DEMMS: Evaluation of Multimedia Systems

Robert Villa

February 2008

UNIVERSITY of GLASGOW

Evaluation

- What are the Evaluation lectures about:
  - When to evaluate
  - What kinds of evaluation are possible
    - Predictive evaluations
    - Traditional user experiments
    - Ethnographic style studies
  - Case study describing an example evaluation in detail

Today:

- The role of evaluation
  - Within the larger development effort
- Predictive evaluation
  - Expert reviews
  - Usage simulations
- Traditional user experiments
  - Collecting usage data
- Ethnographic style techniques
  - Very briefly

Next week:

- Lecture, Tuesday 12th Feb:
  - Evaluation case study
- Tutorial, Tuesday 12th Feb:
  - Evaluation case study

What is Evaluation?

- **From HCI:**
  - **"Evaluation is concerned with gathering data about the usability of a design or product by a specified group of users for a particular activity within a specified environment or work context."**
    - **» Preece, page 602**

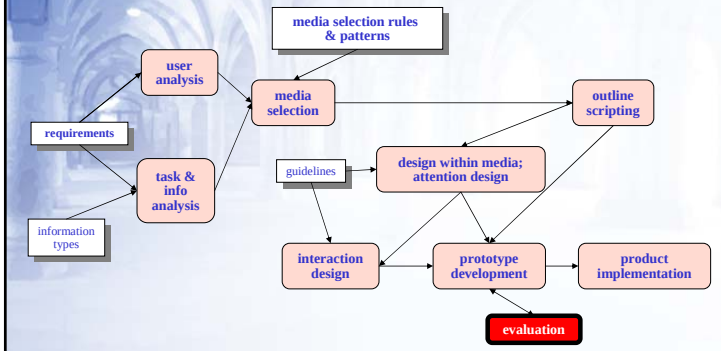
Kinds of Evaluation

- Formative
  - Evaluation which occurs during the design of a product, to guide it's development
  - The principle focus here
- Summative
  - Evaluations which take place after a product has been developed, which judges the finished product

Evaluation within the City Design Method

- The City Design Method has been covered in pervious lectures
  - Dr McGee-Lennon

Prototype development with formative evaluation



Prototyping

- User-centred process
  - Can use storyboards as prototypes for evaluation
  - Mock-ups (few web pages, images, etc.)
- Problems can occur with prototypes
  - False settings (e.g. Ignoring bandwidth issues)

Evaluation in the development life cycle

- Early design stages
  - Predict how well a design works
  - Test out ideas quickly
- Later design stages
  - Identify user difficulties
  - Identify possible improvements
  - Can spend more time on more thorough evaluations

Predictive evaluation

- Does not involve *user testing*
  - Want to try and predict how something works
- Why do it?
  - Quick
  - Cheap

Expert reviews

- A usability expert reviews the system for problems
  - Expert attempts to simulate the behaviour of beginners
- Advantages
  - Efficient: one or two reviewers may identify many problems
  - Experts more forthcoming with information
- Important that the reviewer is not involved with system development

Heuristic evaluation

- Like expert reviews, but inspection is guided by a set of heuristics
  - Heuristics focus on key usability concerns
  - Examples of heuristics:
    - Be consistent
    - Provide clearly marked exits
    - Speak the users' language
    - (Nielsen, 1992)

Walkthroughs

- Determine a task to be done, and the context of the task
  - A expert then "walks through" the task reviewing the actions necessary
- Similar to a review, but with more detailed predictions of what users' do

Simulations

- Given a prototype, automatically simulate users actions with it
  - Requires prototype software
  - Enables a quick "what-if" analysis

Predictive evaluation overview

- Advantages:
  - Relatively fast and cheep (does not require users to test software)
  - Does not require fully working prototypes
  - Can provide allot of feedback from experts
  - May be more appropriate at the start of prototyping and design

