Usability Evaluation

Introduction to User Centred Design

Usability Evaluation is *NOT...*

"I showed my program to three different people and they all said it looked really, really good."

Introduction

- Evaluation is used to:
  1. Identify usability problems
  2. Assess whether the GUI design satisfies usability requirements
  3. Evaluate whether the GUI design will be usable in practice by its intended users
- Step 1 should occur during the design
- Steps 2 and 3 occur towards the end to assess the success of the whole design exercise.

Outline

- Analytic Evaluation
- Expert Evaluation
- Observational Evaluation
- Survey Evaluation
- Experimental Evaluation

Analytic Evaluation

A paper-based analysis of a definition of the user interface, sketched in a natural language or some semi-formal language, e.g., Command Grammar Language or GOMS

- GOOD
  - no need to build prototype
  - no need to arrange user testing
- BAD
  - time consuming
  - requires specialists with background in psychology
  - doesn’t tell us anything about errors or learning behaviour

Analytic Evaluation (2)

- Cognitive Walkthrough
  - A type of analytic evaluation
  - A check for identified psychological criteria during a “walkthrough”
  - Evaluate how well the designed software supports the user in learning to use it
  - Performed by expert in cognitive psychology as applied to interface design

GOMS (Goals, Operators, Methods, and Selection rules)

- Eyes/ears perceive information
- Information enters perceptual processor
- Information enters the visual/auditory image store
- Information is stored in the working memory and long term memory
- Information is analyzed in the cognitive processor and a desired reaction (motor function) is chosen
- Desired motor function is activated in the motor processor
- Desired motor function is applied by user's body

GOMS MTM (Methods-Time Measurement)

- Eye fixation = 230[70, 700] milliseconds
- Eye movement = 30 milliseconds
- Perceptual Processor = 100[50, 200] milliseconds
- Cognitive Processor = 70[25, 170] milliseconds
- Motor Processor = 70[30, 100] milliseconds
  - Can also apply Fitts' Law

KLM (Keystroke Level Model)

- Simpler and faster than GOMS
- Average times as measured by Card, Moran and Newell:
  - Press a key or button
    - Best typist = .08 seconds
    - Good typist = .12 seconds
    - Average skilled typist = .20 seconds
    - Average non-secretary = .28 seconds
    - Typing random letters = .50 seconds
    - Typing complex codes = .75 seconds
    - Worst typist = 1.2 seconds
- Point with a mouse (excluding click) = 1.1 seconds
- Move hands to keyboard from mouse (or vice-versa) = .4 seconds
- Mentally prepare = 1.35 seconds

Outline

- Analytic Evaluation
- Expert Evaluation
- Observational Evaluation
- Survey Evaluation
- Experimental Evaluation

Expert Evaluation

- Expert evaluates interface, and decides what is wrong
- Not empirical research
- Expert must not be part of the design team
- Need interface description, task description, user model
- Also called Heuristic Evaluation
- GOOD
  - efficient, quick, rich source of comments
  - often source of solutions as well as problems
- BAD
  - experts have biases
  - experts are not real users

Expert Evaluation (2)

- Similar to Cognitive Walkthrough, except that a set of usability heuristics (“rules of thumb”) rather than raw psychological theories are applied to the design.
- What guidelines are used? (next slide)

Expert Evaluation – Guidelines (1)

From Shneiderman (*Designing the User Interface*):

1. Strive for consistency
2. Enable frequent users to use shortcuts
3. Offer informative feedback
4. Design dialogues to yield closure
5. Offer simple error handling
6. Permit easy reversal of actions
7. Support internal locus of control
8. Reduce short-term memory load

Expert Evaluation – Guidelines (2)

From Nielsen:

1. Visibility of system status
2. Match between system and the real world
3. User control and freedom
4. Consistency and standards
5. Error prevention
6. Recognition rather than recall
7. Flexibility and efficiency of use
8. Aesthetic and minimalist design
9. Help users recognize, diagnose, and recover from errors
10. Help and documentation

Outline

- Analytic Evaluation
- Expert Evaluation
- Observational Evaluation
- Survey Evaluation
- Experimental Evaluation

Observational Evaluation

Observing user behaviour in lab or workplace setting:

- Direct observation (but Hawthorne effect)
  - *observer-expectancy effect*
- Video recording with playback, participative or not
- Verbal protocols: think aloud, question asking, working in pairs (eavesdropping)
- User notebooks or logs
- Software logging
- Wizard of Oz
- Useful anywhere where evidence is needed of recall of commands, planning, understanding of operations, messages, level of performance of system and user, etc.

