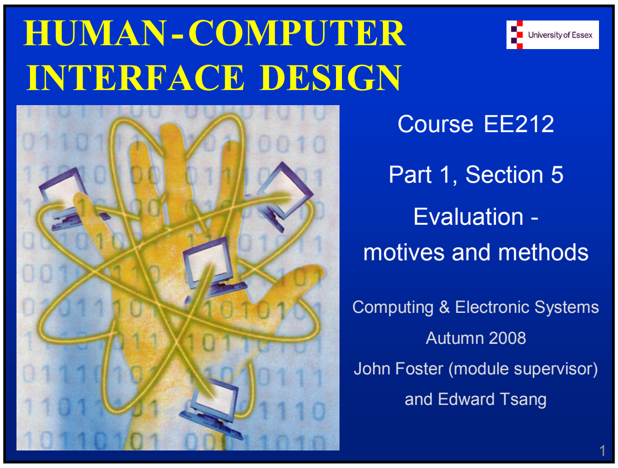


HUMAN-COMPUTER INTERFACE DESIGN

Course EE212

Part 1, Section 5

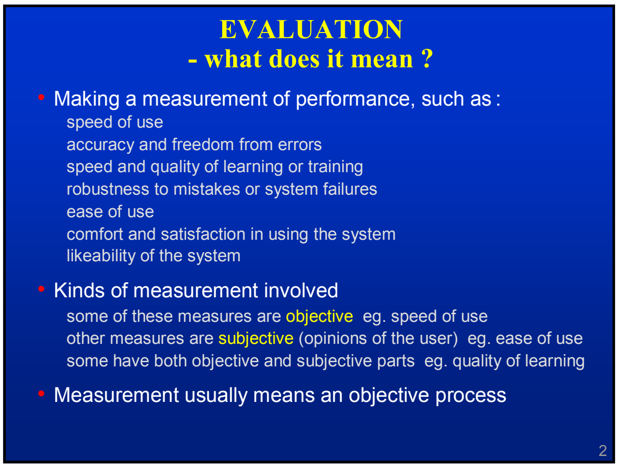
Evaluation - motives and methods

Computing & Electronic Systems

Autumn 2008

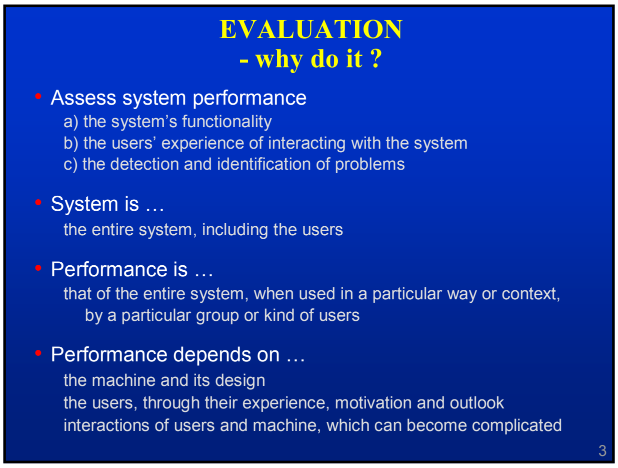
John Foster (module supervisor)

and Edward Tsang



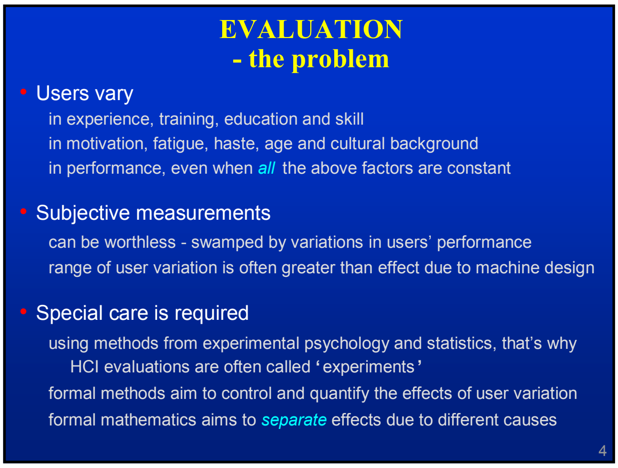
EVALUATION - what does it mean ?

