Evaluation

- Why, What, Where, When to Evaluate
- Evaluation Types
- Evaluation Methods

Why, What, Where, When to Evaluate

Iterative design and evaluation is a continuous process:

Why?

- Designers need to check that they understand users' requirements
- And if the users can use the product and they like it

What?

- Early ideas for conceptual model, e.g., whether a particular screen function is needed
- Early prototypes of the new system, e.g., whether a toy is safe for small children to play with
- Later, more complete prototypes, e.g., whether the size, color, and shape of the casing of a product are liked by people from different age groups living in different countries

Why, What, Where, When to Evaluate

Where?

- Natural environments, e.g., whether children enjoy playing with a new toy and for how long before they get bored, remote studies of online behavior
- Laboratory settings where control is provided, e.g., Web accessibility
- Living laboratories provide the setting of being in an environment, such as at home, while give the ability to control, measure, and record activities

When?

- Early design to clarify design ideas
- Evaluation of a working prototype
- Refining a product
- Finished products can be evaluated to collect information to inform new products

Why, What, Where, When to Evaluate

Bruce Tognazzini, a recognized leader in human/computer interaction design, tells you why you need to evaluate:

"Iterative design, with its repeating cycle of design and testing, is the only validated methodology in existence that will consistently produce successful results. If you don't have user-testing as an integral part of your design process you are going to throw buckets of money down the drain."

For topical discussions about design and evaluation:

www.asktog.com/

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 Types

Based on different objectives, Rubin proposes 4 types:

Exploratory evaluation:

- often informal, conducted early in the system development process
- Aim to
  - Explore interface design features of a prototype
  - Gather feedback on preliminary designs
  - Verify the assumptions about users derived during requirements determination
- The data obtained in an exploratory evaluation are mostly qualitative in nature, and are primarily based on discussions with users
- Prototypes typically used include sketches, scenarios, and interactive paper prototypes

Evaluation Types

Assessment evaluation:

- Are carried out early or midway in the development process after a conceptual model has been created that incorporates information gathered during the exploratory evaluation
- Its aims include:
  - Establish how well user tasks are supported
  - Determine what usability problems may exist
- The evaluation is conducted using user task descriptions, and measures of the level of usability of the system can be obtained
- The outcome of this evaluation may result in a refinement of the system's requirements

Evaluation Types

Comparison evaluation:

- May be performed at any stage in the development process
- When two or more design alternatives exist, either of which may appear possible, an experiment may be developed to compare them directly. Two or more prototypes are constructed, identical in all aspects except for the design issue (type of control, wording of an instruction, etc.)
- Speed and accuracy measures are collected and user preferences solicited

Evaluation Types

Validation evaluation:

- Are conducted toward the end of the development cycle or once the system is in use
- Its purpose is to ascertain that the system meets a predetermined usability objective. This may also be conducted to determine how well all of the components of a system work together. The result of this evaluation determines if the components of the interface meet the required levels of performance
- Always involve all members of the design team in the testing to ensure a common reference point for all. Involving all members also permits multiple insights into the test results from different perspectives of team members

Evaluation Types

Can also be classified as:

- Controlled settings involving users, e.g., usability testing & experiments in laboratories and living labs:
  - Its goal is to bring the lab into the home, e.g., Aware Home was embedded with a complex network of sensors and audio/video recording devices
  - Evaluate people's use of technology in their everyday lives
- Natural settings involving users, e.g., field studies to see how the product is used in the real world
- Settings not involving users, e.g., experts, who are knowledgeable about interaction design and the needs and typical behaviour of users, are employed to predict, analyse and model aspects of the interface analytics

Evaluation Types

Representative approach for each type:

- Usability testing: Quantifying users' performance
- Field studies: Under natural environments
- Analytical evaluation: No users

1. Usability testing

- Goal is to test whether the product being developed is usable by the intended user population to achieve the tasks for which it was designed, i.e., how well users perform tasks with the product
- Involve recording typical users' performance on tasks in controlled settings, e.g., usability lab or other controlled space