Predictive evaluation overview

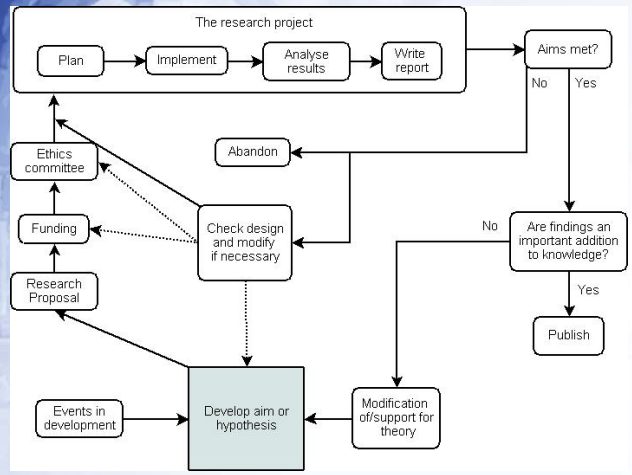
- Disadvantages:
  - The views of experts may not coincide with how your users actually behave
  - Simulations don't necessarily model user's behaviour correctly

User Experiments

- No matter what other kinds of evaluation are carried out, at some point you need to evaluate with real users
  - Traditional lab-based experiments
  - Participative evaluation/design
  - Ethnographic-style work
- Quantitative/Qualitative data

Traditional experiments

- Laboratory setting
- Psychological research is the model
- Generally:
  - Aim is for quantitative results ("hard" evidence)
  - Often relatively narrow domain



Variables

- Independent variables
  - What you manipulate
- Dependent variables
  - Expected to be influenced by the independent variables

Example

- You develop a new type of video browsing interface X. You want to find out if users can browse videos quicker when compared to existing interface Y
- Independent variable:
  - The two different systems X and Y
    - X and Y are the two “levels” of the variable
- Dependent variable:
  - Navigation time

Experimental Design

- Between subject
  - A user does only one condition
- Within subject
  - Users do all conditions
- Matched pairs
  - Users are matched in pairs based on some criteria

Collecting usage data

- Observing users
- Think aloud protocol
- Software logging
- Interviews
- Questionnaires

Observing Users

- Direct observation
  - Watch someone carry out specially devised or normal tasks
  - Obtrusive - Hawthorne effect (1939)
    - Behaviour and performance can be altered when you watch somebody who is aware of being watched

Observing Users (2)

- Indirect observation
  - E.g. video recording or screen recording software
  - Less obtrusive than direct monitoring
- Problems:
  - Lots of data which can be very difficult and time consuming to analyse

Think aloud protocol

- Encourage a user to say out loud what he/she is thinking while carrying out a task
  - Added strain on users (have to talk about what they're doing as well as do it)
  - Can generate lots of feedback about an interface

Software logging

- Software is "instrumented" to generate a time-stamped log of actions
  - Much easier to analyse a log than video
    - E.g. "time on web page" can be calculated if a log contains time stamped browse events
  - Often requires software to be altered
    - Can get general purpose key loggers, browser loggers, etc.

Interviews

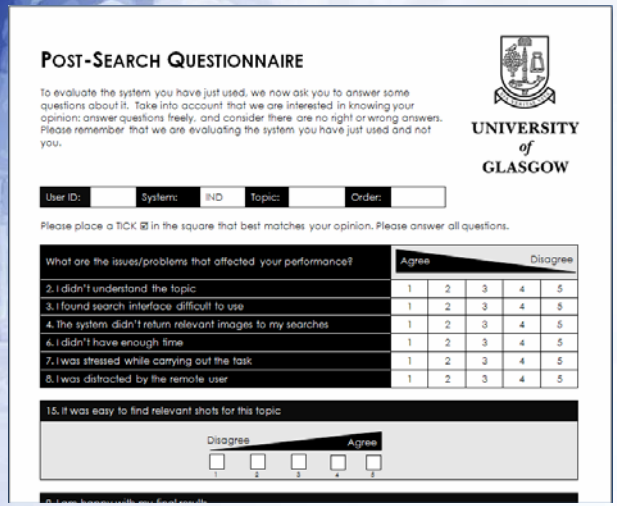
- Structured interviews
  - Predefined questions asked in a set way
    - E.g. Public opinion surveys
  - Important if you want to generate statistics
    - E.g. “X% of people interviewed agreed with ...”
- Flexible interviews
  - Set topics, but interviewer is free to follow interviewee’s replies
  - Often used for requirements gathering and sometimes after more formal evaluations

Questionnaires

- Can be given to a large number of people (e.g. Put on the web)
- Surprisingly difficult to do well
  - Importance is on creating unambiguous questions:
    - Closed questions (multiple choice)
    - Open questions

Questionnaires (cont)

- Different scales can be used in closed questions:
  - Checklist options
    - E.g. Yes/no/don’t know
  - Multi-point rating
    - End points given (e.g very useful/of no use)
  - Likert scale:
    - Multi-point scale where strength of agreement is measured



POST-SEARCH QUESTIONNAIRE

To evaluate the system you have just used, we now ask you to answer some questions about it. Take into account that we are interested in knowing your opinion: answer questions freely, and consider there are no right or wrong answers. Please remember that we are evaluating the system you have just used and not you.