- Making a measurement of performance, such as :
  - speed of use
  - accuracy and freedom from errors
  - speed and quality of learning or training
  - robustness to mistakes or system failures
  - ease of use
  - comfort and satisfaction in using the system
  - likeability of the system
- Kinds of measurement involved
  - some of these measures are objective eg. speed of use
  - other measures are subjective (opinions of the user) eg. ease of use
  - some have both objective and subjective parts eg. quality of learning
- Measurement usually means an objective process



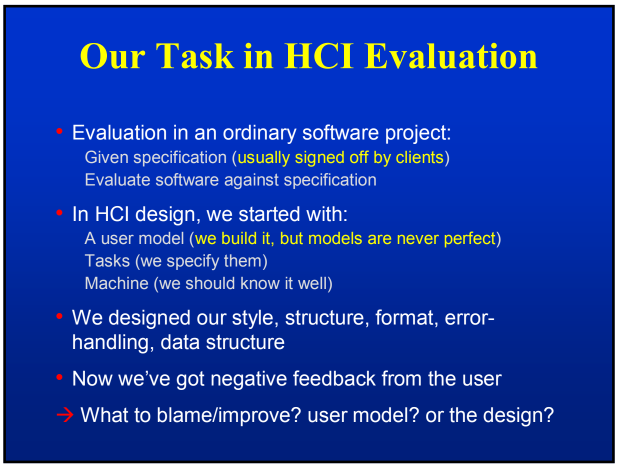
EVALUATION - why do it ?

- Assess system performance
  - a) the system's functionality
  - b) the users' experience of interacting with the system
  - c) the detection and identification of problems
- System is …
  - the entire system, including the users
- Performance is …
  - that of the entire system, when used in a particular way or context, by a particular group or kind of users
- Performance depends on …
  - the machine and its design
  - the users, through their experience, motivation and outlook
  - interactions of users and machine, which can become complicated



EVALUATION - the problem

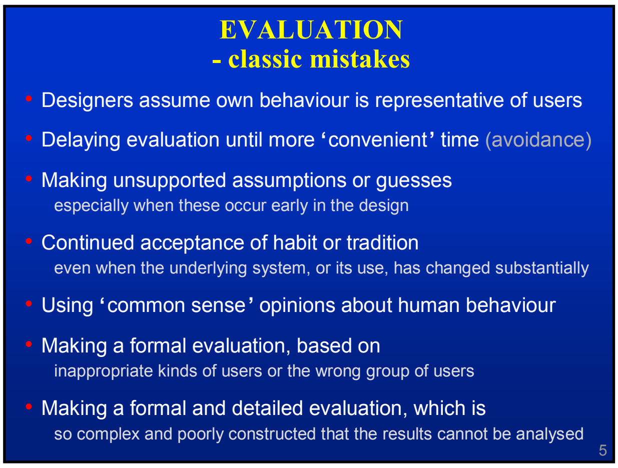
- Users vary
  - in experience, training, education and skill
  - in motivation, fatigue, haste, age and cultural background
  - in performance, even when *all* the above factors are constant
- Subjective measurements
  - can be worthless - swamped by variations in users' performance
  - range of user variation is often greater than effect due to machine design
- Special care is required
  - using methods from experimental psychology and statistics, that's why HCI evaluations are often called 'experiments'
  - formal methods aim to control and quantify the effects of user variation
  - formal mathematics aims to *separate* effects due to different causes