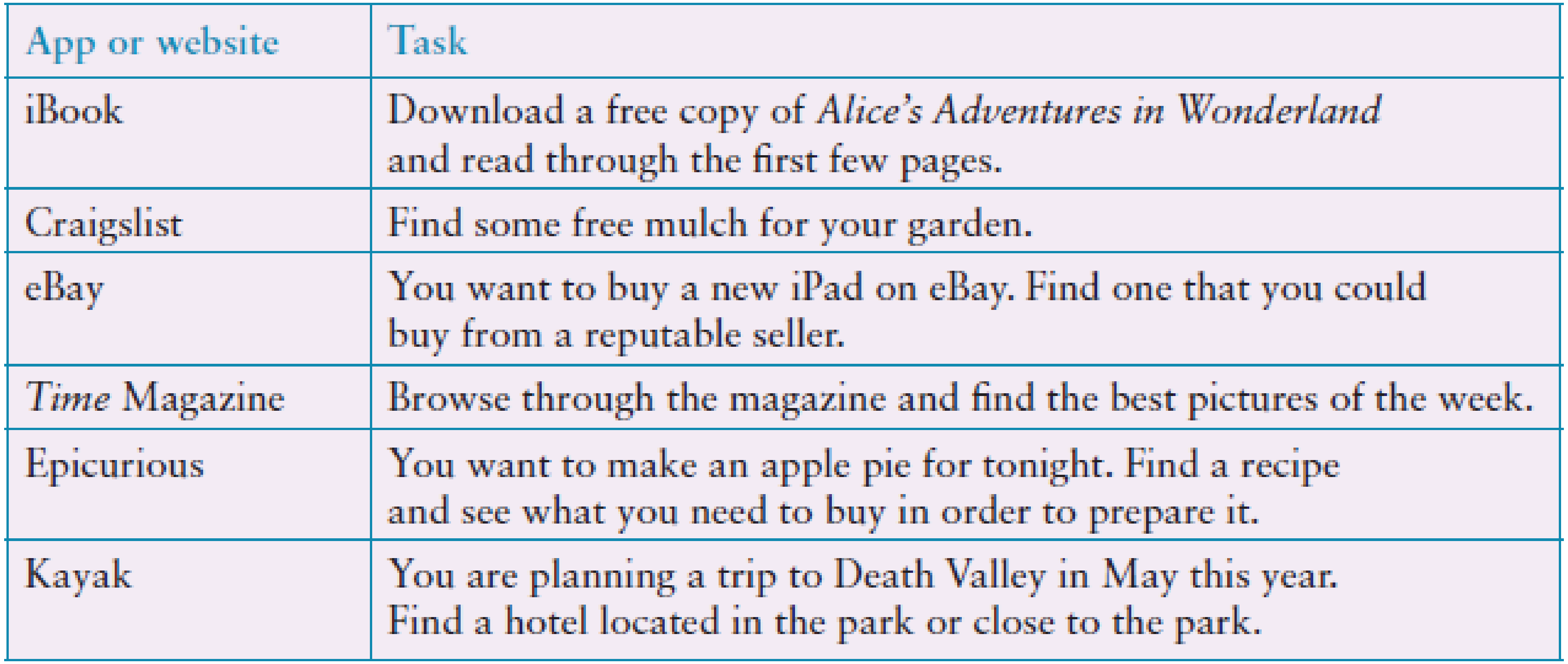
Evaluation Types

- Users are observed and timed in performing the tasks
- Data are recorded on video & key presses are logged, which are then used to calculate performance measures, and to identify & explain errors
- Performance times and numbers are two main performance measures, e.g.,
  - Time to complete a task
  - Time to complete a task after a specified time away from the product
  - Number and type of errors per task
  - Number of errors per unit of time
  - Number of times online help and manuals accessed
  - Number of users making an error
  - Number of users successfully completing a task

Evaluation Types

- Comparison of products or prototypes is common, e.g., what is the optimum number of items in a menu?
- User satisfaction questionnaires & interviews are used to get users' opinions
- Field observations may be used to provide contextual understanding
- Typical settings:
  - Emphasis on selecting representative users and developing representative tasks
  - 5-10 users
  - Tasks usually around 30 minutes
  - Test conditions are the same for every participant
  - Informed consent form explains procedures and deals with ethical issues

Evaluation Types



App or website	Task
iBook	Download a free copy of *Alice's Adventures in Wonderland* and read through the first few pages.
Craigslist	Find some free mulch for your garden.
eBay	You want to buy a new iPad on eBay. Find one that you could buy from a reputable seller.
Time Magazine	Browse through the magazine and find the best pictures of the week.
Epicurious	You want to make an apple pie for tonight. Find a recipe and see what you need to buy in order to prepare it.
Kayak	You are planning a trip to Death Valley in May this year. Find a hotel located in the park or close to the park.

Table 14.1 Examples of some of the tests used in the iPad evaluation (adapted from Budiu and Nielsen, 2010).

Source: Copyright Nielsen Norman Group, from report available at http://www.nngroup.com/reports/.

Evaluation Types

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
- "In the wild" is a term for prototypes being used freely in natural settings
- 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

Evaluation Types

In the wild example: UbiFit Garden

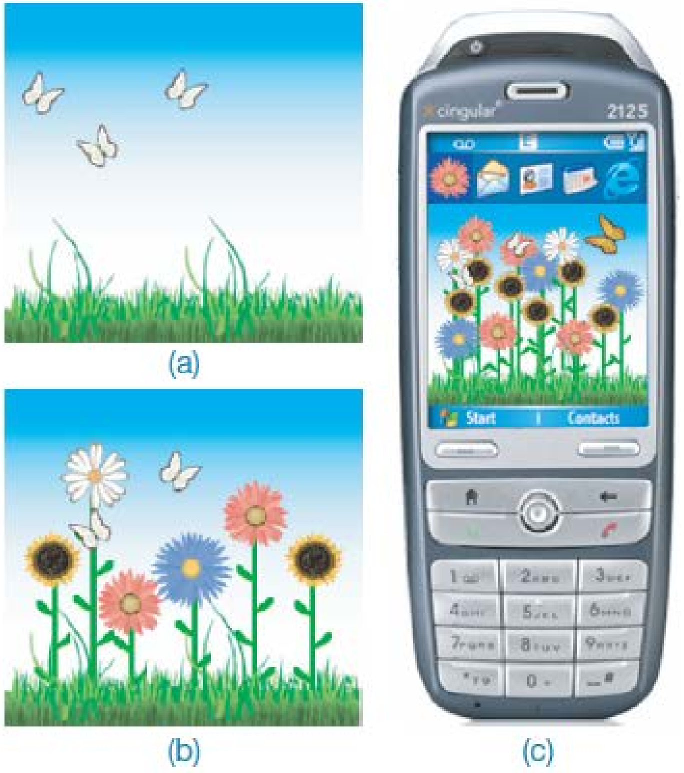


Figure 14.8 UbiFit Garden's glanceable display: (a) at the beginning of the week (small butterflies indicate recent goal attainments; the absence of flowers means no activity this week); (b) a garden with workout variety; (c) the display on a mobile phone (the large butterfly indicates this week's goal was met)

Source: From Consolvo, S., McDonald, D.W., Toscos, T. *et al* (2008) "Activity sensing in the wild: a field trial of UbiFit garden". In: *Proceedings of CHI 2008*, ACM Press, New York, p. 1799.

