Interface Evaluation

Human Computer Interaction

Evaluation Techniques

- Evaluation
  - Tests usability and functionality of system
  - Occurs in laboratory, in the field and/or in collaboration with users
  - Can evaluate both design and implementation
  - Should be considered at all stages in the design life cycle

Determine the goals

- What are the high-level goals of the evaluation?
- Who wants it and why?
- The goals influence the approach used for the study.
- Some examples of goals:
  - Identify the best metaphor on which to base the design.
  - Check to ensure that the final interface is consistent.
  - Investigate how technology affects working practices.
  - Improve the usability of an existing product.

Explore the questions

- *All* evaluations need goals & questions to guide them.
- E.g., the goal of finding out why some customers prefer to purchase paper train tickets rather than e-tickets can be broken down into sub-questions:
  - What are customers' attitudes to these new tickets?
  - Are they concerned about them being accepted?
  - Is the interface for obtaining them poor?
- What questions might you ask about the design of a mobile phone?

Choose the evaluation approach & methods

- The evaluation *approach* influences the *methods* used, and in turn, how data is collected, analysed and presented.
- What are the goals of the system?
- Do we have specific measures like time on task, errors, click-through rates we wish to optimize?
- Are we worried about how the system will work in a real situation?

Practical issues

- Select users.
  - People in key roles are often really busy.
- Stay on budget
- Stay on schedule
- Find evaluators
- Select equipment

Decide about ethical issues

- Develop an informed consent form
- Participants have a right to:
  - Know the goals of the study;
  - Know what will happen to the findings;
  - Privacy of personal information;
  - Leave when they wish;
  - Be treated politely.
- Data protection
  - Transparency, purpose limitation, data minimization, accuracy, storage limitation, confidentiality.

Evaluate, interpret & present data

- The approach and methods used influence how data is evaluated, interpreted and presented.
- The following need to be considered:
  - Reliability: can the study be replicated?
  - Validity: is it measuring what you expected?
  - Biases: is the process creating biases?
  - Scope: can the findings be generalized?
  - Ecological validity: is the environment influencing the findings?

Laboratory studies

- Advantages:
  - specialist equipment available
  - uninterrupted environment
- Disadvantages:
  - lack of context
  - difficult to observe several users cooperating
- Appropriate
  - if system location is dangerous or if it is impractical to observe features of interest in real setting.

Field Studies

- Often involve observation and interviews.
  - Involve both qualitative and quantitative data.
- Advantages:
  - natural environment
  - context retained (though observation may alter it)
  - longitudinal studies possible (intervals, long period).
- Disadvantages:
  - distractions
  - noise
- Appropriate where context is crucial and longitudinal studies

Heuristic Evaluation

- Proposed by Nielsen and Molich.
- Usability criteria (heuristics) are identified
- Design examined by experts to see if these are violated
- Example heuristics
  - system behaviour is predictable
  - system behaviour is consistent
  - feedback is provided
- Heuristic evaluation ‘debugs’ design.

Nielsen's Heuristics

1. **Visibility of system status**

   The system should always keep users informed about what is going on, through appropriate feedback within reasonable time.

2. **Match between system and the real world**

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

3. **User control and freedom**

   Users often choose system functions by mistake and will need a clearly marked "emergency exit" to leave the unwanted state without having to go through an extended dialogue. Support undo and redo.

Heuristics (2)

4. **Consistency and standards**

   Users should not have to wonder whether different words, situations, or actions mean the same thing. Follow platform conventions.

5. **Error prevention**

   A careful design can prevent problems 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 dialogue to another. Instructions for use of the system should be visible or easily retrievable whenever appropriate.

7. **Flexibility and efficiency of use**

   Accelerators -- unseen by the novice user -- may often speed up the interaction for the expert user such that the system can cater to both inexperienced and experienced users. Allow users to tailor frequent actions.

Heuristics (3)

8. **Aesthetic and minimalist design**

   Dialogues should not contain information which is irrelevant or rarely needed. Every extra unit of information in a dialogue 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.

10. **Help and documentation**

    Even though it is better if the system can be used without documentation, it may be necessary to provide help and documentation. Any such information should be easy to search, focused on the user's task, list concrete steps to be carried out, and not be too large.