Our Task in HCI Evaluation

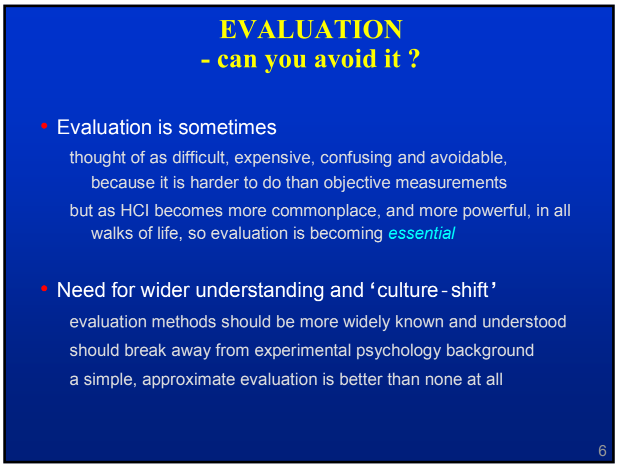
- Evaluation in an ordinary software project:
  - Given specification (usually signed off by clients)
  - Evaluate software against specification
- In HCI design, we started with:
  - A user model (we build it, but models are never perfect)
  - Tasks (we specify them)
  - Machine (we should know it well)
- We designed our style, structure, format, error-handling, data structure
- Now we've got negative feedback from the user

→ What to blame/improve? user model? or the design?



EVALUATION - classic mistakes

- Designers assume own behaviour is representative of users
- Delaying evaluation until more 'convenient' time (avoidance)
- Making unsupported assumptions or guesses
  - especially when these occur early in the design
- Continued acceptance of habit or tradition
  - even when the underlying system, or its use, has changed substantially
- Using 'common sense' opinions about human behaviour
- Making a formal evaluation, based on
  - inappropriate kinds of users or the wrong group of users
- Making a formal and detailed evaluation, which is
  - so complex and poorly constructed that the results cannot be analysed



EVALUATION
- can you avoid it ?

- Evaluation is sometimes

  thought of as difficult, expensive, confusing and avoidable,
  because it is harder to do than objective measurements

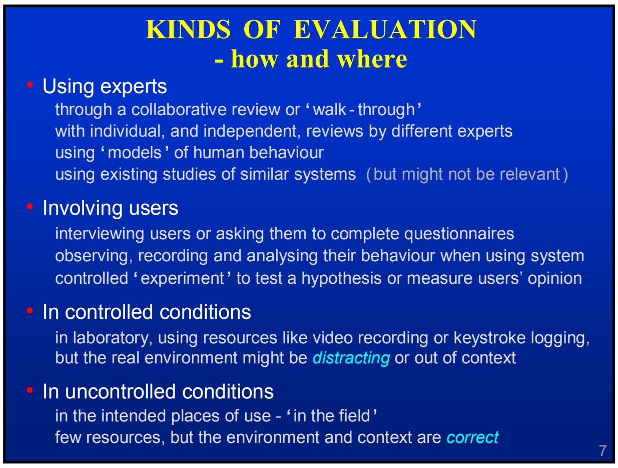
  but as HCI becomes more commonplace, and more powerful, in all walks of life, so evaluation is becoming *essential*

- Need for wider understanding and ‘culture - shift’

  evaluation methods should be more widely known and understood

  should break away from experimental psychology background

  a simple, approximate evaluation is better than none at all



KINDS OF EVALUATION
- how and where

- Using experts

  through a collaborative review or ‘walk - through’
  with individual, and independent, reviews by different experts
  using ‘models’ of human behaviour
  using existing studies of similar systems (but might not be relevant)

- Involving users

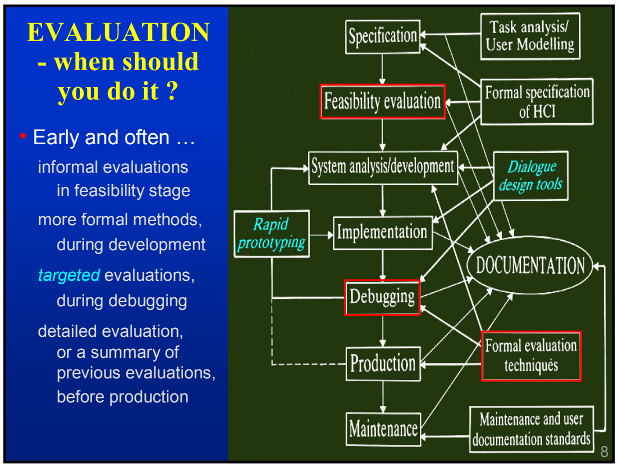
  interviewing users or asking them to complete questionnaires
  observing, recording and analysing their behaviour when using system
  controlled ‘experiment’ to test a hypothesis or measure users’ opinion

- In controlled conditions

  in laboratory, using resources like video recording or keystroke logging,
  but the real environment might be *distracting* or out of context

- In uncontrolled conditions

  in the intended places of use - ‘in the field’
  few resources, but the environment and context are *correct*



EVALUATION
- when should you do it ?