Observational Evaluation (2)

- GOOD
  - real users
- BAD
  - interference with performance
  - post-task rationalization
  - labour intensive

Outline

- Analytic Evaluation
- Expert Evaluation
- Observational Evaluation
- Survey Evaluation
- Experimental Evaluation

Survey Evaluation - Interviews

- Interview users of the system
- Can be structured (planned list of things to ask),
- or unstructured (topics to cover, but no fixed sequence)
- GOOD
  - can obtain in-depth response from user
  - can enable new issues to emerge
- BAD
  - time consuming (expensive)
  - requires training and skill to carry out

Survey Evaluation (Questionnaires)

- Open questions (Can you suggest any improvements?)
- Closed questions (How useful is this particular feature?)
  - Checklists
  - Multipoint scales, including Likert Scale
  - Semantic differential scale
  - Ranked order
- GOOD
  - cheap to administer to a large number of users
  - easy to analyze - unless unstructured responses are allowed
- BAD
  - time and skill required to develop the questionnaire (or commercial set can be used – but this is expensive)
  - will only uncover what is looked for

Outline

- Analytic Evaluation
- Expert Evaluation
- Observational Evaluation
- Survey Evaluation
- Experimental Evaluation

Experimental Evaluation

- Utilizes the scientific method with a controlled experiment (i.e., testing an hypothesis by measuring attributes of subject behaviour)
- Includes
  - Testing of hypothesis
  - Independent variables (varied by experimenter)
  - Dependant variables (performance measurements)
  - Controlled variables (fixed by experimenter)
  - Can the hypothesis be stated in a way that can be tested?
  - Statistical analysis to check reliability of results
  - Pilot studies

Experimental Evaluation (2)

- GOOD
  - reliable results
- BAD
  - need specialist knowledge
  - resources needed to set up experiment
  - can't be used for every design decision
  - works best for narrow questions

Participants

- Formerly “subjects”
- Participants should match the user population
  - Age
  - Education
  - Experience with computers
  - Experience with systems of that type
  - Experience of the task domain
- Generally at least 10 participants required

Independent / Dependent Variables

- Independent variables
  - Manipulated through the design of the experiment; e.g.,
  - interface style (e.g. GUI vs command-line)
  - level of help (e.g., tool tips vs F1)
  - number of menu items (e.g., 4, 8, 16)
  - icon design (e.g., static vs. animated)
- Dependant variables
  - Performance measurements; e.g.,
    - time to complete a task
    - number of errors made
    - user preferences
    - quality of users performance
  - Must be measurable
  - Must be effected by the independent variable
  - As far as possible, must be unaffected by other factors

Hypothesis

- A prediction of the outcome of an experiment.
- States that variation in the independent variable will cause a difference in the dependent variable.
- Experiment is designed to disprove the null hypothesis (i.e., that there is no difference in the dependant variable between levels of the independent variable)

Experimental Design

Beyond the scope of this course…

- in real-life, an expert is hired to design and perform the experiment

But, to whet your curiosity:

- between-groups
- within-groups
- outliers
- statistics & statistical analysis (*T-tests*, *ANOVA*s, etc.)
- interpreting results
- Science!

Data Analysis: ANOVA etc.

- How much of the variation due to chance
- ANOVA report (e.g., from Excel) "$F(2,11) = 6.2, p < 0.05$" means
  - Variation bw groups 6.2 times larger than within
  - More than 19/20 probability it's not due to chance