Usability Evaluation

Why?

- Organizational perspective:
  - To make a better product
    - Is it usable and useful?
    - Does it improve productivity?
  - Reduce development and support costs
- Designer & developer perspective:
  - Understanding the real world
    - What should it do?
  - Compare designs
  - Is it good enough?
  - Does it do what it should do?
Who?

- Who is your target population?
- Do you want expert usability advice or user feedback?
- Who do you have access to?
- Who are you going to select to evaluate?
- Who will conduct the evaluation?

What?

- What do you want to know?
  - What do users want?
  - Is the product usable? Is it useful?
  - What does the system need to do?
  - How easy is the product to learn?
  - How powerful is the product for expert users?
  - Does it do what it claims to?
  - What problems occur when using the system?
- What scope of the product will you evaluate?
  - Full functionality vs. prototype components
Where?

• Where will the study take place?
  – Field study
  – Lab Study

When?

• When in the design/development cycle will your evaluation take place?
• Will there be several iterations of evaluation?
• How long will it last?
Usability Methods

- Inquiry
- Inspections
- Testing

- www.best.com/~jthom/usability/

Usability Methods - Inquiry

- Contextual Inquiry
  - Understanding the context in which a product is used
  - The user is a partner in the design process
  - The usability design process must have a focus
- Field Observation
  - Artifacts are physical object in use at a site
  - Outcroppings are noticeable physical traits that characterize the site
- Interviews and Focus Groups
  - Ask users about their experiences and preferences with your product
  - Focus groups enable you to identify common problems that people experience.
Usability Methods - Inquiry

• Surveys
  – Ad hoc interviews, where the interviewer asks a set list of questions and records the responses.

• Questionnaires
  – A written list of questions that you distribute to your users

• Journaled Sessions
  – They are often used as a remote inquiry method for software user interface evaluation.

Usability Methods - Inquiry

• Self-Reporting Logs
  – Paper and pencil journals where users log their actions and observations while interacting with a product.

• Screen Snapshots
  – The user takes screen snapshots at different times during the execution of a task or a series of tasks.
Usability Methods - Inspection

• Heuristic Evaluation
  – Usability specialists judge whether each element of a UI follows established usability principles.

• Cognitive Walkthrough
  – The person conducting the walkthrough constructs scenarios from a specification or early prototype and role plays the part of a user working with that interface.

• Formal Usability Inspections
  – Adapts the software inspection methodology to usability evaluation.
  – Typically the roles include: moderator, owner, recorder, and inspectors

Usability Methods - Inspection

• Pluralistic Walkthroughs
  – Meetings where users, developers, and usability professionals step through a scenario, discussing and evaluating each element of interaction.

• Feature Inspection
  – Analyze the feature set of a product, usually given end user scenarios.

• Consistency Inspections
  – These are used to ensure consistency across multiple products from the same development effort.
Usability Methods - Inspection

• Standards Inspections
  – Ensure compliance with industry standards. (e.g. software products for the Windows environment should have common elements, such as the same functions on the File menu, Help menu, etc.)

• Guideline Checklists
  – Usually checklists are used in conjunction with many of the usability inspection methods. (The checklist gives the inspectors a basis by which to compare the product.)

Usability Methods - Testing

• #1 determine what you are trying to find out
  – what do you want to know?
  – describe this goal in a few objectives

• #2 design your test
  – identify your users
    • should match the expected user population as closely as possible
    • determine sample size, cost vs. precision -- rule of thumb at least 10 subjects
  – identify your variables
    • independent variables (controlled)
    • dependent variables (measured)
    • Your design must also account for possible confounding factors
Usability Methods - Testing

• #2 design your test (continued)
  – experimental design
    • how you will order and run your experiment to eliminate non-interesting variables from the analysis
      – between-groups and within-groups
      – random assignment and counter-balancing
  – develop the tasks the users will perform
  – specify the experiment equipment
    – methods and data collection techniques to be used
    – physical set-up
  – Identify required research personnel
  – decide how you will analyze the results recorded
    – e.g. will you generate statistics?

