Evaluation

- What, Why, When to Evaluate

- Evaluation Types

- Evaluation Approaches and Methods

What, Why, When to Evaluate

What to evaluate?

- Early ideas for conceptual model

- Early prototypes of the new system

- Later, more complete prototypes

Why?

- Designers need to check that they understand users' requirements

When?

- Early design to clarify design ideas
- Evaluation of a working prototype
- Refining a product
- Exploration of new design concept

Evaluation Types

2 types:

- Formative evaluation: do at different stages of development to check that the product meets users' needs

- Summative evaluation: assess the quality of a finished product

A good example to illustrate them:

"When the cook tastes the soup in the kitchen, that's formative evaluation; when the guests taste the soup at the dinner table, that's summative evaluation."

Evaluation Approaches and Methods

Three main approaches

- Usability testing: Quantifying users' performance
- Field studies: under natural environments
- Analytical evaluation: no users

1. Usability testing

- Involve recording typical users' performance on carefully prepared tasks in controlled settings (laboratory-like conditions that are controlled)
- As the users perform these tasks they are watched & recorded on video & their key presses are logged
- This data is used to calculate performance times, identify errors & help explain why the users did what they did
- User satisfaction questionnaires & interviews are used to get users' opinions

Evaluation Approaches and Methods

2. Field studies

- Evaluations are performed in natural settings (e.g., test an accounting software in an accounting firm)

- The aim is to understand what users do naturally and how technology impacts them

3. Analytical evaluation

- Experts apply their knowledge of typical users, often guided by heuristics, to predict usability problems

- Use user models derived from theory

- Users need not be present

- Relatively quick & inexpensive

Evaluation Approaches and Methods

Characteristics of approaches

	Usability testing	Field studies	Analytical
Users	do task	natural	not involved
Location	controlled	natural	anywhere
When	prototype	early	prototype
Data	quantitative	qualitative	problems
Feed back	measures & errors	descriptions	problems
Type	applied	naturalistic	expert

Evaluation Approaches and Methods

Overview of evaluation methods:

- Observing users (e.g., notes, audio, video)

- Asking users (e.g., interview, questionnaire)

- Asking experts

- Testing users' performance: Measure data from human users to investigate interaction performance

- Modelling users' task performance: Use human-computer interaction models to produce performance

Evaluation Approaches and Methods

Relationship between approaches and methods:

Method	Usability testing	Field studies	Analytical
Observing	X	X	
Asking users	X	X	
Asking experts		X	X
Testing	X		
Modeling			X

Evaluation Approaches and Methods

Techniques for observing users

1. Co-operative evaluation:

- User is observed in performing specified task

- User is asked to describe what he is doing & why, what he thinks is happening, etc.

- User collaborates in evaluation and not an experimental subject

- Both user & evaluator ask each other questions throughout
  (user is encouraged to criticize the system & the evaluator can clarify points of confusion at the time they occur)

Evaluation Approaches and Methods

Advantages:

- Simplicity - require little expertise
- Can provide useful insight
- Can show how system is actually used
- User is encouraged to criticize system
- Clarification possible

Disadvantages:

- Subjective (particularly when number of users are small)
- Act of describing may affect task performance

Evaluation Approaches and Methods

Techniques for asking users:

1. Interview
2. Group Interview
3. Questionnaires

State 4 improvements for the questionnaire:

2. State your age in years ☐

3. How long have you used the Internet? ☐ <1 year
   (*check one only*) ☐ 1–3 years
   ☐ 3–5 years
   ☐ >5 years

4. Do you use the Web to:

   purchase goods ☐
   send e-mail ☐
   visit chatrooms ☐
   use bulletin boards ☐
   find information ☐
   read the news ☐

5. How useful is the Internet to you?

   _____
   _____

Evaluation Approaches and Methods

1. Exact age -> age range (15-24, 25-34, …) in Q.2

   People normally do not want to give their exact ages

2. (<1, 1-3, 3-5, >5) -> (0-1, 2-3, 4-5, >6) in Q.3

   Overlapping in the scales creates confusion, e.g., "3 years" could be in the second or third scales

3. Add "Check as many boxes as you wish" in Q.4

   No instruction about the number of choices

4. Add more space for Q.5

   Small space does not encourage giving opinions

Evaluation Approaches and Methods

Techniques for asking experts:

1. Heuristic evaluation

- Usability inspection technique proposed by Nielsen (1994)

- Experts evaluate (debug) the interface via using a checklist of usability principles called heuristics

- Heuristics are similar to design principles and guidelines

- Heuristics suggested by Nielsen:

    - Visibility of system status (Do user know about what is going on?)

