chapter 9

evaluation techniques

Evaluation Techniques

- Evaluation
  - tests usability and functionality of system
  - occurs in laboratory, field and/or in collaboration with users
  - evaluates both design and implementation
  - should be considered at all stages in the design life cycle

Goals of Evaluation

- assess extent of system functionality
- assess effect of interface on user
- identify specific problems with the system

- The design of the system should enable users to perform their intended tasks more easily
- Matching the use of the system to the user's expectations of the task
- Measuring the user's performance with the system
- How easy the system is to learn, its usability.
- Identify areas of the design that overload the user in some way, perhaps by requiring an excessive amount of information to be remembered
- Etc...

- Expensive to carry out user testing at regular intervals during the design process
- Difficult to get an accurate assessment of the experience of interaction from incomplete designs and prototypes
- Therefore, methods have been proposed to evaluate interactive systems through expert analysis
- Designer, or a human factors expert, taking the design and assessing the impact that it will have upon a typical user.

- Identify any areas that are likely to cause difficulties because they violate known cognitive principles, or ignore accepted empirical results
- Can be used at any stage in the development process; from a design specification, through storyboards and prototypes, to full implementations, making them flexible evaluation approaches.
- Relatively cheap, since they do not require user involvement.
- However, they do not assess actual use of the system, only whether or not a system upholds accepted usability principles.

When to get design critique?

- **Before user testing.** Don't waste users on the small stuff. Critique can identify minor issues that can be resolved before testing, allowing users to focus on the big issues.
- **Before redesigning.** Don't throw out the baby with the bathwater. Critique can help you learn what works and what should change.
- **When you know there are problems, but you need evidence.** Perhaps you've received complaints from customers or found yourself stumbling around your own site. Critique can help you articulate problems and provide you with ammunition for redesign.
- **Before release.** Smooth off the rough edges.

	Qual Research	Quant Research
Questions answered	Why?	How many and how much?
Goals	Both formative and summative: • inform design decisions • identify usability issues and find solutions for them	Mostly summative: • evaluate the usability of an existing site • track usability over time • compare site with competitors • compute ROI
When it is used	Anytime: during redesign, or when you have a final working product	When you have a working product (either at the beginning or end of a design cycle)
Outcome	Findings based on the researcher's impressions, interpretations, and prior knowledge	Statistically meaningful results that are likely to be replicated in a different study
Methodology	• Few participants • Flexible study conditions that can be adjusted according to the team's needs • Think-aloud protocol	• Many participants • Well-defined, strictly controlled study conditions • Usually no think-aloud

Evaluating Designs: Expert Analysis

Cognitive Walkthrough

Heuristic Evaluation

Model based

Review-based evaluation

Cognitive Walkthrough

Proposed by Polson *et al.*

- Walkthroughs require a detailed review of a sequence of actions.
- Experience shows that many users prefer to learn how to use a system by exploring its functionality hands on, and not after sufficient training or examination of a user's manual.
- evaluates design on how well it supports user in learning task
- usually performed by expert in cognitive psychology
- expert 'walks through' design to identify potential problems using psychological principles
- forms used to guide analysis

Cognitive Walkthrough (ctd)

- Walkthroughs require a detailed review of a sequence of actions.
- Sequence of actions / steps that an interface will require a user to perform in order to accomplish some known task.
- The evaluators then ‘step through’ that action sequence to check it for potential usability problems and provide a ‘story’ about why that step is or is not good for a new user.

Cognitive Walkthrough (ctd)

- For each task walkthrough considers
  - what impact will interaction have on user?
  - what cognitive processes are required?
  - what learning problems may occur?
- Analysis focuses on goals and knowledge: does the design lead the user to generate the correct goals?

What is required?

