Evaluation Techniques

OVERVIEW

- Evaluation tests the usability, functionality and acceptability of an interactive system.
- Evaluation may take place:
  - in the laboratory
  - in the field.
- Some approaches are based on expert evaluation:
  - analytic methods
  - review methods
  - model-based methods.
- Some approaches involve users:
  - experimental methods
  - observational methods
  - query methods.
- An evaluation method must be chosen carefully and must be suitable for the job.

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

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 through' design to identify potential problems using psychological principles
- forms used to guide analysis

Cognitive Walkthrough

In the cognitive walkthrough, the sequence of actions refers to the steps that an interface will require a user to perform in order to accomplish some known task.

The evaluators then 'step through' that action sequence to check it for potential usability problems.

Usually, the main focus of the cognitive walkthrough is to establish how easy a system is to learn.

More specifically, the focus is on learning through exploration.

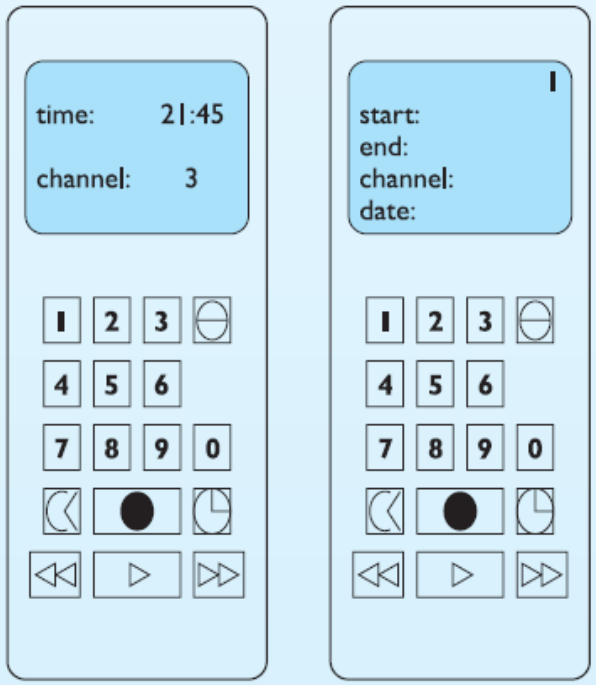
Four things needed in cognitive walkthrough

1. **A specification or prototype of the system**. It doesn't have to be complete, but it should be fairly detailed. Details such as the location and wording for a menu can make a big difference.
2. **A description of the task the user is to perform on the system.** This should be a representative task that most users will want to do.
3. **A complete, written list of the actions needed to complete the task with the proposed system.**
4. **An indication of who the users are and what kind of experience and knowledge the evaluators can assume about them.**

Given the information, the evaluators step through the action sequence (identified in item 3 above) to critique the system. To do this, for each action, the evaluators try to answer the following four questions for each step in the action sequence.

1. **Is the effect of the action the same as the user's goal at that point?**
2. **Will users see that the action is available?**
3. **Once users have found the correct action, will they know it is the one they need?**
4. **After the action is taken, will users understand the feedback they get? If you now assume that the user did manage to achieve the correct action, will he know that he has done so?**

Example



UA 1: Press the 'timed record' button
SD 1: Display moves to timer mode. Flashing cursor appears after 'start:'
UA 2: Press digits 1 8 0 0
SD 2: Each digit is displayed as typed and flashing cursor moves to next position
UA 3: Press the 'timed record' button
SD 3: Flashing cursor moves to 'end:'
UA 4: Press digits 1 9 1 5
SD 4: Each digit is displayed as typed and flashing cursor moves to next position
UA 5: Press the 'timed record' button
SD 5: Flashing cursor moves to 'channel:'
UA 6: Press digit 4
SD 6: Digit is displayed as typed and flashing cursor moves to next position
UA 7: Press the 'timed record' button
SD 7: Flashing cursor moves to 'date:'
UA 8: Press digits 2 4 0 2 0 5
SD 8: Each digit is displayed as typed and flashing cursor moves to next position
UA 9: Press the 'timed record' button
SD 9: Stream number in top right-hand corner of display flashes
UA 10: Press the 'transmit' button
SD 10: Details are transmitted to video player and display returns to normal mode

User's action (UA)
System's display or response (SD)

Having determined our action list we are in a position to proceed with the walkthrough. For each action (1–10) we must answer the four questions and tell a story about the usability of the system. Beginning with UA 1:

UA 1: Press the 'timed record' button

Question 1: Is the effect of the action the same as the user's goal at that point?