Usability Methods - Testing

• #3 get your users
  – selection from population
    • how narrow will your selection be?
    • selection procedure
    • Gather information from your subjects

• #4 setup the test
  – setup all components of the equipment and make sure it is all working

• #5 run the test and collect the data
  – follow your experimental design EXACTLY

• #6 analyze the data
  – find the big problems first
  – summarize the quantitative and qualitative data gathered
Usability Methods - Testing

• Thinking aloud protocol
  – ask the participant to vocalize his/her thoughts, feelings, and opinions while interacting with the product
• Co-discovery method
  – have two participants perform the task together
• Question-asking protocol
  – prompt the user to vocalize his/her thoughts, feeling and opinions by asking direct questions about the product
• Performance measurement
  – usability test designed to determine hard, quantitative data
  – quantitative data is gathered

Theory vs. Reality

• Time constraints
• Financial constraints
• Technology constraints
• Working with human subjects
• Ethical approval
• Video data
• Planning, planning and more planning
Methodology

- Methodology
  - techniques used to measuring phenomena, manipulate phenomena
  - or control the impact of various phenomena

- Methods
  - the tools (instruments, techniques, and procedures) to gather and analyze information
  - each method can do different things
  - methods are inherently flawed, though each is flawed differently
  - use multiple methods can add strength by offsetting each other’s weaknesses

credible empirical knowledge requires consistency or convergence of evidence across studies based on different methods

Research Strategies

- Research evidence in the social and behavioral sciences
  - involves somebody doing something in some situation
  - actors
    - who
    - humans (individuals, groups, organizations, communities) whose behaviour is being studied
  - behaviour
    - what
    - all aspects of states and actions of the humans that might be of interest
  - context
    - when & where
    - all relevant temporal, locational, and situational features of the environment
3 main criteria

- A. Generalizability
  - how generalizable your results are to a population of actors

- B. Precision
  - how precise your measurement of behaviours is as well as precise control over extraneous factors that are not being studies

- C. Realism
  - how realistic the situation or context of where you gathered the evidences is in relation to the contexts to which you want your evidence to apply

- Try to maximize all three criteria, A, B and C

Field strategies

- study “natural” behaviours

- Field study
  - make direct observations of “natural”, ongoing systems, while intruding on and disturbing those systems as little as possible
    - i.e. ethnography, case studies

- Field experiment
  - works within an ongoing natural system as unobtrusively as possible except it manipulates one major feature
  - manipulates one feature in order to assess effect on other behaviours of the system
  - a compromised strategy, the researcher gives up some of the unobtrusiveness of the plain field study, to gain more precision
Experimental strategies

- Involves concocted rather than natural settings

- Lab experiment
  - deliberately constructs a situation or context, defines the rules, and then has individuals perform tasks related to the concocted system
  - can study behaviours with high precision by controlling extraneous factors

- Experimental simulation
  - construct a situation but make it closely resemble a real situation
    - flight simulators
  - fine line between precision and realism

Respondent strategies

- Systematic gathering of responses

- Sample survey
  - obtain evidence to estimate distribution of some variables with a specific population
  - careful sampling of actors from that population and systematically eliciting responses about the material of interest
    - public opinion surveys

- Judgment study
  - researcher concentrates on obtaining information about the properties of a certain set of stimulus
    - psychophysical
Theoretical strategies

- Non-empirical strategies, does not involve “actors behaving in context”

- Formal theory
  - Researchers focuses on formulating general relations among a number of variables of interest
  - These hypotheses are intended to hold over a broad range of populations

- Computer simulation
  - Attempt to model a real world system
  - The system is complete and closed, without any behavior by any participants
  - Any behaviour outcomes are logical predictions from the theory the researcher built into the model

Strategic issues

- Each strategy has inherent weaknesses and potential strengths

- Since all strategies are flawed, but flawed in different ways, need to use more than one strategy, carefully selected to complement each other in their strengths and weaknesses
Comparison techniques

• baserates
  – how often (at what rate or what proportion of time) does Y occur?
  – if you don’t know how often Y occurs in the general case, then how can you decide whether the rate of Y is some particular case is notably high or low
    • high rate of birth defects among infants born to women who worked jobs involving continual use of video display tubes
  – debate on baserates is common in the behavioural sciences
    • how much is too much?