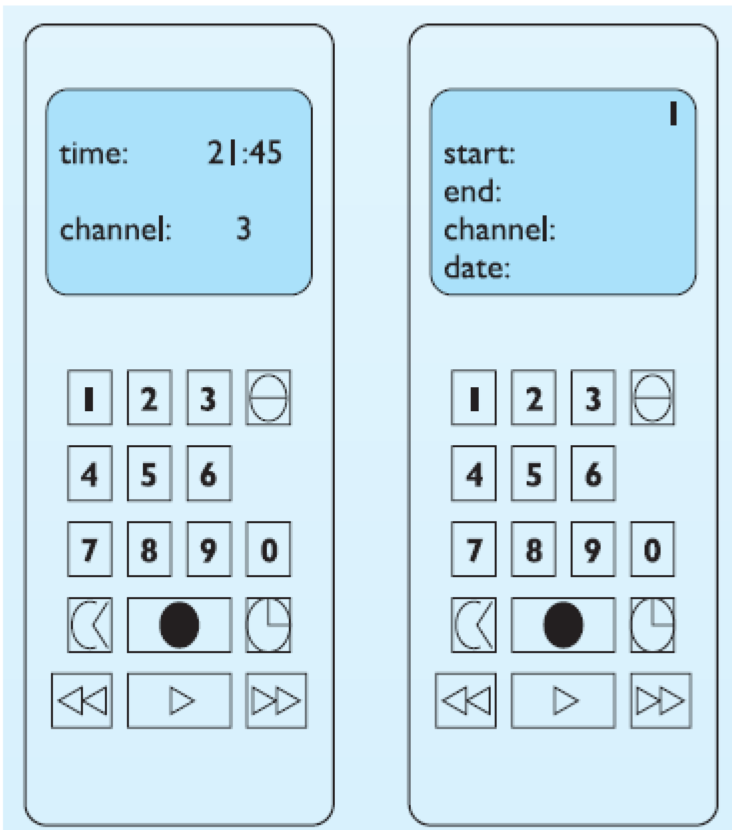
1. A specification or prototype of the system. It doesn't have to be complete, but it should be fairly detailed. Details such as the location and wording for a menu can make a big difference.
2. A description of the task the user is to perform on the system. This should be a representative task that most users will want to do.
3. A complete, written list of the actions needed to complete the task with the proposed system.
4. An indication of who the users are and what kind of experience and knowledge the evaluators can assume about them.

Evaluate

1. Is the effect of the action the same as the user’s goal at that point?
2. Will users **see** that the action is available? (Is it visible?)
3. Once users have found the correct action, will they know it is the one they need? (Recognize)
4. After the action is taken, will users understand the feedback they get? (Both perceive the feedback as well as interpret)

E.g. Timed recording on VCR

- We will assume that the user is familiar with VCRs but not with this particular design.



Program to time-record a program starting at 18.00 and finishing at 19.15 on channel 4 on 24 February 2005.

- Action sequence in terms of the user's action (UA) and the system's display or response (SD).

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

Ask Questions

UA 1: Press the 'timed record' button

Question 1: Is the effect of the action the same as the user's goal at that point?

The timed record button initiates timer programming. It is reasonable to assume that a user familiar with VCRs would be trying to do this as his first goal.

Question 2: Will users see that the action is available?

The 'timed record' button is visible on the remote control.