The timed record button initiates timer programming. It is reasonable to assume that a user familiar with VCRs would be trying to do this as his first goal.

Question 2: Will users see that the action is available?

The 'timed record' button is visible on the remote control.

Question 3: Once users have found the correct action, will they know it is the one they need?

It is not clear which button is the 'timed record' button. The icon of a clock (fourth button down on the right) is a possible candidate but this could be interpreted as a button to change the time. Other possible candidates might be the fourth button down on the left or the filled circle (associated with record). In fact, the icon of the clock is the correct choice but it is quite possible that the user would fail at this point. This identifies a potential usability problem.

Question 4: After the action is taken, will users understand the feedback they get?

Once the action is taken the display changes to the timed record mode and shows familiar headings (start, end, channel, date). It is reasonable to assume that the user would recognize these as indicating successful completion of the first action.

Heuristic Evaluation

- Proposed by Nielsen and Molich.

- Usability experts review your interface and compare it against accepted usability principles
- The general idea behind heuristic evaluation is that several evaluators independently critique a system to come up with potential usability problems.
- Example heuristics
  - system behaviour is predictable
  - system behaviour is consistent
  - feedback is provided

Heuristic Evaluation

- To aid the evaluators in discovering usability problems, a set of 10 heuristics are provided.
- Each evaluator assesses the system and notes violations of any of these heuristics that would indicate a potential usability problem.
- The evaluator also assesses the severity of each usability problem, based on four factors: how common is the problem, how easy is it for the user to overcome, will it be a one-off problem or a persistent one, and how seriously will the problem be perceived?

Heuristic Evaluation

These can be combined into an overall severity rating on a scale of 0–4:

0 = I don't agree that this is a usability problem at all

1 = Cosmetic problem only: need not be fixed unless extra time is available on project

2 = Minor usability problem: fixing this should be given low priority

3 = Major usability problem: important to fix, so should be given high priority

4 = Usability catastrophe: imperative to fix this before product can be released (Nielsen)

Nielsen's Ten Heuristics

1. Visibility of system status. Always keep users informed about what is going on, through appropriate feedback within reasonable time. For example, if a system operation will take some time, give an indication of how long and how much is complete.

2. Match between system and the real world. The system should speak the user's language, with words, phrases and concepts familiar to the user, rather than system-oriented terms. Follow real-world conventions, making information appear in natural and logical order.

3. User control and freedom. Users often choose system functions by mistake and need a clearly marked 'emergency exit' to leave the unwanted state without having to go through an extended dialog. Support undo and redo.

Nielsen's Ten Heuristics

4. Consistency and standards. Users should not have to wonder whether words, situations or actions mean the same thing in different contexts. Follow platform conventions and accepted standards.

5. Error prevention. Make it difficult to make errors. Even better than good error messages is a careful design that prevents a problem from occurring in the first place.

6. Recognition rather than recall. Make objects, actions and options visible. The user should not have to remember information from one part of the dialog to another. Instructions for use of the system should be visible or easily retrievable whenever appropriate.

Nielsen's Ten Heuristics

7. Flexibility and efficiency of use. Allow users to tailor frequent actions. Accelerators – unseen by the novice user – may often speed up the interaction for the expert user to such an extent that the system can cater to both inexperienced and experienced users.

8. Aesthetic and minimalist design. Dialogs should not contain information that is irrelevant or rarely needed. Every extra unit of information in a dialog competes with the relevant units of information and diminishes their relative visibility.

9. Help users recognize, diagnose and recover from errors. Error messages should be expressed in plain language (no codes), precisely indicate the problem, and constructively suggest a solution.

Nielsen's Ten Heuristics

10. Help and documentation. Few systems can be used with no instructions so it may be necessary to provide help and documentation. Any such information should be easy to search, focussed on the user's task, list concrete steps to be carried out, and not be too large.

*Once each evaluator has completed their separate assessment, all of the problems are collected and the mean ratings calculated.

*The design team will then determine the ones that are the most important and will receive attention first.