Guidelines for ambient displays

- Sufficient information design – convey "just enough" information. Too much -> cramped display, too little, less useful system.
- Useful and relevant information. The information should be useful and relevant to the users in the intended setting.
- Easy transition to more in-depth information.
- Display should be unobtrusive and remain so unless it requires the user's attention.
- Users should not have to remember what states or changes in the display mean.

From Mankoff et al., ACM CHI.

Guidelines for tabletop displays

- Support interpersonal interaction.
- Support fluid transitions between activities.
- Support transitions between personal and group work.
- Support transitions between tabletop collaboration and external work
- Support the use of physical objects.
- Provide shared access to physical and digital objects
- Consideration for the appropriate arrangements of users.
- Support simultaneous user actions.

Scott et al., ACM CHI.

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

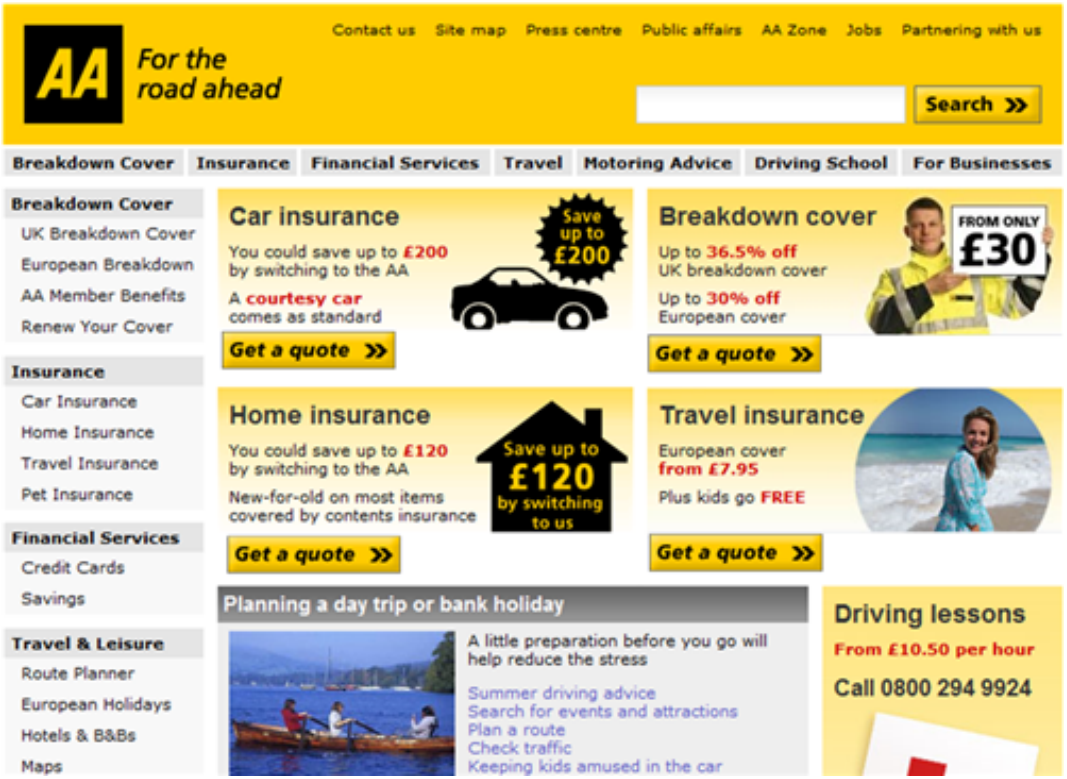
Experimental Factors

- Subjects
  - representative
  - sufficient sample (depends on objective)
- Variables
  - independent variable (IV)
    - characteristic changed to produce different conditions.
    - e.g. interface style, number of menu items.
  - dependent variable (DV)
    - characteristics measured in the experiment
    - e.g. time taken, number of errors.

Experimental Factors (cont.)