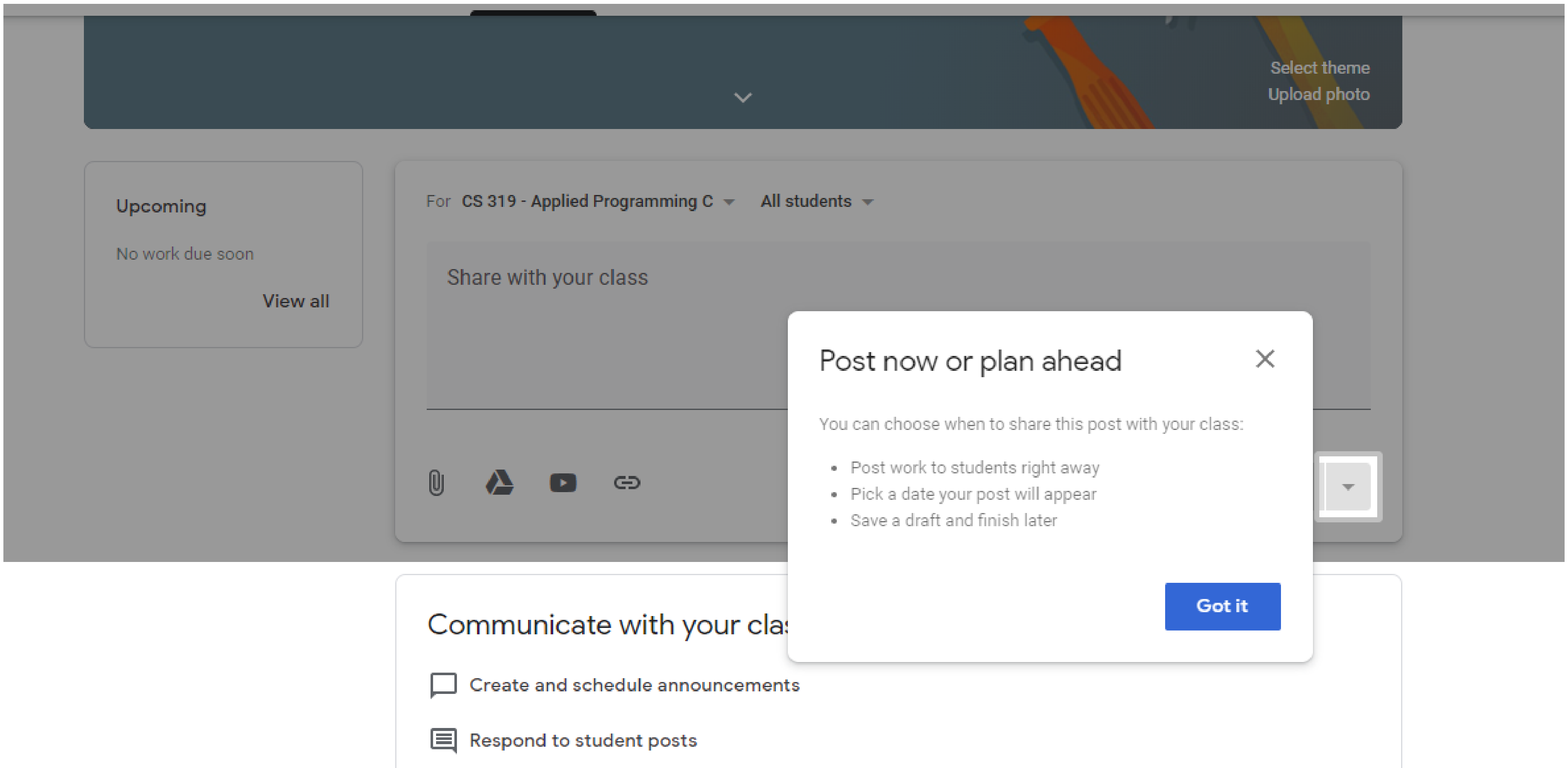
Question 3: Once users have found the correct action, will they know it is the one they need?

It is not clear which button is the 'timed record' button. The icon of a clock (fourth button down on the right) is a possible candidate but this could be interpreted as a button to change the time. Other possible candidates might be the fourth button down on the left or the filled circle (associated with record). In fact, the icon of the clock is the correct choice but it is quite possible that the user would fail at this point. This identifies a potential usability problem.

Question 4: After the action is taken, will users understand the feedback they get?

Once the action is taken the display changes to the timed record mode and shows familiar headings (start, end, channel, date). It is reasonable to assume that the user would recognize these as indicating successful completion of the first action.

- Potential usability problem relating to the icon used on the ‘timed record’ button. We would now have to establish whether our target user group could correctly distinguish this icon from others on the remote.
- For each User Action, the 4 questions from the last slide need to be asked.