    - Match between system and real world (Is the language used at the interface simple?)

    - Consistency and standards (Is the layout consistent?)

Evaluation Approaches and Methods

- **User control and freedom** (Any "back" or "undo" functions?)

- **Help users recognize, diagnose, recover from errors** (Are error messages helpful?)

- **Error prevention** (Is it easy to make errors?)

- **Recognition rather than recall** (Are objects, actions and options always visible?)

- **Flexibility and efficiency of use** (Any shortcuts for faster operation?)

- **Aesthetic and minimalist design** (Is any unnecessary and irrelevant information provided?)

- **Help and documentation** (Is help information provided that can be easily searched and easily followed?)

Evaluation Approaches and Methods

2. Review-based evaluation

- Seek experts' opinion indirectly
- Results reported in the literature are used to support or object parts of design, e.g., ACM Transactions on Computer-Human Interaction
- Need to ensure results are transferable to new design

3. Cognitive walkthroughs

- Proposed by Polson (1992)
- Focus on evaluating designs for ease of learning
- Expert (usually in cognitive psychology)`walks through' design to identify potential problems
- Questions are used to guide analysis

Evaluation Approaches and Methods

Steps involved:

(a) Identify the users and a representative task

(b) Describe the correct action sequence for that task

(c) Designer(s) and expert evaluator(s) come together to do the analysis

(d) For each action in the sequence answer the following questions:

Q1.   Will the correct action be sufficiently evident to user?

e.g., Will the user know what to do to achieve the task?

Q2.   Will user notice that the correct action is available?

e.g. Can user see the button or menu item that he should use for the next action?

Evaluation Approaches and Methods

Q3.   Will the user associate and interpret the response from the action correctly?

e.g. Will the user know from the feedback that he has made a correct or incorrect choice of action?

To summarize: Will the user

- Know what to do?

- See how to do it?

- Understand from feedback whether action was correct or not

Evaluation Approaches and Methods

(e) After performing the walkthrough, record critical information which includes

- The assumptions about what would cause problem & why. This involves explaining why users would face difficulities

- Note about side issues & design changes

- A summary of results

(f) The design is then revised to fix the problems

Example: Find a book at Hong Kong Public Library
http://libcat.hkpl.gov.hk/webpac_eng/wgbroker.exe?new+-access+top.main-page

Task : Find the book "The psychology of everyday things"

Users: Students who use the Web regularly

Evaluation Approaches and Methods

The steps to complete the task are

Step 1. Selecting correct category on Web page
Step 2. Completing the form

Step 1. Selecting correct category on Webpage

Q. Will users know what to do?
A. Yes – they know that they must find "Books".

Q. Will users see how to do it?
A. Yes – they have seen menus before and will know to select the appropriate item and click "Books"

Q. Will users understand from feedback whether the action was correct or not?
A. Yes – their action takes them to a form that they need to complete to search for the book

Evaluation Approaches and Methods

Step 2. Completing the form

Q. Will users know what to do?

A. Yes – the online form is typical in many Webs so they know they have to complete it

Q. Will users see how to do it?

A. Yes – it is clear where the information goes and there is a button "Submit" to tell the system to search for the book

A. Will users understand from feedback whether the action was correct or not?

Q. Yes – they are taken to a description of the book details and check-out records

Evaluation Approaches and Methods

Techniques for user testing:

1. Usability engineering

- Levels of usability are specified quantitatively

- Criteria are specified for judging a product's usability

- Usability specification:

    - Usability attribute/principle – principle to test

    - Measuring concept – More concrete by describing the attribute in terms of the actual product

    - Measuring method – state how the attribute will be measured

    - Now level (value in existing system) / worst case (lowest acceptance value) / planned level (target for the design) / best case (best possible measurement)