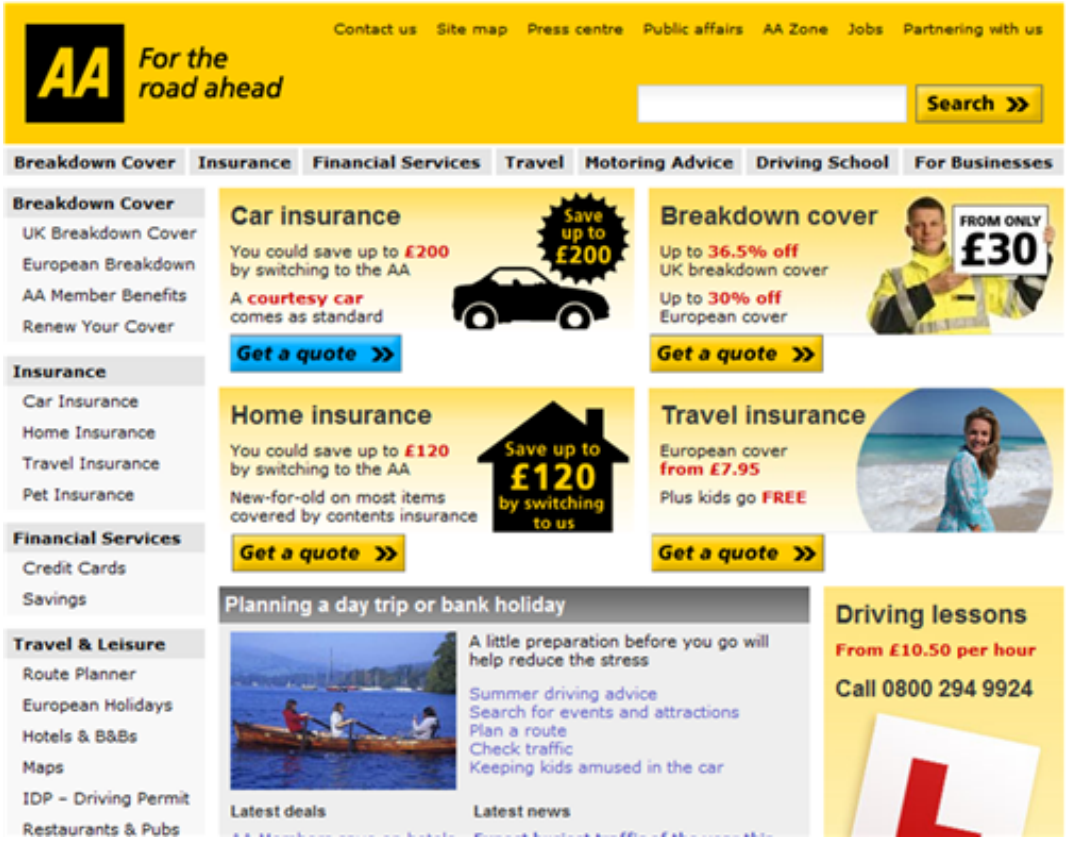
- Hypothesis
  - prediction of outcome framed in terms of IV and DV
  - null hypothesis: states no difference between conditions
  - aim is to disprove this.
- Within subjects (repeated measures) design
  - each subject performs experiment under each condition.
  - transfer of learning possible
  - less costly and less likely to suffer from user variation.
- Between subjects (randomised) design
  - each subject performs under only one condition
  - no transfer of learning
  - more users required
  - variation can bias results.