• differences
  – asks whether Y is present (or at a high value) under conditions where X is present (or at a high value) or vice versa

Comparison techniques (continued)

• correlations
  – asks whether there is a systematic covariation in the values of two (or more) properties
    • if X is high for some case is it likely that Y will also be high for that case
    • and if X is low for some case, is it likely the Y will also be low for that case
  – a high, positive correlation between X & Y means that when X is high, Y is also likely to be high and when X is low, Y is also likely to be low
  – a high, negative correlation between X&Y means high X values will indicate low Y values and vice versa
  – low or no correlation implies that knowing X does not give any indication about the likely value of Y
    • X has no predictive power with respect to Y
    • X & Y do not covary
  – also possible to have a strong, nonlinear correlation
Randomization

• can’t control all potential factors
• random assignment of cases to conditions
  – helps to control these extraneous factors
  – strengthen the credibility of the information
  – effectiveness depends on the number of cases being allocated
  – random allocation procedure: each case must be equally likely to end up in any given combination of conditions
• does not guarantee an equal distribution of any or all potentially extraneous factors, instead makes an unequal distribution highly unlikely (but not impossible)
• if you do end up with an unequal distribution, this could confound your results

Sampling

• Basis for choosing the cases that are to be included in your study out of a larger population of potential cases
  – effects the credibility of your results
  – statistical reasoning requires that the cases in the study be a “random sample” of the population to which the results apply
• “do you have a random sample” vs. “what is the nature of the population of which you actually have a random sample”
• you do not actually “select a random sample”, you select a sample using “a random procedure”
  – no guarantee that the resulting sample will mirror the population
• sample size is important for the credibility of experimental results
Validity

- Internal validity
  - the degree to which the results of a study permit you to make strong inferences about causal relations
    - did it occur by chance?
    - Are other factors causing the effects?
  - Depends on how well you can rule out all other plausible rival hypotheses
- Construct validity
  - how well defined are your theoretical ideas?
- External validity
  - how confident you are that your results will hold upon replication and how generalizable are the results
  - impacted by features of the study (size, nature, sampling, setting, procedures, etc.)
  - no one study has external validity alone

Data collection methods

- Self-Reports
  - i.e., participant fills out a questionnaire
  - participants always know that their behaviour is being recorded
- Observations
  - i.e., Watching someone interact with a piece of software and recording behaviours of interest
  - records of behaviour made by the investigator, an assistant, or by a physical instrument
  - Observations by a Visible Observer
    - participants know that they are being observed
  - Observations by a Hidden Observer
    - participants do not know that they are being observed
Data collection methods (2)

- Archival records
  - analyze material in existing archives
  - records were gathered external to the research activities
  - i.e., public records of births, deaths, marriages, etc.
- Records of Public Behaviour
  - participants were aware that the behaviour was likely to be recorded and used
  - political speeches
- Records of Private Behaviour
  - participant may not know their behaviour will be used later for research (diary)
  - or the results are not affected by the participant’s awareness that the results will become public record (statistical data)

Data collection methods (3)

- Trace measures
  - records of behaviour are made by the behaviour itself but without the participants being aware that they are making the record
  - dog-eared pages
  - like self-reports in that the participants do the recording
  - unlike self-reports in that they are normally not aware that there will be a record of their behaviour and that it will be used for research purposes
Self-reports

- questionnaire responses, interview protocols, rating scales, paper and pencil tests, etc.
- most frequently used type of measure

**Advantages:**
- versatile, low in set-up cost and subsequent cost-per-case
- low "dross rates" (little information gets discarded)