Select theme
Upload photo
Upcoming
No work due soon
View all
For CS 319 - Applied Programming C
All students
Share with your class
Post now or plan ahead
You can choose when to share this post with your class:
Post work to students right away
Pick a date your post will appear
Save a draft and finish later
Got it
Communicate with your clas
Create and schedule announcements
Respond to student posts

Heuristic Evaluation

- Proposed by Nielsen and Molich.
- Is a method for structuring the critique of a system using a set of relatively simple and general heuristics.
- A heuristic is a guideline or general principle or rule of thumb that can guide a design decision or be used to critique a decision that has already been made.

- 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.

- Heuristic evaluation can be performed on a **design specification** so it is useful for evaluating early design. But it can also be used on prototypes, storyboards and fully functioning systems.
- It is therefore a flexible, relatively cheap approach. Hence it is often considered a *discount usability technique*.

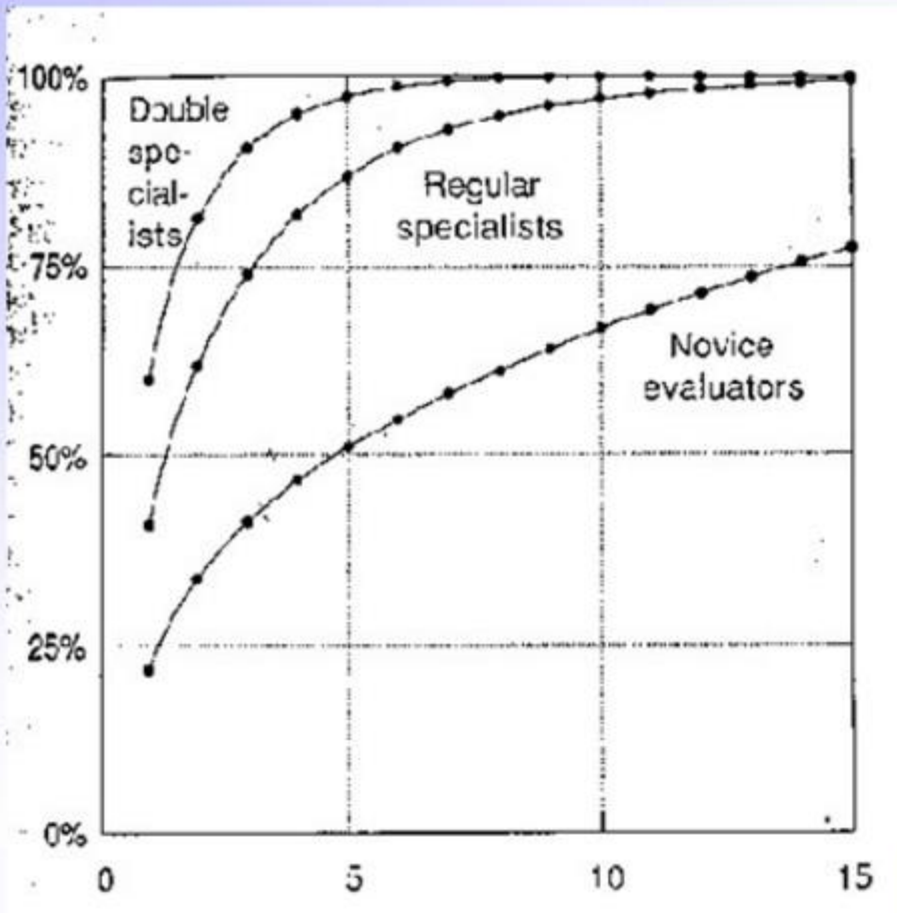
Heuristic Evaluation

- Developed by Jakob Nielsen
- Helps find usability problems in a design
- Small set (3-5) of evaluators examine UI
  - independently check for compliance with usability principles ("heuristics")
  - different evaluators will find different problems
  - evaluators only communicate afterwards
    - findings are then aggregated
- Can perform on working UI or sketches

- Independent first, gather afterwards

- Btw, is this graph okay w.r.t. our criteria?

by different numbers of evaluators (Nielsen)



- 3 ~ 5 evaluators find ~ 75% of the problems.