Example

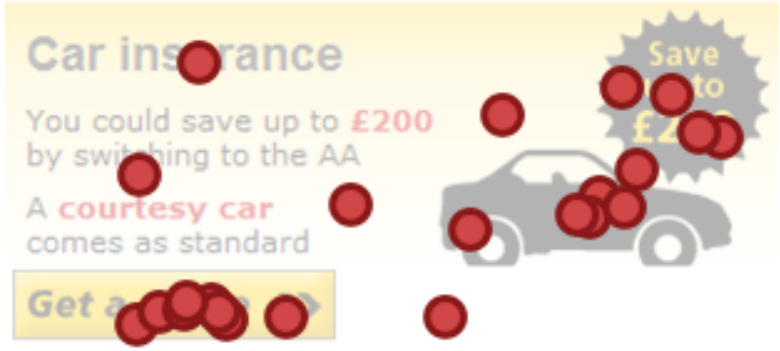


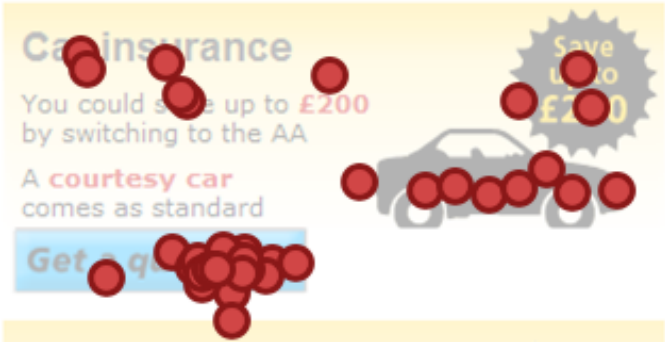
From http://www.goodusability.co.uk/

Example


AA
For the road ahead
Contact us
Site map
Press centre
Public affairs
AA Zone
Jobs
Partnering with us
Search
Breakdown Cover
Insurance
Financial Services
Travel
Motoring Advice
Driving School
For Businesses
Breakdown Cover
UK Breakdown Cover
European Breakdown
AA Member Benefits
Renew Your Cover
Insurance
Car Insurance
Home Insurance
Travel Insurance
Pet Insurance
Financial Services
Credit Cards
Savings
Travel & Leisure
Route Planner
European Holidays
Hotels & B&Bs
Maps
IDP – Driving Permit
Restaurants & Pubs
Car insurance
You could save up to £200 by switching to the AA
A courtesy car comes as standard
Save up to £200
Get a quote
Breakdown cover
Up to 36.5% off UK breakdown cover
Up to 30% off European cover
FROM ONLY £30
Get a quote
Home insurance
You could save up to £120 by switching to the AA
New-for-old on most items covered by contents insurance
Save up to £120 by switching to us
Get a quote
Travel insurance
European cover from £7.95
Plus kids go FREE
Get a quote
Planning a day trip or bank holiday
A little preparation before you go will help reduce the stress
Summer driving advice
Search for events and attractions
Plan a route
Check traffic
Keeping kids amused in the car
Latest deals
Latest news
Driving lessons
From £10.50 per hour
Call 0800 294 9924

What do people notice?





