Evaluation Techniques

From Preece, Rogers & Sharp's *Interaction Design*

029511-1
2013년 가을학기
12/05/2013
박경신

Why, What, Where and When to Evaluate

- Iterative design and evaluation is a continuous process that examines
  - Why – to check users' requirements and that users can use the product and they like it
  - What – a conceptual model, early prototypes of a new system and later, more complete prototypes
  - Where – in natural and laboratory settings
  - When – throughout design; finished products can be evaluated to collect information to inform new products

Evaluation

- Evaluation tests usability and functionality of system
- Should be considered at all stages in design life cycle
  - Not at the end if time permits
  - Evaluates both design and implementation
- In the laboratory or in the field or analytical
  - **Controlled settings**, e.g. **usability testing** & experiments in laboratories and living labs
  - **Natural settings**, e.g. **field studies** to see how the product is used in the real world
  - **Any settings without users**, e.g. consultants critique; to predict analyze & model aspects of the interface **analytics**
- With or without collaboration with users
  - Usually, by designers in the early stage
  - Then, with actual users in the later stage

Goals of Evaluation

- Assess extent and accessibility of **system functionality**
  - In accordance with users' requirements
  - Robustness: task conformance, observability, reachability, ...
- Assess **user's experience** with the interaction
  - Learnability
  - User's satisfaction (enjoyable?)
- **Identify problems** with the system
  - Related to both functionality and usability
  - "Specifically concerned with identifying trouble spots"

Characteristics by Evaluation Approaches

	Usability Testing	Field Studies	Analytical
Users	Do task	Natural	Not involved
Where	Controlled	Natural	Anywhere
When	Prototype	Early	Prototype
Measurement Data	Quantitative	Qualitative	Problems
Feedback	Measures & Errors	Descriptions	Problems
Type	Applied	Naturalistic	Expert

	Usability Testing	Field Studies	Analytical
Observing users	O	O	
Asking users	O	O	
Asking experts		O	O
Testing	O		
Modeling			O

Evaluation Through "Expert Analysis"

- Usually, but not necessarily, in the early design cycles
- By designers and human-factor experts
- Based on cognitive principles and empirical results
- Approaches
  - **Cognitive walkthrough**
  - **Heuristic evaluation**
  - **Model-based evaluation**
  - **Using previous studies in evaluation**

Cognitive Walkthrough

- Proposed by Polson et al.
- **Usually performed by experts in cognitive psychology**
- **Expert walks through with a "detailed review" of a sequence of actions**
  - Sequence of actions are steps to perform to accomplish some known task
- **"The main focus is on how easy a system is to learn"**
  - The focus is on learning through exploration
- Evaluators provide a story about why that step is or is not good for a new user.

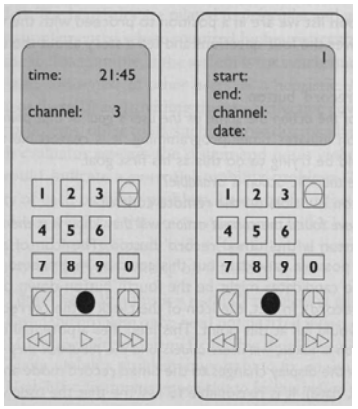
Information to Prepare for Cognitive Walkthrough

- A description of the 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.
- 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.
- A complete, written list of the actions needed to complete the task with the given prototype.
- An indication of who the users are and what kind of experience and knowledge the evaluators can assume about them.

For Each Task, Walkthrough Considers

- Is the effect of the action the same as the user's goal at that point?
  - In other words, users' assumption about an action is correct?
- Will user see that the action is available?
  - E.g., Is a PIP(Picture-In-Picture) button visible on a TV remote?
- Once users have found the correct item, will they know it is the one they need?
  - E.g., Can users recognize a PIP button when it is visible?
- After the action is taken, will users understand the feedback they get?

