrete - finite number of values
  - continuous - any value

24

Analysis - types of test

- Parametric
  - assume known distribution of data, such as normal distribution
  - robust (give reasonable results even if data are not precisely normal) and powerful
- Non-parametric
  - do not assume normal distribution
  - usually based on the ranking of data (e.g. a set of values: 57,32,61,49 reduced to ranking: 3,1,4,2)
  - less powerful (may not detect a difference that a parametric test will detect)
  - more reliable (more resistant to outliers)
- Contingency table
  - classify data by discrete attributes
  - count number of data items in each group

25

Choosing a statistical technique

Independent variable	Dependent variable	
Parametric		
Two valued	Normal	Student's *t* test on difference of means
Discrete	Normal	ANOVA (ANalysis Of VAriance)
Continuous	Normal	Linear (or non-linear) regression factor analysis
Non-parametric		
Two valued	Continuous	Wilcoxon (or Mann–Whitney) rank-sum test
Discrete	Continuous	Rank-sum versions of ANOVA
Continuous	Continuous	Spearman's rank correlation
Contingency tests		
Two valued	Discrete	No special test, see next entry
Discrete	Discrete	Contingency table and chi-squared test
Continuous	Discrete	(Rare) Group independent variable and then as above

26

Analysis of data (ctd)

- What information is required?
  - is there a difference (is one system better than another)?
  - how big is the difference ("selection from five items is 260ms faster than from seven items")?
  - how accurate is the estimate ("selection is faster by 260ms plus or minus 30ms")?
- Parametric and non-parametric tests mainly address first of these questions

27

Experimental studies on groups

More difficult to evaluate groupware environments than single-user ones

Problems with:
- subject groups
- choice of task
- data gathering
- analysis

28

Subject groups

Larger number of subjects
⇒ more expensive

Longer time for a group to `settle down'
… even more variation!

Difficult to timetable the use of shared resources (possibly also used by other people)

So … often only three or four groups

29

The task

- Choosing a suitable task is difficult, as we may want to test a variety of different task types
  - creative, structured, information passing, …
- Must encourage cooperation, to reach consensus or because of distributed control
- Perhaps involve multiple channels of communication in a groupware application
- Options:
  - creative task — e.g. '*write a short report on …*'
  - decision games — e.g. desert survival task
  - control task — e.g. ARKola bottling plant (noise)

30

Data gathering

Several video cameras
+ direct logging of application

Problems:
- Synchronisation of different sources of data gathering
- sheer volume!

One solution:
- record from each participant individually; recreate the situation as it appears to the participant; repeat for all participants

31

Analysis

N.B. vast variation between groups (worse than differences between individuals in single-user experiments): democratic vs autocratic, etc.

Solutions:
- within groups experiments where each group works under several conditions
- micro-analysis (e.g., gaps in speech)
- anecdotal and qualitative analysis looking for critical incidents (interesting events or breakdowns) in the data

Look at interactions between group and communication media as well as applications used

Controlled experiments with a limited number of groups may not be productive ands 'waste' resources!

32

Field studies

Experiments dominated by group formation (often "artificial" combination of people that doesn't necessarily reflect the real working environment)

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
ethnography – very detailed recording of interactions between people and their interactions with the environment and each other

33

Observational methods

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

34

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 used
- Disadvantages
  - subjective and/or selective, depending on the tasks provided
  - act of describing how a task is done may alter task performance

35

Cooperative evaluation

- Variation on think aloud
- User collaborates in evaluation (not simply an experimental participant)
- Both user and evaluator can ask each other questions throughout

- Additional advantages
  - less constrained and easier to use
  - user is encouraged to criticize system
  - the evaluator can clarify points of confusion at the time they occur and thus maximize the potential to identify problems

36

Protocol analysis

- Paper and pencil – cheap, limited to writing speed
- Audio – good for think aloud, difficult to match with other protocols (e.g. handwritten script)
- Video – accurate and realistic, needs special equipment, obtrusive (e.g. ask user not to move!)
- Computer logging – automatic and unobtrusive, large amounts of data difficult to analyze
- User notebooks – coarse and subjective, useful insights, good for longitudinal studies and logging unusual situations
- Mixed use of above techniques in practice
- Audio/video transcription difficult and requires skill
- Some automatic support tools available

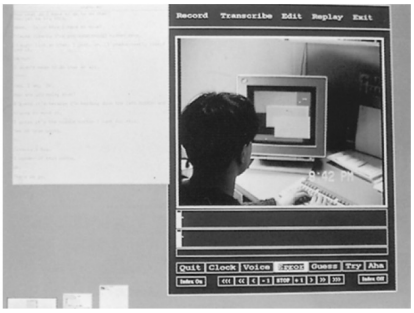
37

Automated analysis – EVA

- Workplace project at Xerox PARC
- 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 (possibly critical) task
- Disadvantages
  - lack of freshness
  - may be post-hoc interpretation of events
  - tagging and annotating events can prevent the evaluator from concentrating on the events themselves