**Disadvantages:**
- potentially reactive
  - the participants are aware their behaviour is being done for the researcher's purposes which may influence how they respond
  - may try to make a good impression
  - give socially desirable answers
  - help the researchers get the answers they are looking for (or hinder)
  - may do this consciously or unconsciously
- as a result, they don’t tend to be very useful forms of evidence

Observations

- watching or recording a participants behaviour

**Advantages:**
- versatile
- data recorded can be very "rich"

**Disadvantages:**
- also potentially reactive
  - the participants are aware their behaviour is being observed which may influence their behaviour
- observer error
- used only on overt behaviour, not on thoughts, feelings or expectations
- costly in time and resources and a high dross rate
- use of hidden observers raises ethical concerns
Trace measures

- physical evidences of behaviour left behind as unintended residue of past behaviour
- Advantages:
  - unobtrusive and therefore not reactive
- Disadvantages:
  - not versatile, not available for many types of data we would like to study
  - difficult to pinpoint the exact cause of the evidence
  - time consuming to gather and process
  - costly and high dross rates
- while this is a potentially strong method, is has been used very little in the social and behavioral sciences

Archival records

- census data, court proceedings, diaries, material from newspapers, magazines, radio, and TV, administrative documents, contracts, etc.
- Advantages:
  - lower cost
  - may be the only way to gain evidence on “past” behaviours or for very large social units
- Disadvantages:
  - reactive (public behaviour)
  - versatility, high doss rates
  - only a loose linkage between the record and the concept represented by it
  - no opportunity to “cross-validate” your findings
Techniques for manipulating variables

• Selection
  – make sure all cases of a condition are alike on a certain variable
    • all six-year olds; all male or female
  – advantage:
    • convenient
  – disadvantage:
    • not a “true experiment”
    • can’t determine why the groups are different, just that the difference may cause an effect

Techniques for manipulating variables (2)

• Manipulation of variables of the system
  – manipulate a given variable and randomly assign participants to each condition
  – advantage:
    • not likely to be costly or time consuming
    • have low dissonance rates
  – disadvantage:
    • applicable for only overt and tangible variables
    • unintended experimental demands (hints as to what the researcher really wants)
      – Hawthorne effect
**Techniques for manipulating variables (3)**

- **Induction**
  - manipulation by less direct interventions
  - #1 misleading instructions to the participant
  - #2 use of false feedback
  - #3 use of experimental confederates
  - advantage:
    - if done correctly, can minimize reactive behaviour
  - disadvantage:
    - involve deception -- raises ethical issues
    - can backfire if the participant figures out the deception

**Observational techniques**

- **Thinking aloud protocol**
  - ask the participant to vocalize his/her thoughts, feelings, and opinions while interacting with the product
- **Co-discovery method**
  - have two participants perform the task together
- **Question-asking protocol**
  - prompt the user to vocalize his/her thoughts, feeling and opinions by asking direct questions about the product
- **Performance measurement**
  - usability test designed to determine hard, quantitative data
  - quantitative data is gathered
Protocol analysis

- field notes (paper and pencil)
  - cheap, flexible
  - hard to get detailed information
  - researcher must be present, obtrusive
- audio recordings
  - can be unobtrusive
  - useful when the user is thinking aloud
  - difficult to record other related information to synchronize later
  - time consuming analysis

Protocol analysis (2)

- video recordings
  - can be unobtrusive
  - can see everything within the camera range
  - limited as to what the camera can look at
  - different perspective than actually being there
  - time consuming analysis
- computer logging
  - cheap, easy, unobtrusive
  - good for longitudinal studies
  - difficult with off-the-shelf software
  - can only give the system perspective (doesn’t know what else is going on in the environment)
  - large amount of data
Protocol analysis (3)

- user notebooks
  - can gather unusual or infrequent feedback
  - useful for longitudinal studies
  - less detail (at a course level)
  - “interpreted” records