- Step through design several times
  - Examine details, flow, and architecture
  - Consult list of usability principles
  - ..and anything else that comes to mind
- Which principles?
  - Nielsen's "heuristics"
  - Category-specific heuristics from e.g., design goals, competitive analysis, existing designs
- Use violations to redesign/fix problems

Phases of Heuristic Evaluation

1. **Pre-evaluation training:** give evaluators needed domain knowledge and information on the scenario
2. **Evaluation:** individuals evaluate and then aggregate results
3. **Severity rating:** determine how severe each problem is (priority). Can do first individually and then as a group
4. **Debriefing:** review with design team

Severity ratings

0 = I don't agree that this is a usability problem at all
1 = Cosmetic problem only: need not be fixed unless extra time is available on project
2 = Minor usability problem: fixing this should be given low priority
3 = Major usability problem: important to fix, so should be given high priority
4 = Usability catastrophe: imperative to fix this before product can be released (Nielsen)

Nielson's ten heuristics

1. **Visibility of system status** Always keep users informed about what is going on, through appropriate feedback within reasonable time. For example, if a system operation will take some time, give an indication of how long and how much is complete.
2. **Match between system and the real world** The system should speak the user's language, with words, phrases and concepts familiar to the user, rather than system-oriented terms. Follow real-world conventions, making information appear in natural and logical order.
3. **User control and freedom** Users often choose system functions by mistake and need a clearly marked 'emergency exit' to leave the unwanted state without having to go through an extended dialog. Support undo and redo.
4. **Consistency and standards** Users should not have to wonder whether words, situations or actions mean the same thing in different contexts. Follow platform conventions and accepted standards.
5. **Error prevention** Make it difficult to make errors. Even better than good error messages is a careful design that prevents a problem 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 dialog to another. Instructions for use of the system should be visible or easily retrievable whenever appropriate.
7. **Flexibility and efficiency of use** Allow users to tailor frequent actions. Accelerators – unseen by the novice user – may often speed up the interaction for the expert user to such an extent that the system can cater to both inexperienced and experienced users.
8. **Aesthetic and minimalist design** Dialogs should not contain information that is irrelevant or rarely needed. Every extra unit of information in a dialog 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** Few systems can be used with no instructions so it may be necessary to provide help and documentation. Any such information should be easy to search, focussed on the user's task, list concrete steps to be carried out, and not be too large.

Heuristics vs. User Testing

- Heuristic Evaluation often faster
  - 1-2 hours each evaluator
- HE results come pre-interpreted
- User testing is more accurate (by def.)
  - takes into account actual users and tasks
  - HE may miss problems & find "false positives"
- Valuable to alternate methods
  - find different problems
  - don't waste participants

Other methods

- Model-based evaluation

  -Cognitive models used to filter design options

  e.g. GOMS (goals, operators, methods and selection) prediction of user performance with a particular interface.

  -Lower-level modeling techniques such as the keystroke-level model provide predictions of the time users will take to perform low-level physical tasks.

  Dialog models can also be used to evaluate dialog sequences for problems, such as unreachable states etc

- **Design rationale** can also provide useful evaluation information

Review-based evaluation

- It is expensive to repeat experiments continually.
- Results from the literature used to support or refute parts of design.
- Experimental psychology and human–computer interaction between them possess a wealth of experimental results and empirical evidence.
- Examples of such issues are the usability of different menu types, the recall of command names, and the choice of icons, etc.
- Care needed to ensure results are transferable to new design.

Evaluating through user Participation

- Expert evaluation are useful as these techniques are for filtering and refining the design, but they are not a replacement for actual usability testing with the people for whom the system is intended: the users.
- User participation in evaluation tends to occur in the later stages of development when there is at least a working prototype of the system in place (Can use prototype based on Wizard of Oz technique).
- Different user evaluation techniques include empirical or experimental methods, observational methods, query techniques, and methods that use physiological monitoring, such as eye tracking and measures of heart rate and skin conductance.