Evaluation Approaches and Methods

Example: Test the usability of an electronic diary device

Attribute: Guessability (defined by usability engineers)

Measuring concept: Ease of first use of system without training

Measuring method: Time to create first entry in diary

Now level: 30 sec. on paper-based system

Worst level: 1 min. (determined before test)

Planned level: 45 sec. (determined before test)

Best case: 30 sec. (determined before test)

If the averaged time (say, from 100 users) is 55 sec., it is usable, although the planned level is not met

Evaluation Approaches and Methods

Measurement methods can be determined from

1. Time to complete a task
2. Per cent of task completed
3. Per cent of task completed per unit time
4. Ratio of successes to failures
5. Time spent in errors
6. Per cent or number of errors
7. Per cent or number of competitors better than it
8. Number of commands used
9. Frequency of help and documentation use
10. Per cent of favourable/unfavourable user comments
11. Number of repetitions of failed commands
12. Number of runs of successes and of failures
13. Number of times interface misleads the user
14. Number of good and bad features recalled by users
15. Number of available commands not invoked
16. Number of regressive behaviours
17. Number of users preferring your system
18. Number of times users need to work around a problem
19. Number of times the user is disrupted from a work task
20. Number of times user loses control of the system
21. Number of times user expresses frustration or satisfaction

Evaluation Approaches and Methods

2. Experiment

- Aim: Answer a question or to test a hypothesis chosen by evaluator

- A number of experimental conditions are considered which differ only in the value of some controlled variables

- Quantitative measurements are collected and analysed statistically

- Factors that are important to reliable experiment:

  - Participants

    - Representative

    - Sufficient sample

Evaluation Approaches and Methods

- Variables

  - Independent variable (IV) – manipulated (controlled) by the evaluator to produce different conditions
    e.g., interface style, number of menu items

  - Dependent variable (DV) – depends on IV and are measured in the experiment
    e.g., time taken, number of errors

- Hypothesis

  - Prediction of outcome in terms of IV and DV

  - Alternative hypothesis: there is difference between conditions

  - Null hypothesis: states no difference between conditions - aim is to disprove this

Evaluation Approaches and Methods

- Allocation of participants

    - Within groups design

        - All participants perform in all conditions

        - Transfer of learning is possible but less costly & less likely to suffer from user variation

    - Between groups design

        - Each participant performs in one condition only

        - No transfer of learning but more users required & variation can bias results

    e.g., assume 2 different conditions and 10 measurements are needed, within groups design requires 10 participants while between groups design requires 20

Evaluation Approaches and Methods

Example:

- Alternative hypothesis: Users will remember the natural icons more easily than the abstract icon
- More easily: the speed at which a user can correctly select an icon
- Independent variables: 2 sets of icons
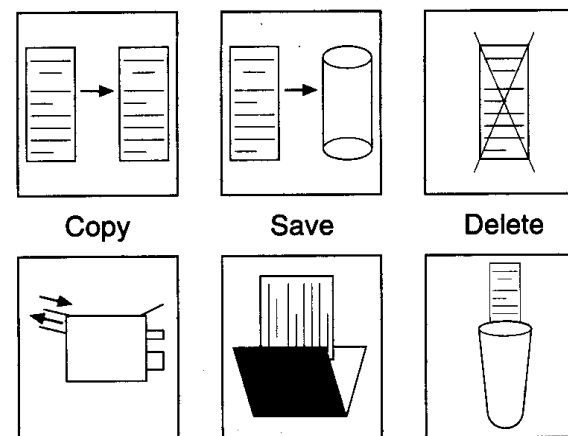- Dependent variables: time & number of mistakes

Figure 11.3 Abstract and concrete icons for file operations

Evaluation Approaches and Methods

Table 11.2		Example experimental results – completion times				
		(1)	(2)	(3)	(4)	(5)
Subject number	Presentation order	Natural (s)	Abstract (s)	Subject mean	Natural (1)–(3)	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)	

Rough conclusion: natural icons require less time in average which means that alternative hypothesis is chosen