Evaluation Types

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

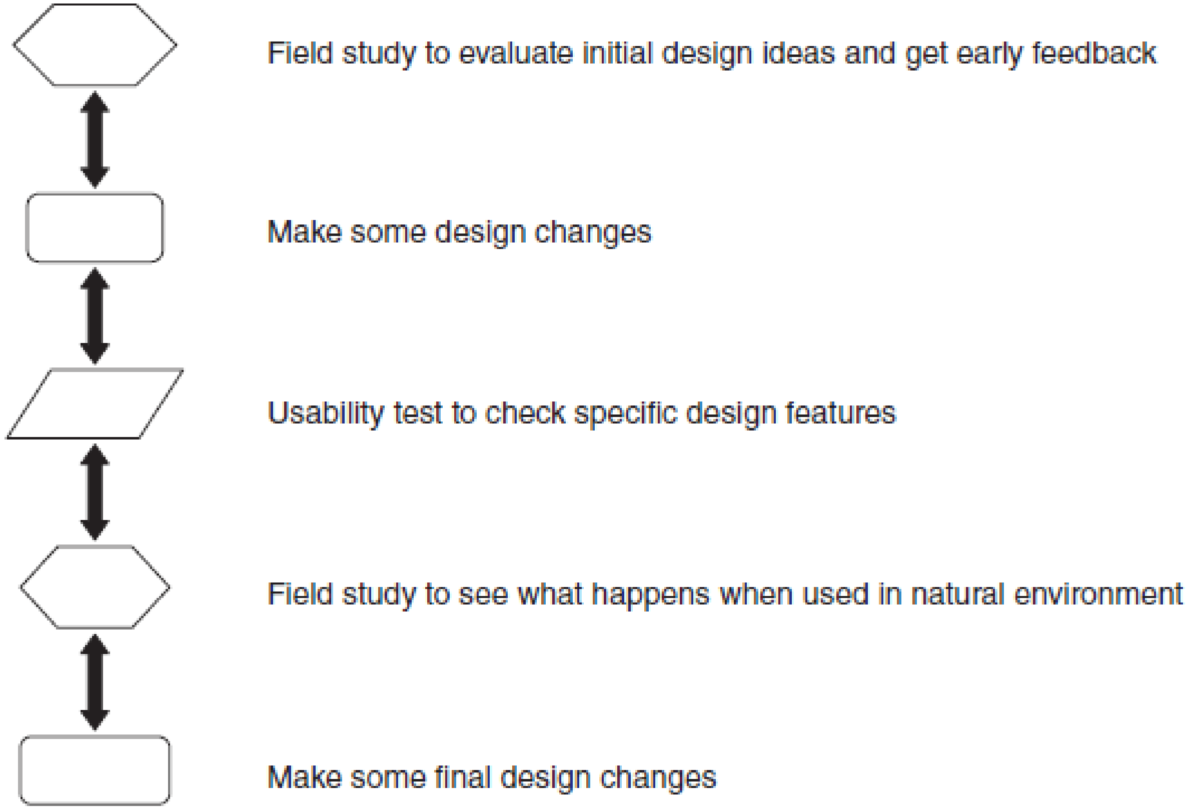
Evaluation Types

Characteristics of these three types:

	Controlled settings	**Natural settings**	**Without users**
Users	do task	natural	not involved
Location	controlled	natural	anywhere
When	prototype	Early	prototype
Data	quantitative	qualitative	problems
Feedback	measures & errors	descriptions	problems

Evaluation Types

It may be advantageous to combine approaches of different types, e.g., usability testing & field studies can compliment



Field study to evaluate initial design ideas and get early feedback
Make some design changes
Usability test to check specific design features
Field study to see what happens when used in natural environment
Make some final design changes

Figure 13.1 Example of the way laboratory-based usability testing and field studies can complement each other

Evaluation Methods

Evaluation can be classified as different 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/predict performance

Evaluation Methods

Relationship between types and methods:

Method	Controlled settings	Natural settings	Without users
Observing users	x	x	
Asking users	x	x	
Asking experts		x	x
Testing	x		
Modeling			x

Evaluation 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 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 is small)
- Act of describing may affect task performance

Evaluation Methods

Techniques for asking users:

1. Interview
2. Group Interview
3. Questionnaires

Beware of participants' rights and getting their consent is needed:

- Participants need to be told why the evaluation is being done, what they will be asked to do and their rights
- Informed consent forms provide this information
- The design of the informed consent form, the evaluation process, data analysis and data storage methods are typically approved by a high authority, e.g., Human Research Ethics Committee:

http://www.rss.hku.hk/HREC/informed-consent-form-adult.doc

Evaluation Methods

Things to consider when interpreting data:

- Reliability or consistency: does the method produce the same/similar results on separate occasions? e.g., a carefully controlled experiment has high reliability, an unstructured interview has low reliability
- Biases: Are there biases that distort the results? e.g., evaluators collecting observational data may consistently fail to notice certain types of behaviour because they do not deem them important
- Scope: How generalizable are the results?

Evaluation Methods

- Validity: does the method measure what it is intended to measure? e.g., if the goal is finding the time for completing a task, it is not appropriate to employ a method that only record the number of user errors

- Ecological validity: does the environment of the evaluation distort the results? e.g., laboratory experiments have low ecological validity because the results are unlikely to represent what happens in the real world while ethnographic studies have high ecological validity as they do not impact the participants or the study location much

Evaluation Methods

Techniques for asking experts:

1. Heuristic evaluation

- Usability inspection technique proposed by Nielsen and his colleagues in 1990s
- Heuristics are similar to design principles and guidelines
- Experts evaluate (debug) the interface via using a checklist of heuristics
- Original set of heuristics for HCI evaluation was distilled from an empirical analysis of 249 usability problems
- Revised version suggested by Nielsen in 2014:
  - Visibility of system status (The system should always keep users informed about what is going on, through appropriate feedback within reasonable time)

Evaluation Methods

- 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)
- 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)
- Consistency and standards (Users should not have to wonder whether different words, situations, or actions mean the same thing. Follow platform conventions)

Evaluation Methods

- Error prevention (Even better than good error messages is a careful design which prevents a problem from occurring in the first place. Either eliminate error-prone conditions or check for them and present users with a confirmation option before they commit to the action)
- Recognition rather than recall (Minimize the user's memory load by making 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)
- Flexibility and efficiency of use (Accelerators 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)

Evaluation Methods

- 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)
- 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)
- 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)

Evaluation Methods

Their findings suggest that 5 evaluators can typically identify around 75% of total usability problems