Example: Programming a Video Recorder by Remote Control



An initial remote control design

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

Action sequence

Heuristic Evaluation

- Proposed by Nielsen and Molich in the early 1990s
- **Critique a system using a set of simple heuristics**
  - Heuristics are guidelines and principles in design
- Can be done with specifications or with prototypes of different levels
- **Several experts, independently (3 to 5), access a system and note violation of any of heuristics**
- Severity rating on a scale of 0 – 4
  - 0 = Not a usability problem at all
  - 1 = Cosmetic problem only
  - 2 = Minor usability problem
  - 3 = Major usability problem
  - 4 = Usability catastrophe

Nielsen's 10 Heurisitics

- Visibility of system status
- Match between system and the real world
- User control and freedom
- Consistency and standards
- Error prevention
- Recognition rather than recall
- Flexibility and efficiency of use
- Aesthetic and minimalist design
- Help users recognize, diagnose and recover from errors
- Help and documentation

3 Stages for Doing Heuristic Evaluation

- Briefing session to tell experts what to do
- Evaluation period of 1-2 hours in which
  - Each expert works separately
  - Take one pass to get a feel for the product
  - Take a second pass to focus on specific features
- Debriefing session in which experts work together to prioritize problems.

Cons and Pros of Heuristic Evaluation

- Best experts have knowledge of application domain and users
- But
  - Can be difficult and expensive to find experts
  - Few ethical and practical issues to consider because users not involved
- Biggest problems
  - Importance problems may get missed
  - Many trivial problems are often identified
  - Experts have biases

Model-Based Evaluation

- Simulation by combining a cognitive model and a design model
- Pros
  - Fast – Evaluation is done in the computer
  - Cheap – No actual participants to pay
- Cons
  - A model is a model; a model cannot capture every aspect of an actual user
- This means it can be used to filter out obvious design problems

GOMS (Goal, Operators, Methods, Selected Rules)

- **Goals**
  - What the user wants to achieve
  - E.g., close-window
- **Operators**
  - Basic actions (visible or not) user performs
  - E.g., press-key, find-command, ...
- **Methods**
  - "Ways to decompose" a goal into sub-goals/operators
  - E.g., menu-method, hotkey-method
- **Selected Rules**
  - Means of choosing between competing methods

GOMS Example

- GOAL: CLOSE-WINDOW
  [select GOAL: USE-MENU-METHOD
  . MOVE-MOUSE-TO-FILE-MENU
  . PULL-DOWN-FILE-MENU
  . CLICK-OVER-CLOSE-OPTION
  GOAL: USE-CTRL-W-METHOD
  . PRESS-CONTROL-W-KEYS]

For a particular user:
Rule 1: Select USE-MENU-METHOD unless another rule applies
Rule 2: If the application is GAME, select CTRL-W-METHOD

Keystroke Level Model (KLM)

- Lowest level of (original) GOMS
- **Six execution phase operators**
  - **Physical motor**
    - **K – Keystroking**
    - **P – Pointing**
    - **H – Homing**
    - **D – Drawing**
  - **Mental**
    - **M - Mental preparation**
  - **System**
    - **R – Response**
- **Times are empirically determined.**
  - **T_execute = TK + TP + TH + TD + TM + TR**

KLM Example

- GOAL: ICONISE-WINDOW
[select GOAL: USE-CLOSE-METHOD
  . MOVE-MOUSE-TO- FILE-MENU
  . PULL-DOWN-FILE-MENU
  . CLICK-OVER-CLOSE-OPTION
  GOAL: USE-CTRL-W-METHOD
  PRESS-CONTROL-W-KEY]
- Compare alternatives:
  - USE-CTRL-W-METHOD vs.
  - USE-CLOSE-METHOD
- Assume hand starts on mouse

USE-CTRL-W-METHOD		USE-CLOSE-METHOD	
H[to kbd]	0.40	P[to menu]	1.1
M	1.35	B[LEFT down]	0.1
K[ctrlW key]	0.28	M	1.35
		P[to option]	1.1
		B[LEFT up]	0.1
Total	**2.03 s**	**Total**	**3.75 s**

Review-Based Evaluation

- Uses results from the literature to support or refute parts of design.
- Instead of having to develop expensive or time consuming experiments
- Care needed to ensure results are transferable to new design (different context).
- Similar to heuristic evaluation?
  - **A review is more specific to a particular context than principles and guidelines!**

Evaluation Through User Participation

- **Usually in the later stages** when there is at least a working prototype of the system is available
- Styles of evaluation
  - Laboratory studies
  - Field Studies
- Empirical methods
  - Participants
  - Variables
  - Hypotheses
  - Experimental design
  - Statistical measures