Contingency table

Button	Noticed	Didn't notice	Total
Yellow	9	51	60
Blue	21	39	60
Total	30	90	120

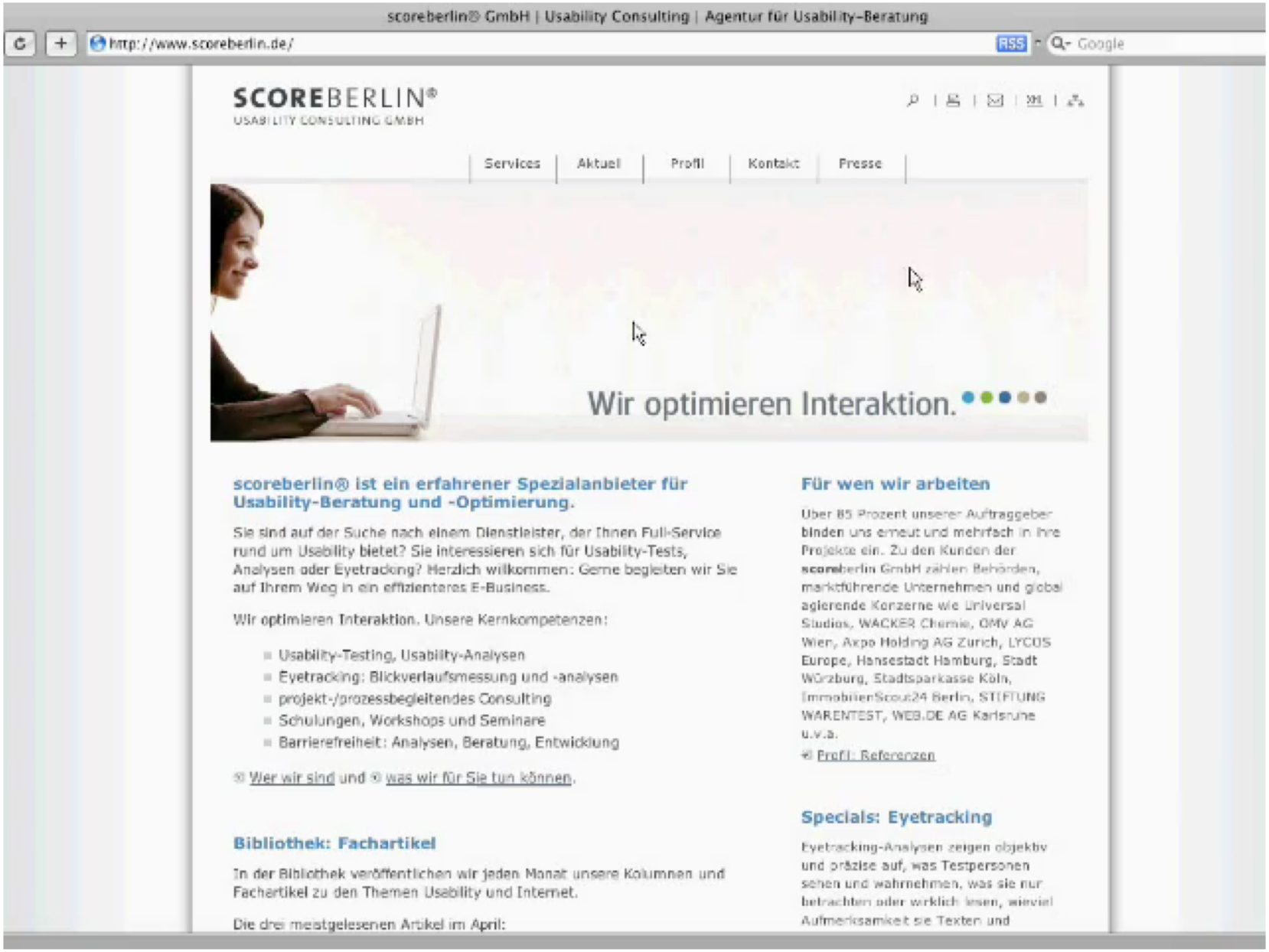
Analysis of data

- Before you start to do any statistics:
  - look at data (outliers?)
  - save original data
- Choice of statistical technique depends on
  - type of data
  - information required
- Type of data
  - discrete - finite number of values
  - continuous - any value

Analysis of data (cont.)

- Contingency table
  - Classify data by several discrete attributes
  - Count number of data items in each group
- What information is required?
  - is there a difference?
  - how big is the difference?
  - how accurate is the estimate?

Eyetracking video



Eye tracking

Observation Methods - Think Aloud

- User observed performing task
- User asked to describe what he is doing and why, what he thinks is happening etc.
- Advantages
  - simplicity - requires little expertise
  - can provide useful insight
  - can show how system is actually used
- Disadvantages
  - subjective
  - selective
  - act of describing may alter task performance

Usefulness of think-aloud protocol

- Think-aloud protocols are of particular value because they focus on the problems a user has;
- When the user is working without difficulty, it may be hard to keep up with them as the user may be unable to communicate as fast as they think and act, unless a specific problem arises which slows them down.
- It is at these times when this method really shines as it allows the observer to correlate the actions and statements of the participant.

Advantages

- Rapid, high-quality, qualitative user feedback (e.g. as compared with questionnaires).
- Data is available from a broad range of sources, such as:
  - Direct observation of what the subject is doing.
  - Hearing what the subject wants, or is trying, to do.
- If the subject gets into difficulties, the observer has the chance to clarify the situation.
- High degree of flexibility; the experiment may easily be steered by the observer.
- The presence of two people allows meaningful, direct dialogue.

Analysis

- Usually done informally at the time by careful note taking but a more quantitative content analysis can be done by examining video evidence after the experiment.
- The purpose of the experiment, e.g. to examine in detail a specific 'corner' of the interface, or to gain an outline of its general efficacy, should be decided upon initially.
- This will allow the observer to choose one of the scenarios given above, and to tailor the protocol accordingly.

Procedure

- Make sure that user realises that the interface, not they, are under scrutiny.
- That the user should at all times comment liberally on his/her actions, intentions and thoughts.
- That the user is at ease. This involves explaining that you may give only a bare minimum of help to the user, and apologising in advance for this. The user should try to find their own way as much as possible.
- Any help given to the user should be carefully thought out, in order for its effects to be recorded as part of the experiment.

Note taking

- The main task of the observer is to jot down what happens.
- The moment when an observer notes what a subject is doing/saying only occurs once and thus needs to be recorded.
- This can be aided by a structured data sheet, can include categories to observe and also prompts.
- Note taking is still relevant even when videoing the experiment as the video representation cannot capture everything and there is no way of clarifying ambiguities when watching a video.

Direction

- Furthermore, the protocol may be used in two distinct scenarios.

1. The observer specifies a definite task to be accomplished by the subject.
   - This allows the observer to concentrate on a specific task they are interested in.
2. 'Open-ended'; no task is specified, and the user is free to choose their own task.
   - Allows the observer to concentrate on naturally occurring problems.

