HCI Lecture 12
Evaluation

Barbara Webb

Key points:

- The problem with anecdotes
- Evaluation by designers or usability experts
- Evaluating during use:
  - Cooperative
  - Ethnographic
  - Automated
- Evaluation after use:
  - Post task walkthroughs
  - Interviews
  - Surveys

An anecdote

- "I downloaded your shareware program for playing music files, but I had to look up the manual to work out that the button marked ATT could be used to attenuate the sound! You should change it!"
- Do you think you need to redesign the system?

Questions

- **Who** is giving this feedback? A design expert? A fellow programmer? A typical user or member of a target user group? Is it just one person or a significant proportion of your users?
- **When** are you getting this feedback? On an early prototype or an established product?
- **How** has this evaluation been arrived at? Is it by comparison to some guidelines, or from a simulated walkthrough? Is it from use in a realistic context ('ecological validity')?
- **What** has been used as a measure? Quantitative (time to complete task, error rate) or qualitative (ease of use ratings)? Compared to recommendations or alternatives? Consistency?
- **Also** what are the costs of making the change?
- How would you go about getting better information about the need for a change?

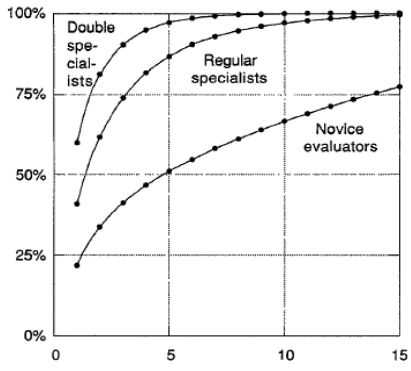
Evaluation by designers: Guideline-based Evaluation

- Interface developers can apply a set of explicit design guidelines (c.f. lecture 10) e.g. Nielsen's 10 heuristics:
  - provide feedback
  - speak the user's language
  - provide clearly marked exits
  - be consistent
  - prevent errors
  - minimize user memory load
  - provide shortcuts
  - use simple and natural dialogue design
  - provide error recovery
  - provide help

'Discount' methods

(Nielsen & Molich, 1990)

- Severity ratings may be given, e.g., on 0-4 scale

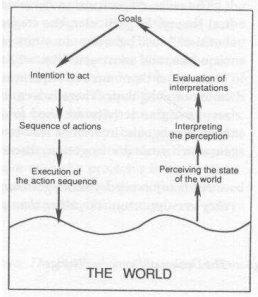


Average proportion of usability problems found as a function of number of evaluators in a group performing the heuristic evaluation.

How Many Heuristic Evaluators and How Skilled?

Evaluation by designers: Cognitive walkthrough

- Cognitive walkthrough (Polson et al 1992) involves stepping through an interaction sequence from the user's perspective
- Organised around Norman's description of task execution:
  1. Goal (question 1, 9a)
  2. Planning (question 2, 3)
  3. Action specification (question 4, 5, 7)
  4. Action execution (question 6)
  5. Perception of outcome (question 8)
  6. Interpretation of the outcome (question 8b)
  7. Evaluation of the outcome (question 8a, 9b)



Cognitive Walkthrough Form

1. Description of user's immediate goal:
2. (First/next) action user should take:
   - Obvious that action is available? Why/Why not?
   - Obvious that action is appropriate to goal? Why/Why not?
3. How will user access description of action?
   - Problem with accessing? Why/Why not?
4. How will user associate description with command?
   - Problem with associating? Why/Why not?
5. All other available commands less appropriate?
   - For each, why/why not?
6. How will user execute the command?
   - Problems? Why/Why not?
7. If time outs, time for user to decide before time out?
   - Why/Why not?
8. Execute the action. Describe system response:
   - Obvious that progress has been made toward goal? Why/Why not?
   - User can access needed info. in system response? Why/Why not?
9. Describe appropriate modified goal, if any:
   - Obvious that goal should change? Why/Why not?
   - If task completed, is it obvious? Why/Why not?