Laboratory Studies

- Goals and questions focus on how well users perform tasks with the product
- Comparison of products or prototypes common
- **Focus is on time to complete task and number types of errors**
- Data collected by video and interaction logging
- Testing is central
- User satisfaction questionnaires and interviews provide data about users' opinions

Usability Lab with Observers Watching a User & Assistant

Cons and Pros of Laboratory Studies

- Advantages:
  - Specialist equipment available: Recording equipment, two-way mirrors, instrumented computers, …
  - Uninterrupted environment
- Disadvantages:
  - Lack of context, e.g., filing cabinets, wall calendars, books, interruptions, …
  - Difficult to observe several users cooperating
- Appropriate
  - If system location is dangerous or impractical
  - For constrained single user systems
  - For controlled experiments

Field Studies

- **Field studies are done in natural settings**
- "In the wild" is a term for prototypes being used freely in natural settings
- Aim to understand what users do naturally and how technology impacts them
- Field studies are used in product design to
  - Identify opportunities for new technology
  - Determine design requirements
  - Decide how best to introduce new technology
  - Evaluate technology in use

Cons and Pros of Field Studies

- Advantages:
  - Natural environment
  - Context retained (though observation may alter it)
  - Longitudinal (long-term) studies possible
- Disadvantages:
  - Distractions
  - Noisy
- Appropriate
  - Where context is crucial
  - For longitudinal studies

Experimental Evaluation

- Controlled evaluation of specific aspects of interactive behavior
- 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 behavioral measure are attributed to different conditions

Experimental Factors

- Participants
  - 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

Participants

- Match the expected user population
  - Age & sex
  - Level of education
  - Experience with computers
  - Knowledge of task domain, ...
- Sample size large enough for statistically significant conclusion
  - **5-10 users typically selected**
  - 5 may be good enough to reveal usability problems
  - 10 may be good enough for many statistical analysis
- **Informed consent form (IRB)** explains procedures and deals with ethical issues

Variables

- **Independent variable (IV)**
  - Characteristic changed to produce different **conditions**
  - E.g., interface style, level of help, number of menu items, icon design, ...
- **Dependent variable (DV)**
  - Characteristics measured in the experiment
  - E.g., the speed of menu selection, ...
  - **E.g., time taken, number of errors, user preference, quality of user's performance, ...**

Hypothesis

- Prediction of outcome
  - Framed in terms of IV and DV
  - E.g. "Error rate will increase as font size decreases"
- **Null hypothesis**
  - States no difference between conditions
  - Aim is to disprove this
  - E.g., Null hypothesis = "No change with font size"

Experimental Design

- **Within groups design**
  - Each subject performs experiment under each condition.
  - Transfer of learning possible
  - Less costly and less likely to suffer from user variation.
- **Between groups design**
  - Each subject performs under only one condition
  - No transfer of learning
  - More users required
  - Variation can bias results.

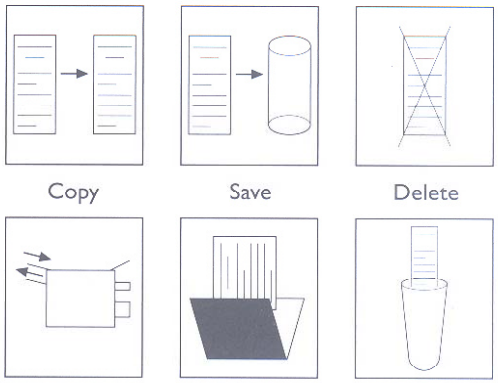
Analysis of Data

- Before you start to do any statistics...
  - Look at data
  - Save original data
- Type of data
  - Discrete - finite number of values
  - Continuous - any value
- Choice of statistical technique depends on
  - Type of data
  - Information required
    - Is there a difference? (A is faster than B)
    - How big is the difference? (A is faster by 120ms)
    - How accurate is the estimate? (... within +-9ms with 95% confidence)

Analysis - Types of Test

- Parametric
  - If normal distribution can be assumed
  - Powerful: more likely to discriminate cases
- Non-parametric
  - If normal distribution CANNOT be assumed
  - Usually based on the ranking of the data
  - Less powerful
  - Less assumption (which is more reliable)
- Contingency table
  - Classify data by discrete attributes
  - Count number of data items in each group