- Early and often …

  informal evaluations
  in feasibility stage

  more formal methods,
  during development

  targeted evaluations,
  during debugging

  detailed evaluation,
  or a summary of
  previous evaluations,
  before production

Specification

Task analysis/
User Modelling

Feasibility evaluation

Formal specification
of HCI

System analysis/development

Dialogue
design tools

Rapid
prototyping

Implementation

DOCUMENTATION

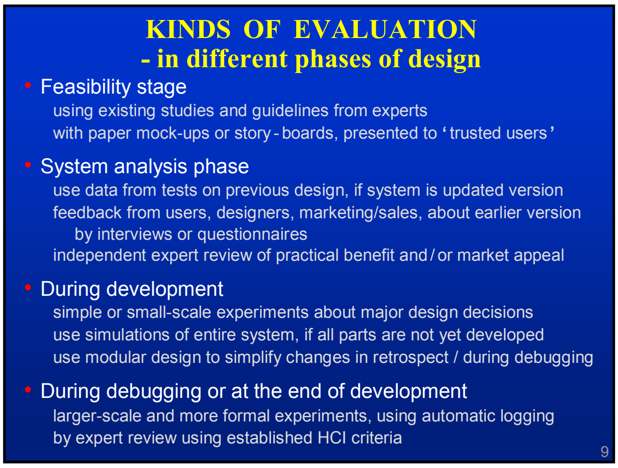
Debugging

Formal evaluation
techniquès

Production

Maintenance

Maintenance and user
documentation standards



KINDS OF EVALUATION
- in different phases of design

- Feasibility stage

  using existing studies and guidelines from experts
  with paper mock-ups or story - boards, presented to ‘trusted users’

- System analysis phase

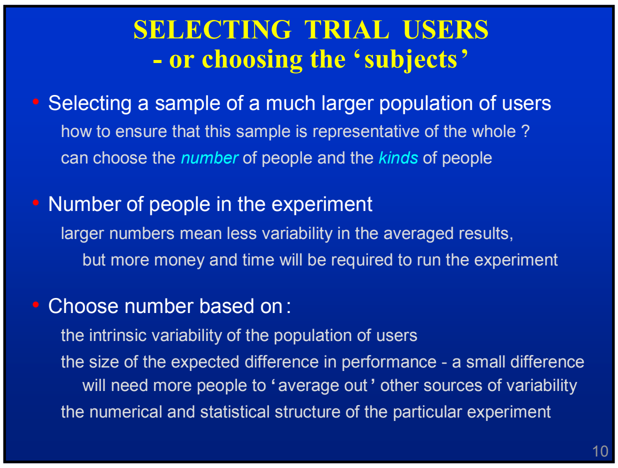
  use data from tests on previous design, if system is updated version
  feedback from users, designers, marketing/sales, about earlier version
  by interviews or questionnaires
  independent expert review of practical benefit and / or market appeal

- During development

  simple or small-scale experiments about major design decisions
  use simulations of entire system, if all parts are not yet developed
  use modular design to simplify changes in retrospect / during debugging

- During debugging or at the end of development

  larger-scale and more formal experiments, using automatic logging
  by expert review using established HCI criteria



SELECTING TRIAL USERS
- or choosing the ‘subjects’

- Selecting a sample of a much larger population of users

  how to ensure that this sample is representative of the whole ?