Actions/choices should be ranked according to percentage of users expected to have problems:
0 = none
1 = some
2 = more than half
3 = most

Evaluation by designers

✓ Can occur at any stage, but particularly useful early in design (e.g. can apply to mock-up, prototype or not fully functional systems)
✓ Will identify major problems before users are involved
✓ Will flag potential areas to check in user evaluation

✘ Requires expertise (and experts may be expensive) or significant problems may be missed
✘ Can get large variability between different evaluators
✘ May also generate 'false positives' i.e. problems (according to the guidelines or the cognitive analysis) that would not really be problems in normal use

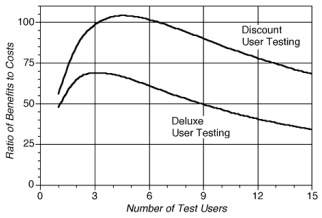
During use: Cooperative Evaluation

- User carries out task with direct feedback to/from evaluator
- E.g.
  - Get users to ask for help when they get stuck
  - Ask users what the commands mean
  - After users read task, ask users how they might solve it
  - As users consider each command, ask what they think it does
  - When users have entered a command, ask what they think it has done and what the response indicates

During use: Cooperative Evaluation

✓ Creates dialogue between evaluators/designers and users
✓ Helps provide intentional context necessary to interpret user behaviour
✓ May learn a lot from just a few users

✘ May affect normal behaviour
✘ Intensive effort for user and evaluator
✘ May be hard to analyse results



During use: Ethnographic Studies

- Discrete observation
  - "hanging around"
  - Can be simple or complex measurement (e.g. time to complete interaction, apparent degree of frustration or enjoyment)
- Ideally should be in normal use situation (e.g. within the company, in the home)

✓ Excellent method for understanding the "mundane features", "common sense logic"of everyday work

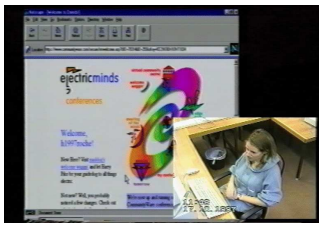
✘ May be hard to keep track of what is going on (record for later analysis?)
✘ Requires significant time commitment
✘ May not know what user was trying to do, what caused errors
✘ May be ethical issues if users are not aware of observation, or interference if they are aware.

During use: Automated

- Activity logs
  - Again could be simple (number of 'undo' commands) or rich (times for every keystroke – when and for how long did they hesitate?)
- Eye tracking (see lecture 1)
- Video and audio recording with automated analysis (e.g. instrumented meeting room, ICCS)

After use: Post-task walkthrough

- Combine pros (and cons) of co-operative and ethnographic by not interfering during use, but instead recording and then play back to user asking about intentions and reasons.
- E.g. re-enactment protocol
  - Subjects were videoed using the www
  - Later, they were shown the video and asked to comment on their actions



Re-enactment protocol

- Subject A downloading Web page from the Times newspaper Web site.
- 25 seconds through the data transfer process subject A moves hand to mouse and starts to move the cursor.
- Stops for 10 seconds then places the cursor over the stop button and waits a further 23 seconds.
- The Web page then starts to render and is completed in 5 seconds.
- Afterwards, subject A is asked about their behaviour:

Re-enactment protocol

A: I had been waiting a while and so I thought something may be wrong and I was going to stop it but then I thought it was about 10 o'clock so lots of people in their offices would be reading the Times.
I: So therefore, it would be slow?
A: Well yes that is what I thought. I was afraid to stop it if it just going to appear in a few seconds. When lots of people use a web page its slow right.
I: And if nobody is using it its fast?
A: Yes that's what I think, when something is obscure I am probably the only person in the world looking at it so it will be fast.
I: OK, so how long do you think you would have waited for before stopping?
A: Well I was very unsure whether to stop the thingy, didn't know whether it was right thing to do. I was deliberating over the fact, but I really was unsure what was happening and why it was slow. Then it just appeared.

After use: Interviews

- Asking user to reflect on their experience
  - Can vary questions to suit the context, or follow up issues in detail
  - Understanding preferences, impressions and attitudes
  - Revealing unanticipated problems
- N.B. Interviews may used to 'administer' questionnaire or may be more or less 'guided' by preset questions
  - Improves consistency but may reduce depth

After use: Surveys and questionnaires

- Questionnaire types
  - Background
  - Attitudes
- Questionnaire design
  - Open-ended, qualitative
    - Don't restrict answers to evaluators' expectations, but difficult to analyse
  - Rating scale
    - Easier to analyse, but needs careful design
    - Typically 5 or 7 point scale measuring agreement/disagreement with statement(s)
  - Multiple choice: select one or as many as apply
- There are several 'standard' usability questionnaires in the public domain
  - These can be adapted for specific needs

After use: Surveys and questionnaires

Example:

Indicate your agreement or disagreement with the following statements by circling the appropriate number.