38

Experimental Video Annotator



EVA: an automatic protocol analysis tool. Source: Wendy Mackay

39

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

40

Query techniques

Interviews
Questionnaires

41

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

42

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

43

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

44

Physiological methods

Eye tracking
Physiological measurement

45

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

46

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

47

Choosing an evaluation method

Factors distinguishing evaluation techniques
A classification of evaluation techniques

48

Factors distinguishing evaluation techniques

- Design vs implementation
  - the stage in the design process at which evaluation is required
  - at design level, information is collected to feed the implementation
  - at implementation level, there is a physical artifact to use
  - but early evaluation brings the greatest pay-off since any problems can be easily resolved at that stage
- Laboratory vs field studies
  - ideally include both types, lab based (that allows control experimentation) and field based (that offers a natural working environment
- Subjective vs objective
  - ideally include both types, the former can detect problems the latter cannot, the latter avoid bias

49

Factors distinguishing evaluation techniques (ctd)

- Qualitative vs quantitative measures
  - the former is numeric and can be easily analysed using statistical techniques
  - the latter is non-numeric and can provide important detail not determined by numbers
  - often the former is used with objective techniques and the latter with subjective ones
- Information provided
  - depending on the kind of evaluation, low-level information may be required (e.g. which font is more readable) or high-level (e.g. "is the system usable?")
- Immediacy of response
  - some methods record information at the time of the evaluation (e.g. think loud) and others rely on the user's recollection of events (e.g. post-talk walkthrough)

50

Factors distinguishing evaluation techniques (ctd)

- Intrusiveness
  - related to the previous factor of immediacy
  - usually techniques that produce immediate measurements are intrusive, i.e. obvious to the user during the interaction, and thus susceptible of influencing his behaviour
- Resources
  - equipment, time, money, participants, expertise of evaluator, context
  - when resources are limited, a choice must be made in a way that the most effective and useful information can be generated, under the circumstances

51

Classification of analytic evaluation techniques

	Cognitive walkthrough	Heuristic evaluation	Review based	Model based
Stage	Throughout	Throughout	Design	Design
Style	Laboratory	Laboratory	Laboratory	Laboratory
Objective?	No	No	As source	No
Measure	Qualitative	Qualitative	As source	Qualitative
Information	Low level	High level	As source	Low level
Immediacy	N/A	N/A	As source	N/A
Intrusive?	No	No	No	No
Time	Medium	Low	Low–medium	Medium
Equipment	Low	Low	Low	Low
Expertise	High	Medium	Low	High

52

Classification of experimental and query evaluation techniques

	Experiment	Interviews	Questionnaire
Stage	Throughout	Throughout	Throughout
Style	Laboratory	Lab/field	Lab/field
Objective?	Yes	No	No
Measure	Quantitative	Qualitative/ quantitative	Qualitative/ quantitative
Information	Low/high level	High level	High level
Immediacy	Yes	No	No
Intrusive?	Yes	No	No
Time	High	Low	Low
Equipment	Medium	Low	Low
Expertise	Medium	Low	Low

53

Classification of observational evaluation techniques

	Think aloud[1]	Protocol analysis[2]	Post-task walkthrough
Stage	Implementation	Implementation	Implementation
Style	Lab/field	Lab/field	Lab/field
Objective?	No	No	No
Measure	Qualitative	Qualitative	Qualitative
Information	High/low level	High/low level	High/low level
Immediacy	Yes	Yes	No
Intrusive?	Yes	Yes[3]	No
Time	High	High	Medium
Equipment	Low	High	Low
Expertise	Medium	High	Medium

1 Assuming a simple paper and pencil record
2 Including video, audio and system recording
3 Except system logs

54

Classification of monitoring evaluation techniques

	Eye tracking	Physiological measurement
Stage	Implementation	Implementation
Style	Lab	Lab
Objective?	Yes	Yes
Measure	Quantitative	Quantitative
Information	Low level	Low level
Immediacy	Yes	Yes
Intrusive?	No[1]	Yes
Time	Medium/high	Medium/high
Equipment	High	High
Expertise	High	High

1 If the equipment is not head mounted

55

Summary

- Evaluation is an integral part of the design process and should take place throughout the design life cycle
- It can take place in a specialist laboratory or in the user's workplace
- A design can be evaluated by analytic techniques before implementation or by experimental and observational techniques once a prototype is available
- The choice of the evaluation method depends on what exactly is required of the evaluation as well as available resources

56