Use of video recording

- Useful because session can be recorded for later analysis
- With the observers usual position beside the participant they may miss something
- **BUT**, for the reasons specified above, note taking is still vital
- An effective alternative to the videoing of a session may be to simply repeat the session with different participants

Use of Prompts

- There is a considerable difference between prompting and biasing the user, basically say as much as necessary to keep user happy without helping and making suggestions. Such questions may include:
  - What are you thinking now?
  - Why did you do that?

Providing help

- If a subject is completely stuck a decision may have to be made whether to help the participant or not. It is up to the judgement of the observer, but if a decision is made to help the following should be noted:
  - Ask what the user would do if the observer wasn't there
  - Take a note of what you said and what happened afterwards
  - Was the problem solved?

Observation Methods - Cooperative evaluation

- Variation on think aloud
- User collaborates in evaluation
- Both user and evaluator can ask each other questions throughout
- Additional advantages
  - less constrained and easier to use
  - user is encouraged to criticise system
  - clarification possible

Observation Methods - Protocol analysis

- Record of evaluation session known as *protocol*.
- Paper and pencil
  - cheap, limited to writing speed
- Audio
  - good for think aloud, difficult to match with other protocols
- Video
  - accurate and realistic, obtrusive
- Computer logging
  - automatic and unobtrusive, large amounts of data difficult to analyse

Protocol Analysis (contd).

- User notebooks
  - coarse and subjective, useful insights, good for longitudinal studies
- Mixed use in practice.
- Transcription of audio and video difficult and requires skill.
- Many automatic support tools available

Protocol Analysis (contd.)

- Post task walkthrough
  - user explains each action after the event
  - used to fill in intention
- Advantages
  - analyst has time to focus on relevant incidents
  - avoid excessive interruption of task
- Disadvantages
  - lack of freshness
  - may be post-hoc interpretation of events

Query Techniques - 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

Query Techniques - Questionnaires

- Set of fixed questions given to users
- Advantages
  - quick and reaches large user group
  - can be analysed more rigorously
- Disadvantages
  - less flexible
  - less probing

Questionnaires (contd.)

- Need careful design
  - what information is required?
  - how are answers to be analyzed?
- Styles of question
  - general (eg. user profile)
  - open-ended (Can you suggest any improvements?)
  - scalar (eg. Likert scale).
  - multi-choice
  - ranked (good for preferences, forces decision).
- Pilot study (comprehensible etc.).

System Usability Scale

Participants asked to score 10 items from Strongly Agree to Strongly disagree:

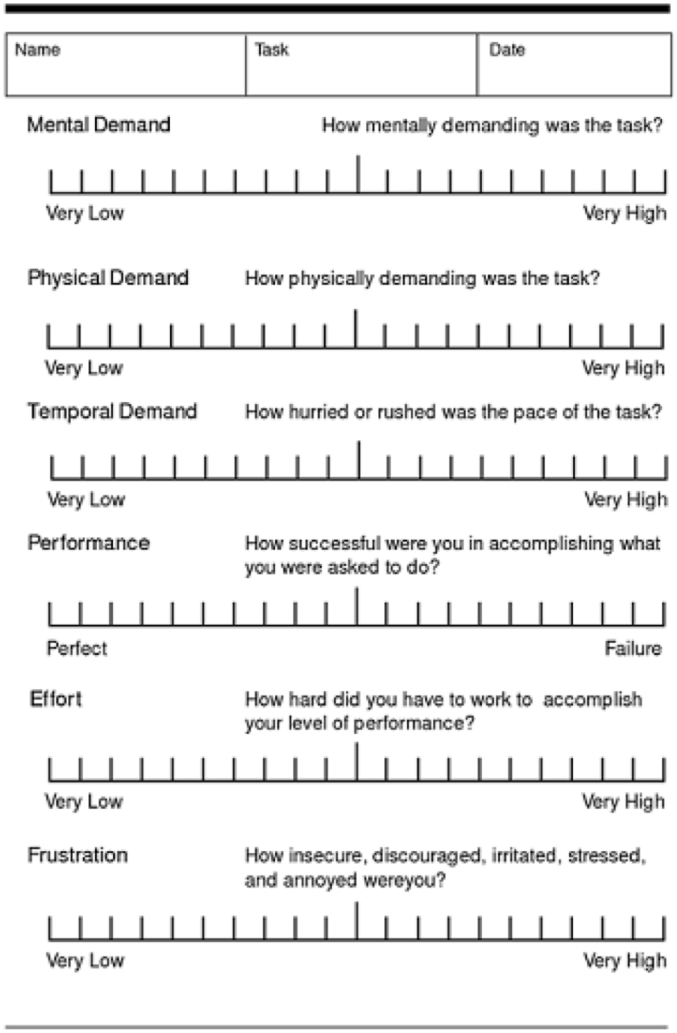
- I think that I would like to use this system frequently.
- I found the system unnecessarily complex.
- I thought the system was easy to use.
- I think that I would need the support of a technical person to be able to use this system.
- I found the various functions in this system were well integrated.
- I thought there was too much inconsistency in this system.
- I would imagine that most people would learn to use this system very quickly.
- I found the system very cumbersome to use.
- I felt very confident using the system.
- I needed to learn a lot of things before I could get going with this system.