Example: Evaluating Icon Designs



Example: Evaluating Icon Designs

Table 9.2 Example experimental results – completion times

Participant number	Presentation order	(1) Natural (s)	(2) Abstract (s)	(3) Participant mean	(4) Natural (1)–(3)	(5) Abstract (2)–(3)
1	AN	656	702	679	–23	23
2	AN	259	339	299	–40	40
3	AN	612	658	635	–23	23
4	AN	609	645	627	–18	18
5	AN	1049	1129	1089	–40	40
6	NA	1135	1179	1157	–22	22
7	NA	542	604	573	–31	31
8	NA	495	551	523	–28	28
9	NA	905	893	899	6	–6
10	NA	715	803	759	–44	44
mean (μ)		698	750	724	–26	26
s.d. (σ)		265	259	262	14	14
			s.e.d. 117		s.e. 4.55	
Student's *t*			0.32 (n.s.)		5.78 (p<1%, two tailed)	

Studies of Groups of Users

- New problems with ...
  - Participant groups
  - Experimental task
  - Data gathering
  - Analysis
  - Field studies with groups

Participant Groups

- Larger number of subjects
- Longer time to `settle down' (some rapport to develop)
- Difficult to timetable
- So ... often only three or four groups

The Task

- Choose a task that encourages cooperation
  - That requires **consensus**
  - That requires information **distributed** among participants
- Make all channels utilized
  - If a task can be done only through video/voice channel, it is in fact testing a video conferencing system.
- Options:
  - Creative task e.g. 'Write a short report on ...'
  - Decision games e.g. Desert survival task
  - Control task e.g. Arkola bottling plant

Data Gathering

- Several video cameras + direct logging of application
- Problems
  - How to synchronize all of them?
  - How to handle/analyse huge amount of data?
- A possible alternative
  - Focus on the participants individually
  - Recreate the situation as it appeared to a participant
  - Repeat, if desired, for each participant

Analysis

- Vast variation between groups
  - Group variation > Sum of individual variances
  - Due to different relationship, different interaction styles
- Solutions
  - **Within groups experiments**
    - Beware of common problems with within-groups experiments
  - Micro-analysis (e.g., Gaps in speech)
    - Normal distribution, less dependent on social differences
  - Opt for an anecdotal and qualitative analysis
    - E.g. interesting events or breakdowns
    - Social differences are a part of study
- Controlled experiments may `waste' resources!

Field Studies of Groups

- Beware that you may end up with studying the process of group formation, instead of interaction between actual groups (in their context)
- Field studies more realistic
  - Distributed cognition -> work studied in context
  - Real action is situated action
  - Physical and social environment both crucial
- Contrast
  - Psychology – Controlled experiment
  - Sociology and anthropology – Open study and rich data

Observational Methods

- Think aloud
- Cooperative evaluation
- Protocol analysis
- Automated analysis
- Post-task walkthroughs

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

Cooperative Evaluation

- Variation on think aloud
- User sees himself a collaborator in evaluation
- **Both user and evaluator can ask each other questions throughout**
- Additional advantages
  - Less constrained and easier to use
  - User is encouraged to criticize system
  - Evaluator can clarify points of confusion → identify problem areas

Protocol Analysis

- Protocol: record of evaluation session
- Methods
  - Paper and pencil – cheap, limited by 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 unobtrusive, 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.

Automatic Protocol Analysis Tools

- Experimental Video Annotator
- Workplace project (Xerox PARC)
- DRUM

Post-Task Walkthroughs

- Data obtained by observation may lack interpretation
- Transcript played back to participant for comment
  - Immediately → fresh in mind
  - Delayed → evaluator has time to identify questions
- Useful to identify reasons for actions and alternatives considered
- Necessary in cases where think aloud is not possible

Post-Task Walkthrough

- 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
- 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

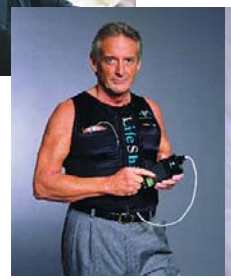
- Set of fixed questions given to users
- Advantages
  - Quick and reaches large user group
  - Can be analysed more rigorously
- Disadvantages
  - Less flexible
  - Less probing

Questionnaires

- Need careful design
  - What information is required?
  - How are answers to be analysed?
- Styles of question
  - General
  - Open-ended
  - Scalar
  - Multi-choice
  - Ranked

Physiological methods

- Eye tracking
- Physiological measurement



Seeing Machines, Facelab

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
  - Number of fixations
    - The more fixations, the less efficient the search strategy
  - Fixation duration
    - Longer fixations may indicate difficulty with a display
  - Scan path
    - Indicating areas of interest, search strategy and cognitive load
    - Moving straight to a target with a short fixation at the target is the optimal scan path...