  can choose the *number* of people and the *kinds* of people

- Number of people in the experiment

  larger numbers mean less variability in the averaged results,

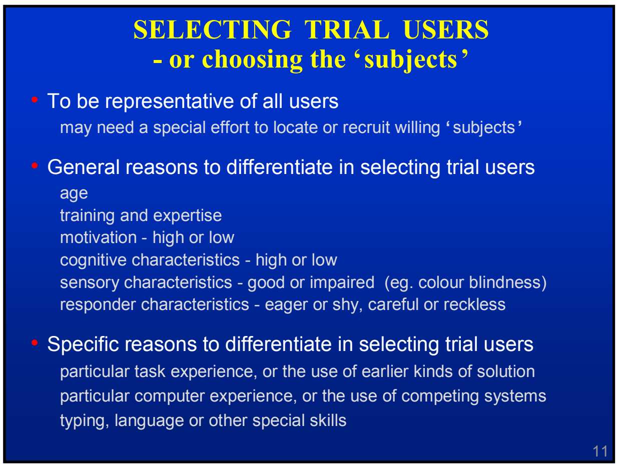
  but more money and time will be required to run the experiment

- Choose number based on :

  the intrinsic variability of the population of users

  the size of the expected difference in performance - a small difference
  will need more people to ‘average out’ other sources of variability

  the numerical and statistical structure of the particular experiment



SELECTING TRIAL USERS
- or choosing the ‘subjects’

- To be representative of all users

  may need a special effort to locate or recruit willing ‘subjects’

- General reasons to differentiate in selecting trial users

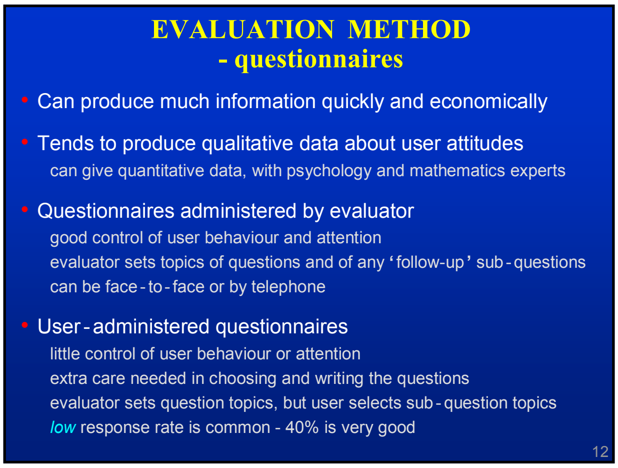
  age
  training and expertise
  motivation - high or low
  cognitive characteristics - high or low
  sensory characteristics - good or impaired (eg. colour blindness)
  responder characteristics - eager or shy, careful or reckless

- Specific reasons to differentiate in selecting trial users

  particular task experience, or the use of earlier kinds of solution

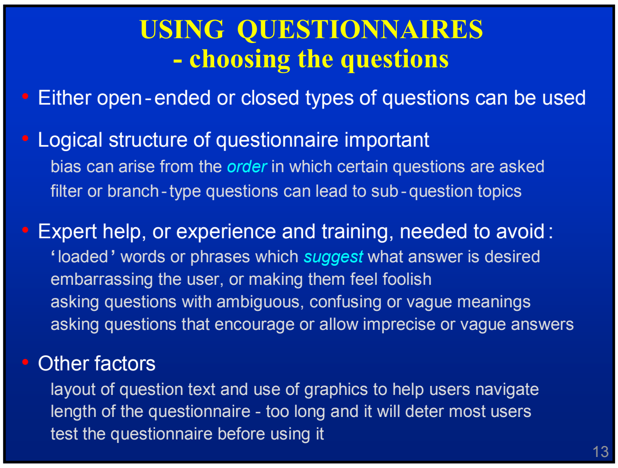
  particular computer experience, or the use of competing systems