NASA Task Load Index (TLX)

Figure 8.6

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.



- Provides an overall workload rating based on 6 subscales.
- Developed over a 3 year project including 40 experiments.

Choosing an Evaluation Method

- When in cycle is evaluation carried out? design vs implementation
- What style of evaluation is required? laboratory vs field
- How objective should the technique be? subjective vs objective – multiple evaluators.
- What type of measures are required? qualitative vs quantitative
- What level of information is required? high level vs low level
- What level of interference? obtrusive vs unobtrusive
- What resources are available? time, subjects, equipment, expertise

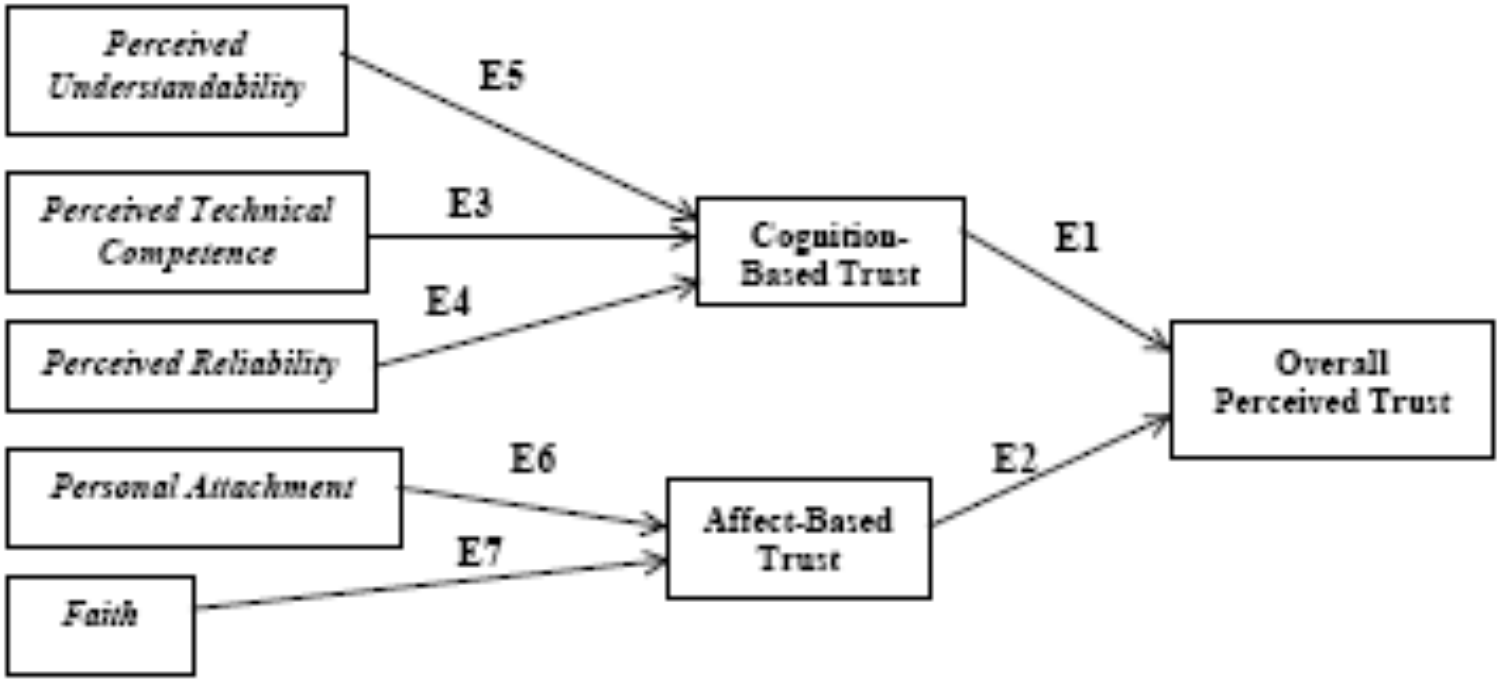
Evaluating user experience?