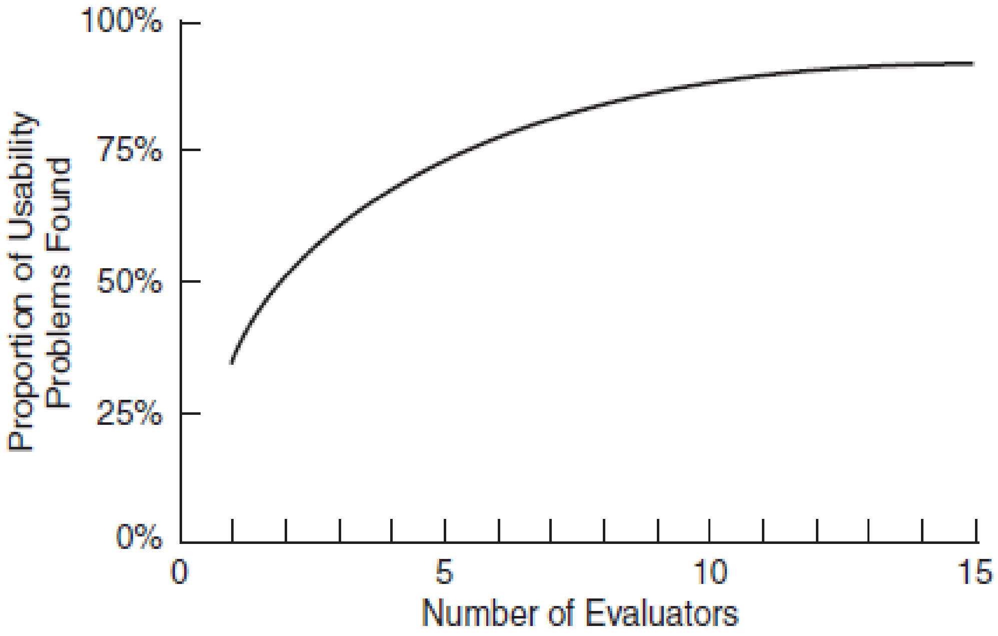


Figure 15.1 Curve showing the proportion of usability problems in an interface found by heuristic evaluation using various numbers of evaluators. The curve represents the average of six case studies of heuristic evaluation

Source: Usability Inspection Methods, J. Nielson & R.L. Mack ©1994. Reproduced with permission of John Wiley & Sons Inc.

Evaluation Methods

- Category-specific heuristics can be applied to specific class of product
- Exactly which heuristics are appropriate and how many are needed for different products depend on the goal of evaluation, but most sets of heuristics have between 5 to 10 items
- Website heuristics by Budd (2007)
  - Design for user expectations
    - Choose features that will help users achieve their goals
    - Use common web conventions
    - Make online processes work in a similar way to their offline equivalents
    - Don’t use misleading labels or buttons

Evaluation Methods

- Clarity

  –Write clear, concise copy

  –Only use technical language for a technical audience

  –Write clear and meaningful labels

  –Use meaningful icons

- Minimize unnecessary complexity and cognitive load

  –Remove unnecessary functionality, process steps and visual clutter

  –Use progressive disclosure to hide advanced features

  –Break down complicated processes into multiple steps

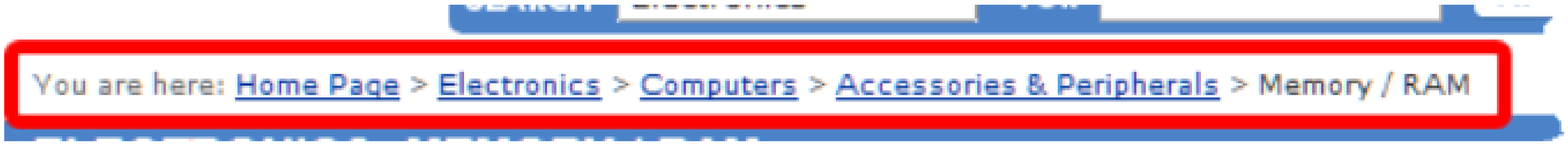
  –Prioritise using size, shape, colour, alignment and proximity

Evaluation Methods

- Efficiency and task completion
  - –Provide quick links to common features/functions
  - –Provide advanced features like the ability to delete multiple messages
  - –Pre-check common options, like opt-out of marketing emails
  - –Allow defaults to be changed, cancelled or overridden
  - –Remove unnecessary steps
- Help users notice, understand and recover from errors
  - –Visually highlight errors
  - –Provide feedback close to where the error occurred
  - –Use clear messages and avoid technical jargon

Evaluation Methods

- Provide users with context
  - –Provide a clear site name and purpose
  - –Highlight the current section in the navigation
  - –Provide a breadcrumb trail



You are here: Home Page > Electronics > Computers > Accessories & Peripherals > Memory / RAM

  - –Appropriate feedback messages
  - –Show number of steps in a process
  - –Reduce perception of latency by providing visual cues (e.g. progress indicator) or by allowing users to complete other tasks while waiting
- Promote a pleasurable and positive user experience
  - –Create a pleasurable and attractive design
  - –Provide easily attainable goals
  - –Provide rewards for usage and progression

Evaluation Methods

- Consistency and standards

  –Use common naming conventions such as “log in”?

  –Place items in standard locations like search boxes at the top right of the screen

  –Use the right interface element or form widget for job

  –Create a system that behaves in a predictable way

  –Use standard processes and web patterns

- Prevent errors

  –Disable irrelevant options

  –Accept both local and international dialling codes

  –Provide examples and contextual help

  –Check if a username is already being used before the user registers

Evaluation Methods

- 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 discuss findings and to prioritize problems, and suggest solutions

Evaluation Methods

Advantages:

- Few ethical & practical issues to consider because users are not involved
- Best experts have knowledge of application domain & users

Disadvantages:

- Can be difficult & expensive to find experts
- Important problems may get missed
- Many trivial problems (not about usability) are often identified
- Experts have biases

Evaluation 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, Lewis, Rieman, Wharton (1992)
- Involve walking through a task with the product and noting problematic usability features
- Focus on evaluating designs for ease of learning

Evaluation Methods

- Designer presents an aspect of the design & usage scenarios
- Expert, possibly in cognitive psychology, is told the assumptions about user population, context of use, task details
- One or more experts walk through a task in the design prototype with the scenario:
  - Each task involves a sequence of actions. In each task, (i) “what impact will interaction have on user?” (ii) “what cognitive processes are required?” (iii) “what learning problems may occur?” are considered
  - Individual task actions are examined and the expert tries to establish a logical reason why the user would perform each examined action

Evaluation Methods

- Actions are compared to the user’s goals and knowledge
- Discrepancies and potential problems can be identified
- Questions are used to guide analysis

Steps involved:

(a) Identify the users and a representative task

(b) Describe the correct action sequence for that task

(c) Evaluator(s) come together to do the analysis

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

Q1. Will the correct action be sufficiently evident to user? (Will the user know what to do to achieve the task?) *e.g., In cash withdrawal using ATM, can the user know the first step is to insert the bank card?*

Evaluation Methods

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?