Model-Based Evaluation

Certain cognitive and design models provide a means of combining design specification and evaluation into the same framework

Model-Based Evaluation

- GOMS (goals, operators, methods and selection) model predicts user performance with a particular interface and can be used to filter particular design options.
- Design rationale provides a framework in which design options can be evaluated. By examining the criteria that are associated with each option in the design, and the evidence that is provided to support these criteria, informed judgments can be made in the design.
- Dialog models can also be used to evaluate dialog sequences for problems, such as unreachable states, circular dialogs and complexity. Models such as state transition networks are useful for evaluating dialog designs prior to implementation.

Using previous studies in evaluation

- A final approach to expert evaluation exploits this inheritance, using previous results as evidence to support aspects of the design.
- It is expensive to repeat experiments continually and an expert review of relevant literature can avoid the need to do so.
- It should be noted that experimental results cannot be expected to hold randomly across contexts.
- The reviewer must therefore select evidence carefully, noting the experimental design chosen, the population of participants used, the analyses performed and the assumptions made.

Evaluating through user Participation

Laboratory studies

Field studies

Laboratory studies

Advantages

- may contain audio/visual recording and analysis facilities, two-way mirrors, instrumented computers and the like, which cannot be replicated in the work environment.
- The participant operates in an interruption-free environment.

Disadvantages

- Lack of context – for example, filing cabinets, wall calendars, books or interruptions – and the unnatural situation may mean records a situation that never arises in the real world.
- Difficult to observe several people cooperating on a task in a laboratory situation
- However, some situations where laboratory observation is the only option, for example, if the system is to be located in a dangerous or remote location, such as a space station. Also some very constrained single-user tasks may be adequately performed in a laboratory.

Field Studies

- The second type of evaluation takes the **designer or evaluator out into the user's work environment in order to observe the system in action**
- Advantages:
  - natural environment
  - it allows us to study the interaction as it occurs in actual use
- Disadvantages:
  - distractions
  - noise

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
- Variables
  - things to modify and measure
- Hypothesis
  - what you'd like to show
- Experimental design
  - how you are going to do it

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 data- can only take certain values.

Example: the number of students in a class (you can't have half a student).

Continuous data- can take any value (within a range)

Examples: A person's height: could be any value (within the range of human heights), not just certain fixed heights. Time in a race: you could even measure it to fractions of a second,

Experimental studies on groups

More difficult than single-user experiments

Problems with:

- subject groups
- choice of task
- data gathering
- analysis

Experimental studies on groups: Subject groups

Experiments in group working are often longer than the single-user equivalents as we must allow time for the group to 'settle down' and some rapport to develop.

Experimental studies on groups: Experimental task

- Choosing a suitable task is also difficult. We may want to test a variety of different task types: creative, structured, information passing, and so on.
- The tasks must encourage active cooperation, because information and control is distributed among the participants.

Experimental studies on groups: Data gathering

- In group setting - use several video cameras as well as direct logging of the application
- Problem: So for a three-person group, we are trying to synchronize the recording of six or more video sources and three keystroke logs. The technical problems are clearly massive.

Sol:

- Focus on the participants individually, recording, for each one, the video images that are being relayed as part of the system and the sounds that the participant hears.
- These can then be synchronized with the particular participant's keystrokes and additional video observations.

Experimental studies on groups: Analysis

- Problems with taking groups of users and putting them in an experimental situation.
- If the groups are randomly mixed, then we are effectively examining the process of group formation, rather than that of a normal working group.
- Even where a pre-existent group is used, excluding people from their normal working environment can completely alter their working patterns.

Observational Methods

Think Aloud

Cooperative evaluation

Protocol analysis

Automated analysis

Think Aloud

Think aloud is a form of observation where the user is asked to talk through what he is doing as he is being observed; for example, describing what he believes is happening, why he takes an action, what he is trying to do.