Physiological Measurements

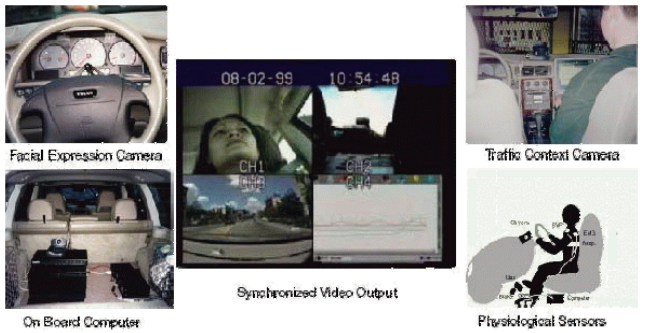
- Emotional response is linked to physical changes
- Which interaction events cause a user stress or which promote relaxation?
- Measurements include
  - Heart activity, including blood pressure, volume and pulse: Electrocardiogram (ECG)
  - Activity of sweat glands: galvanic skin response (GSR)
  - Electrical activity in muscle: electromyogram (EMG)
  - Electrical activity in brain: electroencephalogram (EEG)
- Some difficulty in interpreting these physiological responses - more research needed

Physiological Signals

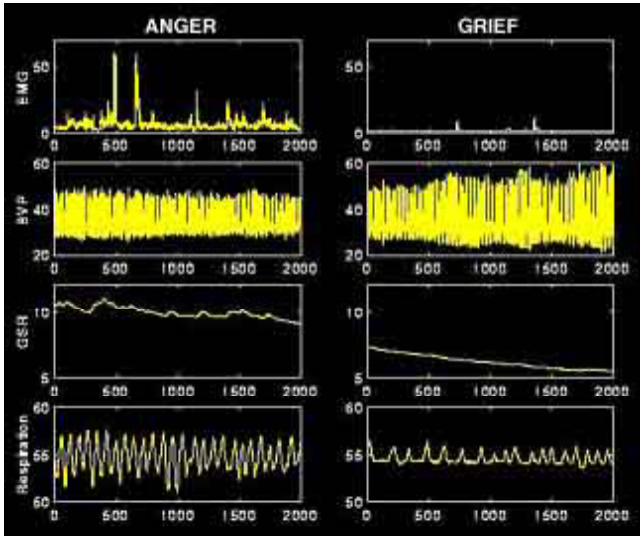
- Heart activity
  - Indicated by blood pressure, volume and pulse.
  - Respond to stress or anger
- Activity of the sweat glands
  - Indicated by skin resistance or galvanic skin response.
  - Indicate levels of arousal and mental effort
- Electrical activity in muscle
  - Measured by EMG
  - Reflect involvement in a task
- Electrical activity in the brain
  - Measured by EEG
  - Associated with decision making, attention and motivation

Detecting Driver Stress

- Four types of physiological sensors:
  - EKG, EMG, a respiration sensor, and two GSR on both the right hand and the left foot.
- Camera to capture facial expression, road condition, …
- Audio to capture ambient noise and driver's voice.



Emotion Recognition in an Actor



Choosing an Evaluation Method

- Factors to consider
  - When in process: design vs. implementation
  - Style of evaluation: laboratory vs. field
  - How objective: subjective vs. objective
  - Type of measures: qualitative vs. quantitative
  - Level of information: high level vs. low level
  - Level of interference: obtrusive vs. unobtrusive
  - Resources available: time, subjects, equipment, expertise

References

- Preece, Rogers & Sharp, Interaction Design: Beyond Human-Computer Interaction, Chapter 12,13,14,15 http://www.id-book.com