UNIVERSITY of GLASGOW

User ID: | System: IND | Topic: | Order:

Please place a TICK ☑ in the square that best matches your opinion. Please answer all questions.

What are the issues/problems that affected your performance?	Agree				Disagree
2. I didn't understand the topic	1	2	3	4	5
3. I found search interface difficult to use	1	2	3	4	5
4. The system didn't return relevant images to my searches	1	2	3	4	5
6. I didn't have enough time	1	2	3	4	5
7. I was stressed while carrying out the task	1	2	3	4	5
8. I was distracted by the remote user	1	2	3	4	5

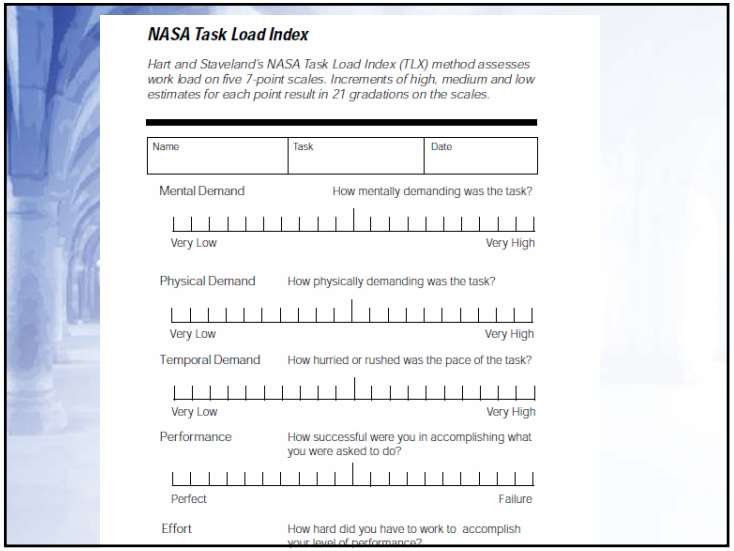
15. It was easy to find relevant shots for this topic

Disagree Agree

1 2 3 4 5

Standard questionnaires

- Standard questionnaires have been developed, which can be re-used
  - NASA-TLX
    - Level of task load of a user
  - QUIS
    - "Questionnaire for user Interaction Satisfaction"
    - Assess user's subjective satisfaction with aspects of a user interface



NASA Task Load Index

Hart and Staveland's NASA Task Load Index (TLX) method assesses work load on five 7-point scales. Increments of high, medium and low estimates for each point result in 21 gradations on the scales.

Name	Task	Date

Mental Demand — How mentally demanding was the task?
Very Low — Very High

Physical Demand — How physically demanding was the task?
Very Low — Very High

Temporal Demand — How hurried or rushed was the pace of the task?
Very Low — Very High

Performance — How successful were you in accomplishing what you were asked to do?
Perfect — Failure

Effort — How hard did you have to work to accomplish your level of performance?

Common Style of Experiment

- Often with Multimedia/HCI experiments:
  - Purpose is to determine if a system or interface is "better" than an old one
  - Within subject designs
  - Independent variables:
    - Two or more "systems" or "interfaces"
    - One or more tasks (e.g. four different search task)
  - Dependent variables:
    - Time
    - Task performance (where it can be measured)

Common Style of Experiment (cont)

- Uses questionnaires:
  - Entry questionnaire:
    - general information about the user (gender, languages, etc.)
  - Post-task questionnaire:
    - user perception of the task/system/etc.
  - Exit questionnaire:
    - User perceptions of the different systems etc.

Next week ...

- We'll go through an example case study

Ethnographic style studies

- Lab evaluations have been criticised:
  - The lab is not like the real world
  - No account of context
  - Artificial tasks
  - Not possible to control everything
- In response, some argue for:
  - ethnographic style studies where researchers study the use of systems in situ

Ethnographic style studies (cont)

- In reality this generally means:
  - The experimenter must go into the work environment and observe users working
- Issues:
  - Takes lots of time
  - Typically generates qualitative rather than quantitative data