- How easy was it to learn how to use and subsequently use?
- Was it effective? Did people ever get confused?
- How enjoyable, pleasurable, engaging was it?
- What errors, mistakes, frustrations and misunderstandings were experienced?
- Did people notice it?
- Was it informative?
- How much time did people spend switching between the system and other things of interest?

Trust

- ***Madsen and Gregor (2000) define human computer trust as the extent to which a user is confident in, and willing to act on the basis of, the recommendations, actions, and decisions of an artificially intelligent decision aid.***
- Trust involves both the user's confidence in the system and their willingness to act on the system's decisions and advice

Trust



Madsen and Gregor, 2000.

Human Computer Trust Scale

1. ***Perceived Reliability***
   - R1 - The system always provides the advice I require to make my decision.
   - R2 - The system performs reliably.
   - R3 - The system responds the same way under the same conditions at different times.
   - R4 - I can rely on the system to function properly.
   - R5 - The system analyzes problems consistently.
2. ***Perceived Technical Competence***
   - T1 - The system uses appropriate methods to reach decisions.
   - T2 - The system has sound knowledge about this type of problem built into it.
   - T3 - The advice the system produces is as good as that which a highly competent person could produce.
   - T4 - The system correctly uses the information I enter.
   - T5 - The system makes use of all the knowledge and information available to it to produce its solution to the problem.

HCT scale

3. ***Perceived Understandability***
   - U1 - I know what will happen the next time I use the system because I understand how it behaves.
   - U2 - I understand how the system will assist me with decisions I have to make.
   - U3 - Although I may not know exactly how the system works, I know how to use it to make decisions about the problem.
   - U4 - It is easy to follow what the system does.
   - U5 - I recognize what I should do to get the advice I need from the system the next time I use it.
4. ***Faith***
   - F1 - I believe advice from the system even when I don't know for certain that it is correct.
   - F2 - When I am uncertain about a decision I believe the system rather than myself.
   - F3 - If I am not sure about a decision, I have faith that the system will provide the best solution.
   - F4 - When the system gives unusual advice I am confident that the advice is correct.
   - F5 - Even if I have no reason to expect the system will be able to solve a difficult problem, I still feel certain that it will.

HCT scale

5. *Personal Attachment*

- P1 - I would feel a sense of loss if the system was unavailable and I could no longer use it.
- P2 - I feel a sense of attachment to using the system.
- P3 - I find the system suitable to my style of decision making.
- P4 - I like using the system for decision making.
- P5 - I have a personal preference for making decisions with the system.

Trust checklist

Checklist for Trust between People and Automation

Below is a list of statement for evaluating trust between people and automation. There are several scales for you to rate intensity of your feeling of trust, or your impression of the system while operating a machine. Please mark an "x" on each line at the point which best describes your feeling or your impression.

(Note: not at all=1; extremely=7)

1 The system is deceptive

1 2 3 4 5 6 7

2 The system behaves in an underhanded manner

1 2 3 4 5 6 7

3 I am suspicious of the system's intent, action, or outputs

1 2 3 4 5 6 7

4 I am wary of the system

1 2 3 4 5 6 7

5 The system's actions will have a harmful or injurious outcome

1 2 3 4 5 6 7

6 I am confident in the system

1 2 3 4 5 6 7

7 The system provides security

1 2 3 4 5 6 7

8 The system has integrity

1 2 3 4 5 6 7

9 The system is dependable

1 2 3 4 5 6 7

10 The system is reliable

1 2 3 4 5 6 7

11 I can trust the system

1 2 3 4 5 6 7

12 I am familiar with the system

1 2 3 4 5 6 7

Jian et al. United States Airforce Laboratory Technical Report & International Journal of Cognitive Ergonomics.

How would you evaluate...

A website for selling consumer electronics?

A videogame?

A handheld system for a parcel delivery service?

A location based (mobile) game?