The system tells me what to do at every point.
Disagree 1 2 3 4 5 Agree

It is easy to recover from mistakes
Disagree 1 2 3 4 5 Agree

I always know what the system is doing
Disagree 1 2 3 4 5 Agree

Etc.

- Could be based on general usability heuristics
- Could be designed to suit specific system characteristics

After use: Surveys and questionairres

✓ Can administer to large numbers
✓ Potential to automate administration and analysis using web forms

- Limited quality of feedback
  - Depends on having right questions
  - May only capture *whether* there were problems, rather than *why*
- Self-selecting bias
  - those who fill in may not be typical, e.g. provide feedback only if experienced a problem
- For this (and all other methods) remains issue of analysing and intepreting the data.

Types of data

- We have discussed a wide range of methods to collect data from users, e.g.:
  - Observe during use: thinking aloud, or in natural situation; data-logging
  - Collect post-task information through playback, interview or survey
- Each method could potentially provide both qualitative and quantitative data:
  - **Quantitative** – has structure of integers or real numbers, e.g. number of errors (discrete), time to complete task (continuous)
    - *Not* just "comprised of numeric values" e.g. numeric labels
  - **Qualitative** – everything else, e.g. categories of action, preferences
    - But note may be able to quantify the number of people falling into a category or expressing the same preference.

? Are rating scales quantitative? E.g.

The system tells me what to do at every point (circle a number)

Disagree 1 2 3 4 5 Agree

Qualitative/interpretive analysis

- Typically used with methods like co-operative evaluation:
  - Raw data is small number of richly described interactions
- May use to identify and list potential problems:
  - C.f. Nielsen's suggestion in 'discount evaluation' that will find most problems with first 3-5 users
  - Look for recurring patterns or critical incidents
  - Examples to illustrate to designer that problems really exist
- May use to gain insight into qualitative aspects of the task, the user or the environment
  - Identify goals, background assumptions, semantic knowledge requirements, working practices
  - Might feed into formal task analysis, user modelling or other theoretical basis
- May try to draw more general conclusions about interaction ⇨

Qualitative/interpretive analysis

- Example: Mack, Lewis and Carol (1983) recorded interactions of experienced typists with early word processing program:

P: (Participant types two lines. At the end of the second line, she types a comma instead of a period and then presses return which positions the cursor at the beginning of a new blank line. Participant notices the typing error.)

P: Oh. I see. So now ....

E: What are you thinking?

P: I made a mistake up here. Now if I want to go back, I guess I would . .. (looks in manual for information).

E: What are you looking at? Page 3-4? (Participant says nothing.) What is that telling you?

P: Well, I'm trying to figure out how to go back to correct that mistake. Am I supposed to correct my mistakes yet? Or am I supposed to just not worry about the mistakes? Or... I'm going to try to go back.

P: (Participant presses backspace and incurs an error which is signaled by a beep. This is because backspace will not move the cursor beyond the left edge of the screen.)

P: Woo! It didn't like that!

Qualitative/interpretive analysis

- On the basis of these transcripts Mack *et al.* drew a number of general conclusions about learning issues in HCI:

1. Learning is difficult – evidence of frustration, difficulty remembering
2. Learners lack basic knowledge – 'who is the printer?' 'return'
3. Learners make ad hoc interpretations - explain away errors
4. Learners generalise from what they know – space bar on typewriter moves the cursor, rather than 'inserting' a space
5. Learners have trouble following directions – will try to anticipate; but if follow directions blindly will not learn
6. Problems interact – can't make decision about what to correct
7. Interface features may not be obvious – significance of beep?
8. Help facilities don't always help – have to know what you are looking for

Qualitative/interpretive analysis

- Issues with qualitative analysis
  - Validity: how much can we generalise from a few examples; are these interactions representative?
  - Reliability: would another evaluator reach the same conclusions?

References

Nielsen, J., & Molich, R. (1990). Heuristic evaluation of user interfaces. Paper presented at the ACM CHI'90 Conference on Human Factors in Computing Systems, Seattle, WA.
(also a handy summary by Nielsen of the main issues here: http://www.useit.com/papers/guerilla_hci.html)

Polson, P. et al (1992) Cognitive walkthroughs: a method for theory based evaluation of user interfaces. *International Journal of Man-Machine Studies*, 36:741-773

Hughes, J. et al (1995) The Role of Ethnography in Interactive Systems Design. *ACM Interactions* April 1995

See also:

- Dix et. al. chapter 9