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 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 difficulties
- Note about side issues & design changes
- A summary of results

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

Evaluation Methods

Example: Forwarding phone calls
Task: Forwarding all phone calls to Ext. 1234

The steps to complete the task are

Step 1. Pick up handset (Phone: dial tone)
Step 2. Press *2 (Phone: dial tone)
Step 3. Press 1234 (Phone: beep beep beep)
Step 4. Hang up the handset

Assume that the instruction of “FWD = *2” is available on the interface

Consider Step 2 only, the possible answers can be:

Q. Will users know what to do?
A. Yes – Although there is no FWD key, there is clear instruction of “*2” on the interface

Evaluation Methods

Q. Will users see how to do it?

A. Yes – The keys “*” and “2” are visible and thus they can press them

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

A. No – There is no feedback, i.e., no change in the tone

To fix this issue, we can change the dial tone to another tone at Step 2

Example: Find a book at Hong Kong Public Library
https://www.hkpl.gov.hk/en/index.html

Task: Find “The Psychology of Everyday Things”

Users: Students who have extensive Web surfing experience

Evaluation Methods

The steps to complete the task are

Step 1. Type the book name on the search bar
Step 2. Press the search button

Step 1. Type the book name on the search bar

Q. Will users know what to do?

A. Yes – As they have extensive Web surfing experience, they know that they need to type the book name

Q. Will users see how to do it?

A. Yes – They see “Find books, music, video and more” and thus they type there

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

A. Yes – They understand as what they type appear on the search bar

Evaluation Methods

Step 2. Press the search button

Q. Will users know what to do?

A. Yes – They know to activate the search after typing the book name

Q. Will users see how to do it?

A. Yes – There is a typical search icon for pressing. Also, moving the pointer over the icon shows “Search”

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

Q. Yes – they are taken to another Web page showing the (relevant) book details and availabilities

Evaluation Methods

Advantages:

- Walkthroughs permit a clear evaluation of the task flow early in the design process, before empirical user testing is possible
- The earlier a design flaw can be detected, the easier it is to be fixed. They can also be used to evaluate alternative design solutions
- They are more structured than other evaluation methods (such as heuristic evaluation), being less likely to suffer from subjectivity because of the emphasis on user tasks
- They are very useful for assessing exploratory learning, first-time use of a system without formal training

Evaluation Methods

Disadvantages:

- Very time-consuming and laborious to do
- Evaluators need a good understanding of the cognitive process
- Studies have found that cognitive walkthroughs appear to detect far more problems than actually exist, compared to performance-based usability testing results (Koyani et al., 2004). In these studies only about 25% of predicted problems turned to be actual problems in a usability test

Evaluation Methods

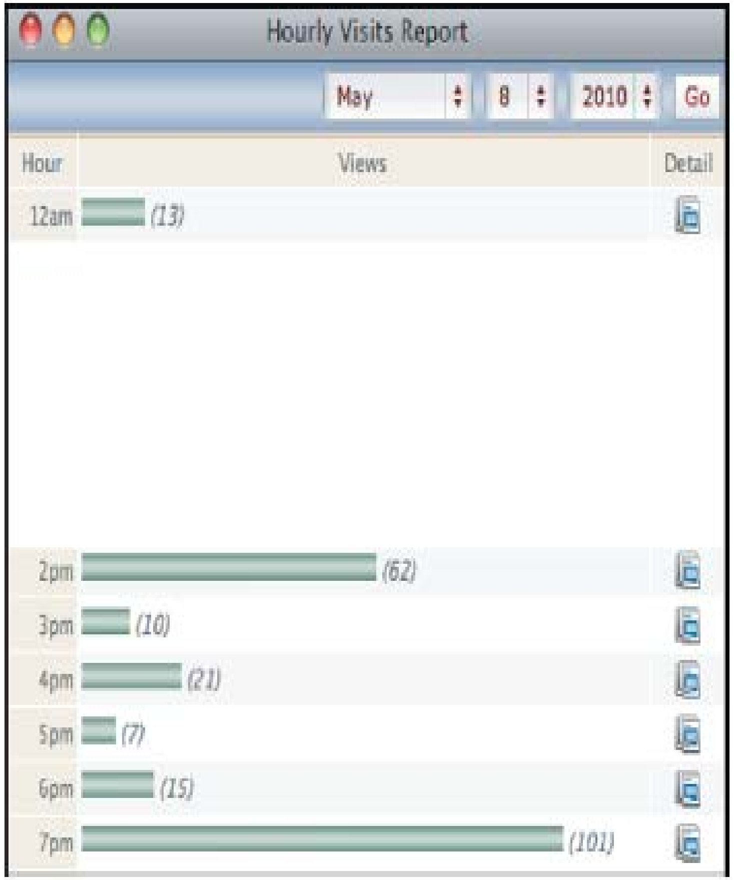
5. Analytics

- Method for evaluating user traffic through a system or part of a system, e.g., systems involve selling a product or service can find out what customers do and want, which is important for improving the product or service
- Particularly useful for evaluating the usability of Website via logging user activity, counting and analysing data in order to understand what parts of the Website are being use and when
- Recently applied to understand how learners in massive open online courses interact with the corresponding systems, e.g., what are the characteristics of learners who complete the course compared with those who do not complete it?

Evaluation Methods

- Companies may develop their own analytics tool or use the services of companies which specialize in providing analytics and the analysis necessary to understand large volumes of data, e.g., Google and KickFire
  - e.g., KickFire applied Web analytics to analyse and improve Website performance of Mountain Wines, via providing information including overview of the number of page views of its Website per day, hour-by-hour traffic with additional details for a selected day, and IP addresses of the traffic are located, which visitors are new to the site, and which are returners