- Advantages
  - simplicity - requires little expertise
  - can provide useful insight into problems with an interface
  - can show how system is actually use
- Disadvantages
  - the information provided is often subjective and may be selective, depending on the tasks provided.
  - The process of observation can alter the way that people perform tasks and so provide a biased view

Cooperative evaluation

- variation on think aloud
- user collaborates in evaluation
- both user and evaluator can ask each other questions throughout (typically of the 'why?'

or 'what-if ?' type)

Advantages:

-the process is less constrained and therefore easier to learn to use by the evaluator

-the user is encouraged to criticize the system

-the evaluator can clarify points of confusion at the time they occur and so maximize the effectiveness of the approach for identifying problem areas

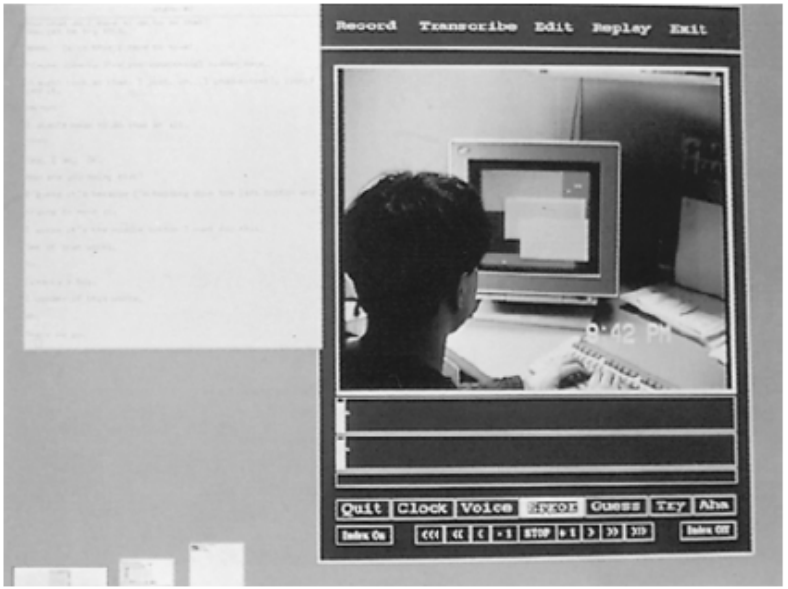
Protocol analysis

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

- Mixed use in practice.
- audio/video transcription difficult and requires skill.
- Some automatic support tools available

automated analysis – EVA

- *EVA* (*Experimental Video Annotator*) is a system that runs on a multimedia workstation with a direct link to a video recorder.
- The evaluator can devise a set of buttons indicating different events. These may include timestamps and snapshots, as well as notes of expected events and errors.
- The buttons are used within a recording session by the evaluator to annotate the video with notes.



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

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

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

Styles of question that can be included in the questionnaire

General These are questions that help to establish the background of the user and his place within the user population. They include questions about age, sex, occupation, place of residence, and so on. They may also include questions on previous experience with computers, which may be phrased as open-ended, multi-choice or scalar questions

Open-ended These ask the user to provide his own unprompted opinion on a question, for example 'Can you suggest any improvements to the interface?'. They are useful for gathering general subjective information but are difficult to analyze in any rigorous way, or to compare, and can only be viewed as supplementary. They are also most likely to be missed out by time-conscious respondents! However, they may identify errors or make suggestions that have not been considered by the designer. A special case of this type is where the user is asked for factual information, for example how many commands were used.

Scalar These ask the user to judge a specific statement on a numeric scale, usually corresponding to a measure of agreement or disagreement with the statement. For example,

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

Styles of question that can be included in the questionnaire

Multi-choice Here the respondent is offered a choice of explicit responses, and may be asked to select only one of these, or as many as apply. For example,

How do you most often get help with the system (tick one)?

- Online manual ❑
- Contextual help system ❑
- Command prompt ❑
- Ask a colleague ❑

Which types of software have you used (tick all that apply)?

- Word processor ❑
- Database ❑
- Spreadsheet ❑
- Expert system ❑
- Online help system ❑
- Compiler ❑

Styles of question that can be included in the questionnaire

Ranked These place an ordering on items in a list and are useful to indicate a user's preferences. For example,

Please rank the usefulness of these methods of issuing a command (1 most useful, 2 next, 0 if not used).

- Menu selection ❑
- Command line ❑
- Control key accelerator ❑

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
  - 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

Figure 9.5 Eye-tracking equipment. Source: Courtesy of J. A. Renshaw

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

Data Lab Psychophysiology equipment showing some of the sensors and a typical experimental arrangement with sensors attached to the participant's fingers and the monitoring software displayed on the evaluator's machine.
Source: Courtesy of Dr R. D. Ward