Evaluation Approaches and Methods

Techniques for user modelling

1. GOMS

- Proposed by Card, Moran and Newell in 1983
- Most well-known predictive modelling method in HCI
- Stand for goals, operators, methods, selection
  - Goal: what the user wants to achieve (e.g., find a website about HCI design)
  - Operators: basic actions user performs to attain the goal (e.g., press keyboard key, click mouse)
  - Methods: learned procedures for accomplishing the goals (e.g., type "human computer interaction", press "search" button)
  - Selection: means of choosing between methods

Evaluation Approaches and Methods

Example: Delete text in a paragraph using WORD

Goal: delete text in a paragraph

Menu-Option Method:
Step 1.   Highlight text
Step 2.   Execute "Cut" command in "Edit" menu

Delete-Key Method:
Step 1.   Press "Delete" key to delete character one by one

Operators to use in above methods:

Click mouse
Drag cursor over text
Select menu
Move cursor
Press keyboard key

Evaluation Approaches and Methods

Selection of methods:

Rule 1: Use Menu-Option Method if large amount of text is to be deleted

Rule 2: Use Delete-Key Method if small amount of text is to be deleted

- Uses of GOMS:

  - Provide measures of performance (e.g., ↑Steps in method $\Rightarrow$ ↑short term memory requirement)

  - Provide suggestions for improving the design (e.g., old-version ATMs returned cards in the last step)

Evaluation Approaches and Methods

2. Keystroke level model (KLM)

- A very low-level GOMS model which can provide actual numerical predictions of user performance
- 7 execution phase operators:
  - Physical motor - K keystroking, actually striking keys

    B pressing a mouse button

    P pointing a target

    H homing, switching hand between mouse & keyboard

    D drawing lines using the mouse
  - Mental   -   M mentally preparing
  - System   -   R system response (can be ignored)

Evaluation Approaches and Methods

- Times are empirically determined:

$$T_{execute} = T_K + T_B + T_P + T_H + T_D + T_M + T_R$$

Table 6.1 Times for various operators in the KLM (adapted from Card, Moran and Newell [37])		
Operator	Remarks	Time (s)
K	Press key	
	good typist (90 wpm)	0.12
	poor typist (40 wpm)	0.28
	non-typist	1.20
B	Mouse button press	
	down or up	0.10
	click	0.20
P	Point with mouse	
	Fitts' law	$0.1 \log_2 (D/S + 0.5)$
	average movement	1.10
H	Home hands to and from keyboard	0.40
D	Drawing – domain dependent	–
M	Mentally prepare	1.35
R	Response from system – measure	–

Evaluation Approaches and Methods

Example: Delete the word "not" from the following sentence

I do not like using keystroke level model

Assumptions:

- User's hands at keyboard at the beginning

- User is a good typist

Which of the following methods is faster? Menu-Option Method or Delete-Key Method?

Evaluation Approaches and Methods

Menu-Option Method:

Mentally prepare	M	1.35
Switch to mouse	H	0.40
Move cursor to just before "not"	P	1.10
Hold mouse button down	B	0.10
Drag the mouse across "not" and one space	P	1.10
Release mouse button	B	0.10
Move cursor to "Edit"	P	1.10
Click mouse	2B	0.20
Move cursor to "Cut" option	P	1.10
Click mouse	2B	0.20

Hence the total execution time is 6.75 sec.

Evaluation Approaches and Methods

Delete-Key Method:

Mentally prepare	M	1.35
Switch to mouse	H	0.40
Move cursor to just before "not"	P	1.10
Click mouse	2B	0.20
Switch to keyboard	H	0.40
Press "Delete" (for "n")	K	0.12
Press "Delete" (for "o")	K	0.12
Press "Delete" (for "t")	K	0.12
Press "Delete" (for "space")	K	0.12

Hence the total execution time is 3.93 sec.

$\Rightarrow$ Delete-Key Method is faster

Evaluation Approaches and Methods

Limitations of GOMS and KLM:

- Does not allow for errors in the execution

- Difficult to apply in large-scale dialogs

- Unpredictable factors such as individual differences among users, fatigue, mental workload, etc.

3. Fitts' Law (1954)

- The law predicts that the time to point at an object using a device is a function of the distance from the target object & the object's size

- Useful for evaluating systems for which the time to locate an object is important such as handheld devices like mobile phone

4. Hick's Law