Evaluation Methods



Hourly Visits Report

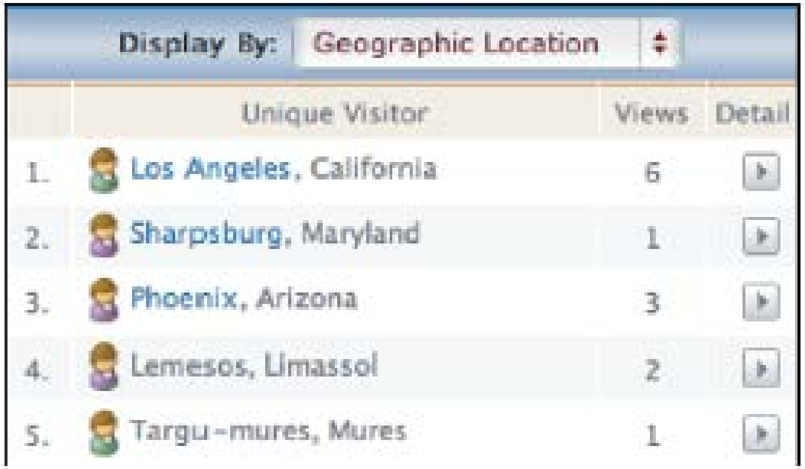
May 8 2010 Go

Hour	Views	Detail
12am	(13)	
2pm	(62)	
3pm	(10)	
4pm	(21)	
5pm	(7)	
6pm	(15)	
7pm	(101)	

Figure 15.5 Clicking on May 8 provides an hourly report from midnight until 10.00 p.m. (only midnight and 2.00 p.m.–7.00 p.m. shown)

Source: http://www.visistat.com/tracking/monthly-page-views.php

Evaluation Methods



Display By: Geographic Location

	Unique Visitor	Views	Detail
1.	Los Angeles, California	6	▸
2.	Sharpsburg, Maryland	1	▸
3.	Phoenix, Arizona	3	▸
4.	Lemesos, Limassol	2	▸
5.	Targu-mures, Mures	1	▸

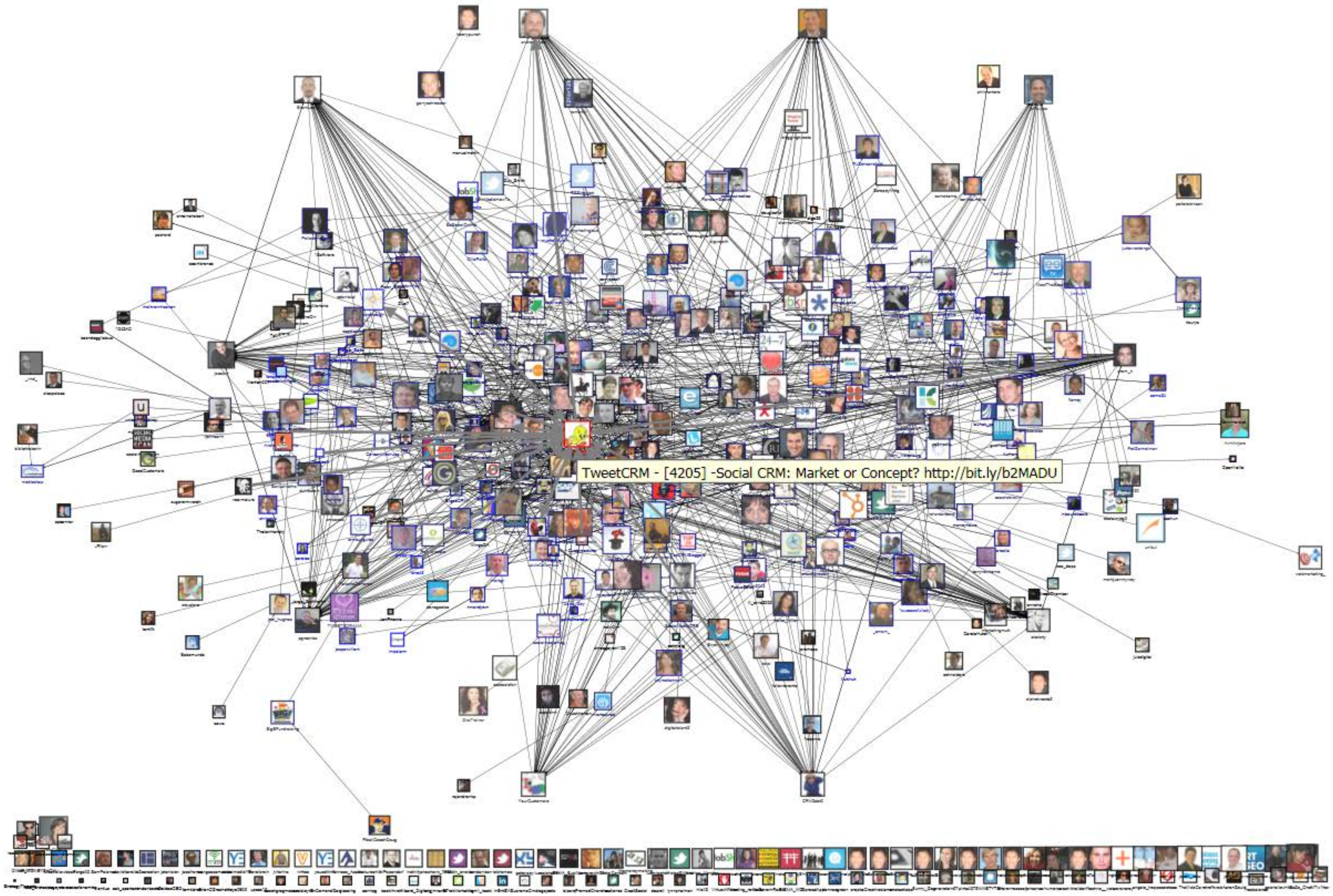
Figure 15.6 Clicking on the icon for the first hour in Figure 15.5 shows where the IP addresses the 13 visitors to the website are located

Source: http://www.visistat.com/tracking/monthly-page-views.php

Visitor IP addresses or their physical locations can be obtained

Evaluation Methods

- Other analytics include visual analytics:


TweetCRM - [4205] -Social CRM: Market or Concept? http://bit.ly/b2MADU

Evaluation 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 Methods

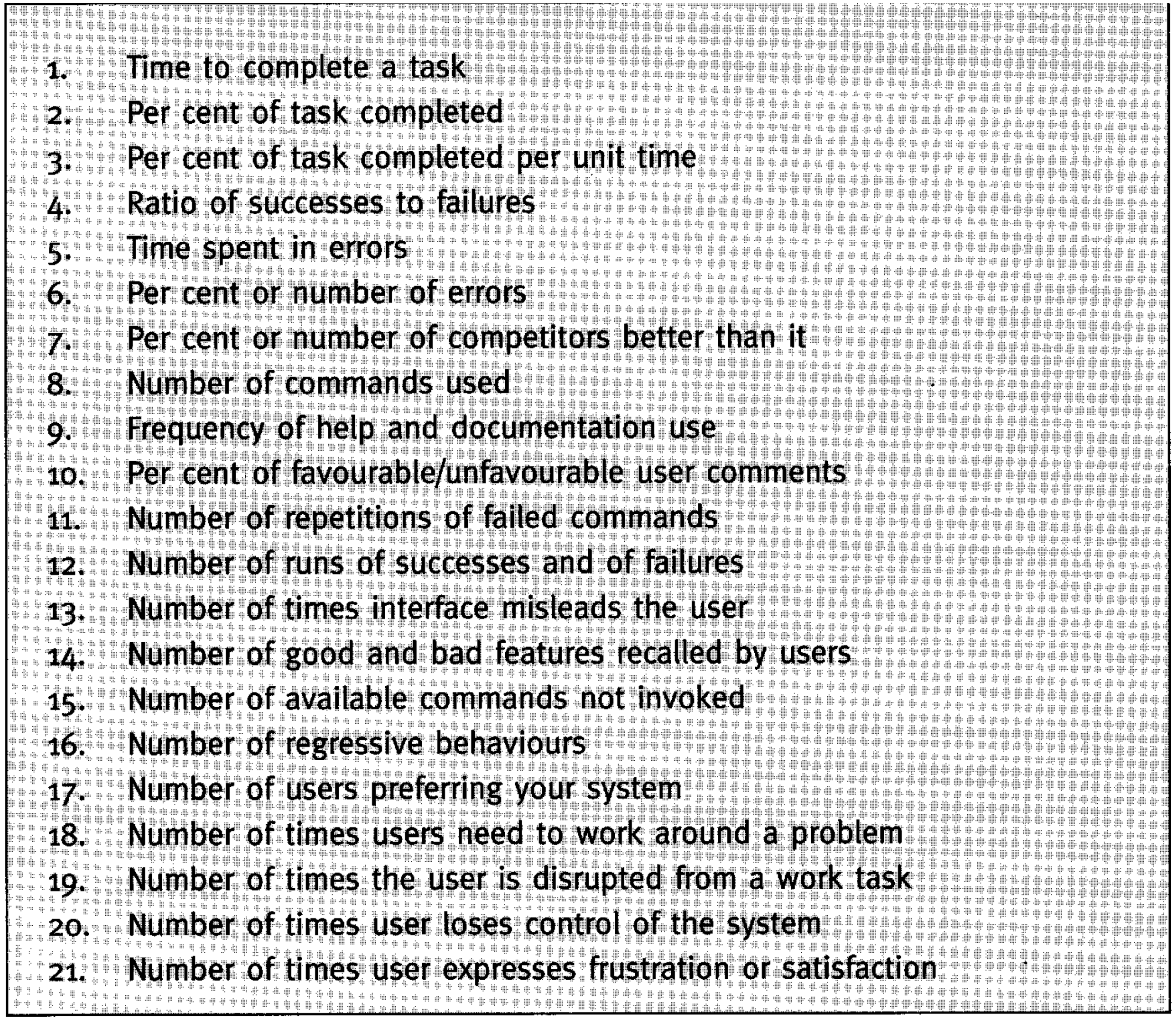
Example: Test the usability of an electronic diary device

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
- Predict the relationship between two or more variables
- Quantitative measurements are collected
- Analysed and validated statistically & replicable.
- 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 type, 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