Laboratory studies

- Users are taken out of their normal work environment to take part in controlled tests.
- May contain sophisticated audio/visual recording and analysis facilities, two-way mirrors, instrumented computers.
- Advantages:
  - specialist equipment available
  - uninterrupted environment
- Disadvantages:
  - lack of context
  - difficult to observe several users cooperating
- Appropriate
  - If system location is dangerous
  - We may deliberately want to manipulate the context in order to uncover problems or observe less used procedures.
  - We may want to compare alternative designs within a controlled context.

Field Studies

- The second type of evaluation takes the designer or evaluator out into the user's work environment in order to observe the system in action.
- On balance, field observation is to be preferred to laboratory studies as it allows us to study the interaction as it occurs in actual use.
- However, we should remember that even in field observations the participants are likely to be influenced by the presence of the analyst and/or recording equipment, so we always operate at a slight remove from the natural situation, a sort of Heisenberg uncertainty principle.

- Advantages:
  - natural environment
  - interruptions
  - context retained (though observation may alter it)
  - longitudinal studies possible
- Disadvantages:
  - distractions
  - noise
- Appropriate
  - where context is crucial for longitudinal studies

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 <u>which differ only in the value of some controlled variable</u>.
- changes in behavioural measure are attributed to different conditions

Important Experimental factors

- Subjects
  - who – representative, sufficient sample
- Variables
  - things to modify and measure (e.g. interface style, level of help, layout, number of menu items and icon design.)
- Hypothesis
  - what you'd like to show
- Experimental design
  - how you are going to do it

Subjects/ Participants

- Choice of participants is vital to the success of any experiment.
- If participants are not actual users, they should be chosen to be of a similar characteristics (similar age and level of education as the intended user group).
- ~3-5?

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, user preferences

Hypothesis

- A hypothesis is a prediction of the outcome of an experiment
  - 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 hyp. = “no change with font size”

Experimental design

- The first phase in experimental design then is to choose the hypothesis: to decide exactly what it is you are trying to demonstrate
- clarify the independent and dependent variables
- consider your participants: how many are available and are they representative of the user group?

- There are at least two conditions: the *experimental condition* (in which the variable has been manipulated) and the *control*, which is identical to the experimental condition except for this manipulation.
- This control serves to ensure that it is the manipulation that is responsible for any differences that are measured.
- There may, of course, be more than two groups, depending on the number of independent variables and the number of levels that each variable can take.

Experimental design

- between groups design
  - each subject performs under only one condition
  - no transfer of learning
  - more users required
  - variation can bias results.
- within groups design
  - each subject performs experiment under each condition.
  - transfer of learning possible
  - less costly and less likely to suffer from user variation.

Analysis of data

- Before you start to do any statistics:
  - look at data
  - 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 - types of test

- parametric
  - assume normal distribution
  - robust
  - powerful
- non-parametric
  - do not assume normal distribution
  - less powerful
  - more reliable
- contingency table
  - classify data by discrete attributes
  - count number of data items in each group

Analysis of data (cont.)

- What information is required?

Is there a difference? For example, is one system better than another? Techniques that address this are called *hypothesis testing*. The answers to this question are not simply yes/no, but of the form: 'we are 99% certain that selection from menus of five items is faster than that from menus of seven items'.

How big is the difference? For example, 'selection from five items is 260 ms faster than from seven items'. This is called *point estimation*, often obtained by averages.

How accurate is the estimate? For example, 'selection is faster by 260 ± 30 ms'. Statistical answers to this are in the form of either measures of variation such as the *standard deviation* of the estimate, or *confidence intervals*. Again, the answers one obtains are probabilistic: 'we are 95% certain that the difference in response time is between 230 and 290 ms'.