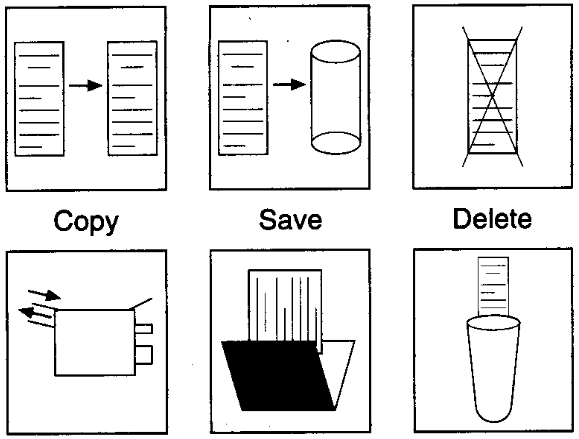


Figure 11.3 Abstract and concrete icons for file operations

Evaluation Approaches and Methods

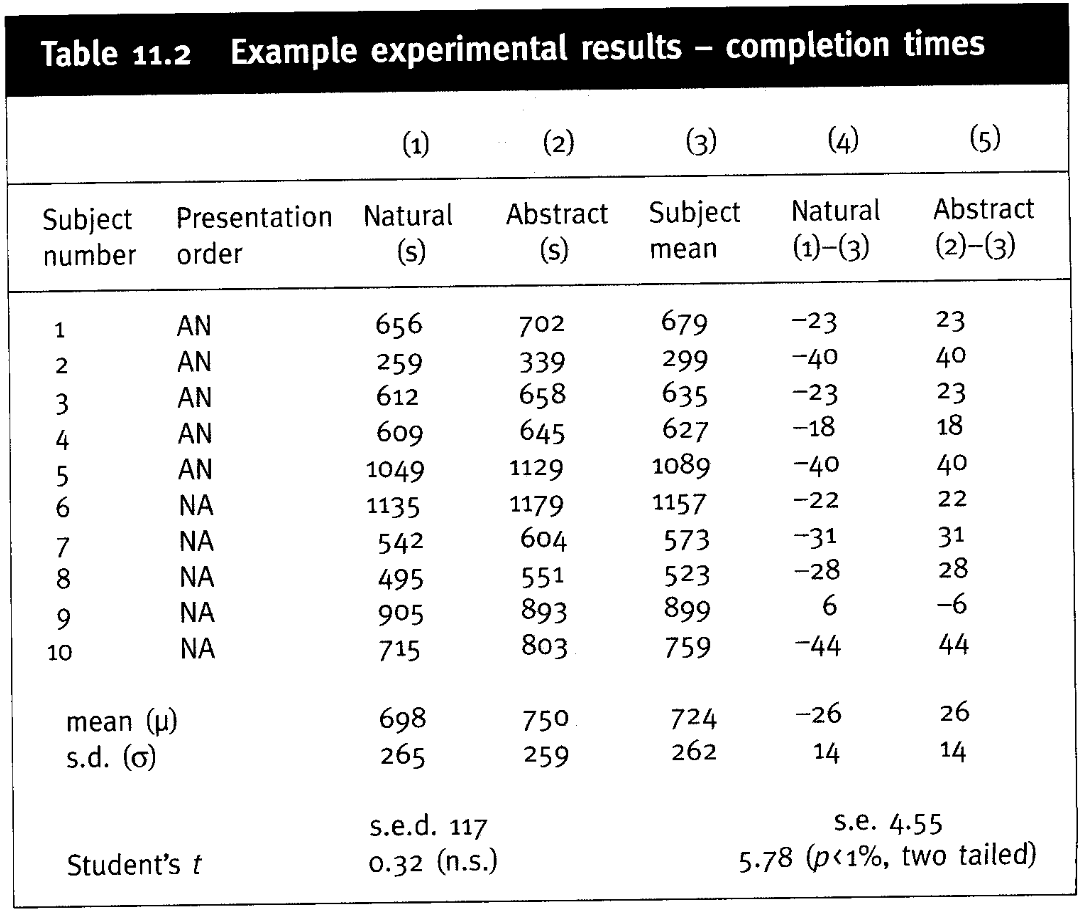


Table 11.2 Example experimental results – completion times

		(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:

- Provide a way of evaluating products or designs without directly involving users via prediction
- Less expensive than user testing
- Usefulness limited to systems with predictable tasks - e.g., telephone answering systems, mobiles, cell and smart phones, while they are difficult to apply in large-scale dialogs
- Based on expert error-free behavior, i.e., errors are not allowed in the execution
- Unpredictable factors such as individual differences among users, fatigue, mental workload, etc.

Evaluation Methods

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 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 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 ⇒ ↑short term memory requirement)
    - Provide suggestions for improving the design (e.g., old-version ATMs returned cards in the last step)

Evaluation 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 Methods

- Times are empirically determined:

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

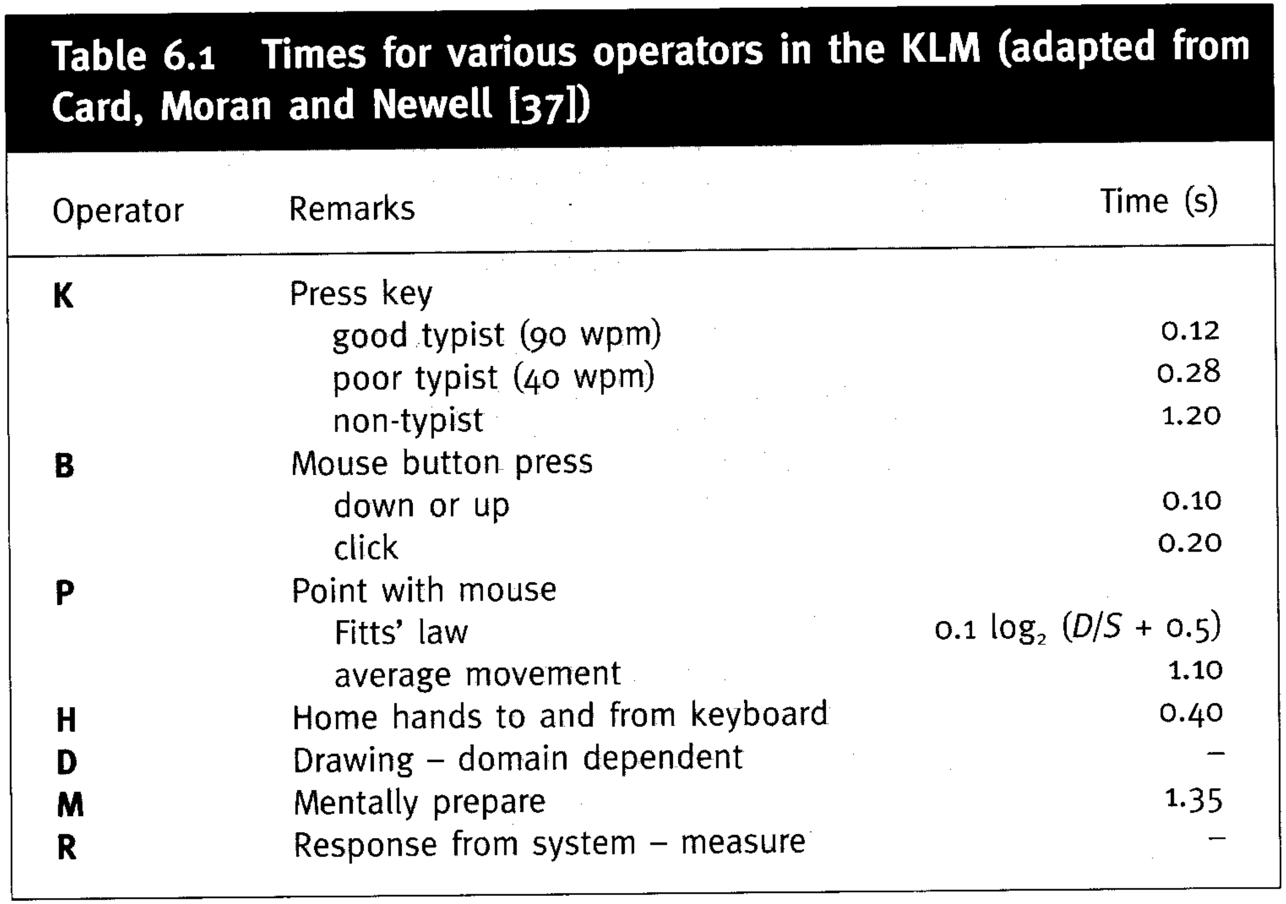


Table 6.1 Times for various operators in the KLM (adapted from Card, Moran and Newell [37])

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 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 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 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 Methods

3. Fitts' Law

- 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
- The further away and the smaller the object, the longer the time to locate it and point to it
- Useful for evaluating systems for which the time to locate an object is important such as handheld devices
- Other relevant applications include:
  - Predict expert text entry rates for several input methods on a 12-key cell phone keypad
  - Compare different ways of mapping Chinese characters to the keypad of cell phones
  - Evaluate tilt as an input method for mobile devices with built-in accelerators