- Parametric and non-parametric tests mainly address first of these

Statistical Techniques

Table 9.1 Choosing a statistical technique

Independent variable	Dependent variable	
Parametric		
Two valued	Normal	Student's *t* test on difference of means
Discrete	Normal	ANOVA (ANalysis Of VAriance)
Continuous	Normal	Linear (or non-linear) regression factor analysis
Non-parametric		
Two valued	Continuous	Wilcoxon (or Mann–Whitney) rank-sum test
Discrete	Continuous	Rank-sum versions of ANOVA
Continuous	Continuous	Spearman's rank correlation
Contingency tests		
Two valued	Discrete	No special test, see next entry
Discrete	Discrete	Contingency table and chi-squared test
Continuous	Discrete	(Rare) Group independent variable and then as above

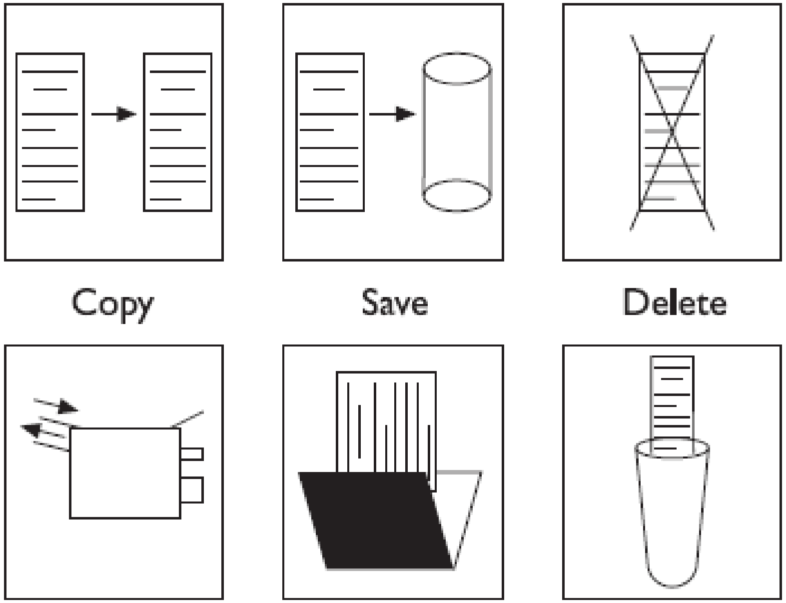


Figure 9.3 Abstract and concrete icons for file operations

- Hypothesis?
- IV, DV?
- Experimental setup?
- The user is presented with a task (say 'delete a document') and is required to select the appropriate icon.
- In order to avoid learning effects from icon position, the placing of icons in the block can be randomly varied on each presentation.
- And one other thing….

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)	

- Possibly, looking at the errors we may find that the natural icons have *more* errors – it could well be that they are more rapidly, but less accurately, remembered. It is always worth keeping in mind the difference between the intended purpose of the experiment (to see which is better remembered) and the actual measurements (speed and accuracy).

Experimental studies on groups

More difficult than single-user experiments

Problems with:

- subject groups
- choice of task
- data gathering
- analysis

Subject groups

larger number of subjects

$\Rightarrow$ more expensive

longer time to `settle down'

… even more variation!

difficult to timetable

so … often only three or four groups

The task

must encourage cooperation

perhaps involve multiple channels, e.g. talk and edit a document

- For example, in the case of shared application with video, it should not be possible (or at least not easy) to perform the task without using the application, otherwise we are simply investigating video conferencing.

options:

- creative task e.g. '*write a short report on …*'
- decision games e.g. desert survival task. List important items
- control task e.g. ARKola bottling plant. Control different parts of a system together.

- In ARKola, a soft drinks factory was modeled, with two users attempting to optimize the factory's output, working remotely from each other and using an audio/video link. Input machines supplied raw materials while output machines capped the bottles and shipped them out. Each machine had an on/off switch and a rate control, with a sound that indicated its status; for example, the bottle dispenser made the sound of clinking glass, with a rhythm that indicated its operating speed. Splashing sounds indicated spilled liquids, while breaking glass showed that bottles were being lost. The users monitored the status of the plant by listening to the auditory clues, and were able to help each other more effectively, since they found it easier to monitor their own machines without having to spend time looking at them, and could hear when something had gone wrong with their partner's part of the system.

Data gathering

several video cameras
+ direct logging of application

problems:

- synchronisation
- sheer volume!

one solution:

- record from each perspective

Analysis

N.B. vast variation between groups

solutions:

- within groups experiments
- micro-analysis (e.g., gaps in speech)
- anecdotal and qualitative analysis

look at interactions between group and media

controlled experiments may `waste' resources!

Field studies

Experiments dominated by group formation

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 (point of view of user selected)
  - act of describing may alter task performance

- https://www.coursera.org/lecture/ui-testing/usability-lab-example-1-user-1-c050l

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 criticize system
  - clarification possible

Protocol analysis

- paper and pencil – cheap, limited to 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.
- Some automatic support tools available

automated analysis – EVA (*Experimental Video Annotator)*

- Workplace project
- Post task walkthrough
  - user reacts on 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

post-task walkthroughs

- 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

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 analyzed more rigorously
- Disadvantages
  - less flexible
  - less probing

Questionnaires (ctd)

- Need careful design
  - what information is required?
  - how are answers to be analyzed?
- Styles of question
  - general
  - open-ended
  - Scalar

    It is easy to recover from mistakes.
    Disagree 1 2 3 4 5 Agree

  - multi-choice

    How do you most often get help with the system (tick one)?
    Online manual ❑
    Contextual help system ❑
    Command prompt ❑
    Ask a colleague ❑

  - ranked

    Please rank the usefulness of these methods of issuing a command (1 most useful, 2 next, 0 if not used).
    Menu selection ❑
    Command line ❑
    Control key accelerator ❑

Physiological methods

Eye tracking
Physiological measurement

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
  - fixations: eye maintains stable position. Number and duration indicate level of difficulty with display
  - saccades: rapid eye movement from one point of interest to another
  - scan paths: moving straight to a target with a short fixation at the target is optimal

physiological measurements

- emotional response linked to physical changes
- these may help determine a user's reaction to an interface
- measurements include:
  - heart activity, including blood pressure, volume and pulse.
  - 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

Choosing an Evaluation Method

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

Table 9.4 Classification of analytic evaluation techniques

	Cognitive walkthrough	Heuristic evaluation	Review based	Model based
Stage	Throughout	Throughout	Design	Design
Style	Laboratory	Laboratory	Laboratory	Laboratory
Objective?	No	No	As source	No
Measure	Qualitative	Qualitative	As source	Qualitative
Information	Low level	High level	As source	Low level
Immediacy	N/A	N/A	As source	N/A
Intrusive?	No	No	No	No
Time	Medium	Low	Low–medium	Medium
Equipment	Low	Low	Low	Low
Expertise	High	Medium	Low	High

Table 9.5 Classification of experimental and query evaluation techniques

	Experiment	Interviews	Questionnaire
Stage	Throughout	Throughout	Throughout
Style	Laboratory	Lab/field	Lab/field
Objective?	Yes	No	No
Measure	Quantitative	Qualitative/ quantitative	Qualitative/ quantitative
Information	Low/high level	High level	High level
Immediacy	Yes	No	No
Intrusive?	Yes	No	No
Time	High	Low	Low
Equipment	Medium	Low	Low
Expertise	Medium	Low	Low

Table 9.6 Classification of observational evaluation techniques

	Think aloud[1]	Protocol analysis[2]	Post-task walkthrough
Stage	Implementation	Implementation	Implementation
Style	Lab/field	Lab/field	Lab/field
Objective?	No	No	No
Measure	Qualitative	Qualitative	Qualitative
Information	High/low level	High/low level	High/low level
Immediacy	Yes	Yes	No
Intrusive?	Yes	Yes[3]	No
Time	High	High	Medium
Equipment	Low	High	Low
Expertise	Medium	High	Medium

1 Assuming a simple paper and pencil record
2 Including video, audio and system recording
3 Except system logs

Table 9.7 Classification of monitoring evaluation techniques

	Eye tracking	Physiological measurement
Stage	Implementation	Implementation
Style	Lab	Lab
Objective?	Yes	Yes
Measure	Quantitative	Quantitative
Information	Low level	Low level
Immediacy	Yes	Yes
Intrusive?	No[1]	Yes
Time	Medium/high	Medium/high
Equipment	High	High
Expertise	High	High

1 If the equipment is not head mounted