  typing, language or other special skills



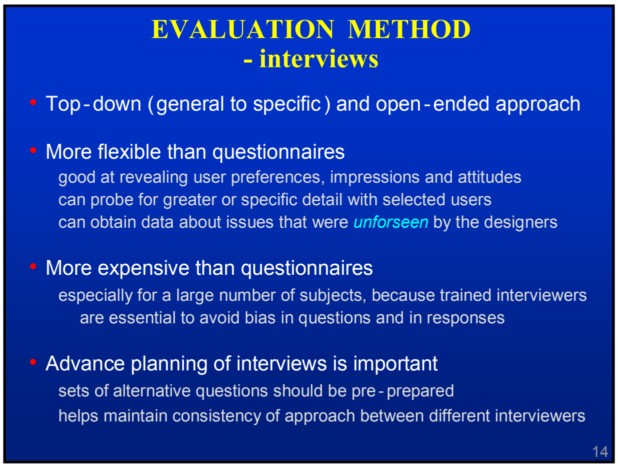
EVALUATION METHOD
- questionnaires

- Can produce much information quickly and economically
- Tends to produce qualitative data about user attitudes
  - can give quantitative data, with psychology and mathematics experts
- Questionnaires administered by evaluator
  - good control of user behaviour and attention
  - evaluator sets topics of questions and of any 'follow-up' sub-questions
  - can be face-to-face or by telephone
- User-administered questionnaires
  - little control of user behaviour or attention
  - extra care needed in choosing and writing the questions
  - evaluator sets question topics, but user selects sub-question topics
  - *low* response rate is common - 40% is very good



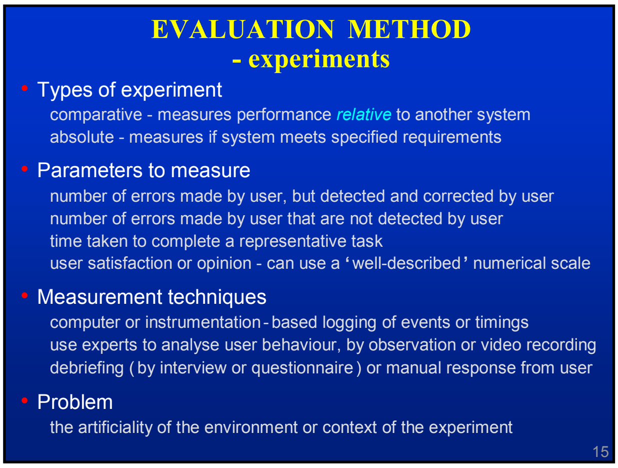
USING QUESTIONNAIRES
- choosing the questions

- Either open-ended or closed types of questions can be used
- Logical structure of questionnaire important
  - bias can arise from the *order* in which certain questions are asked
  - filter or branch-type questions can lead to sub-question topics
- Expert help, or experience and training, needed to avoid:
  - 'loaded' words or phrases which *suggest* what answer is desired
  - embarrassing the user, or making them feel foolish
  - asking questions with ambiguous, confusing or vague meanings
  - asking questions that encourage or allow imprecise or vague answers
- Other factors
  - layout of question text and use of graphics to help users navigate
  - length of the questionnaire - too long and it will deter most users
  - test the questionnaire before using it



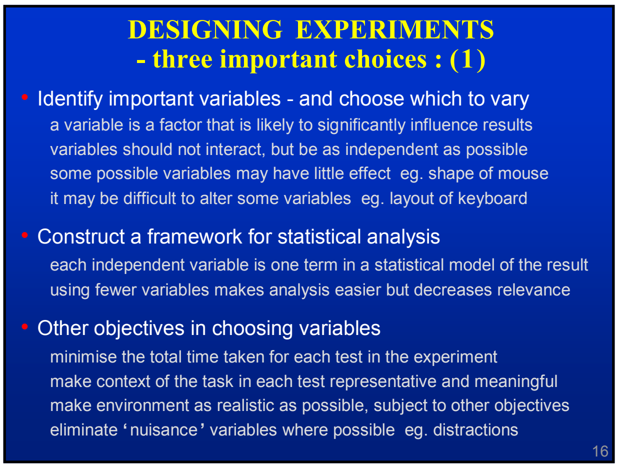
EVALUATION METHOD
- interviews

- Top-down (general to specific) and open-ended approach
- More flexible than questionnaires
  - good at revealing user preferences, impressions and attitudes
  - can probe for greater or specific detail with selected users
  - can obtain data about issues that were *unforseen* by the designers
- More expensive than questionnaires
  - especially for a large number of subjects, because trained interviewers are essential to avoid bias in questions and in responses
- Advance planning of interviews is important
  - sets of alternative questions should be pre-prepared
  - helps maintain consistency of approach between different interviewers



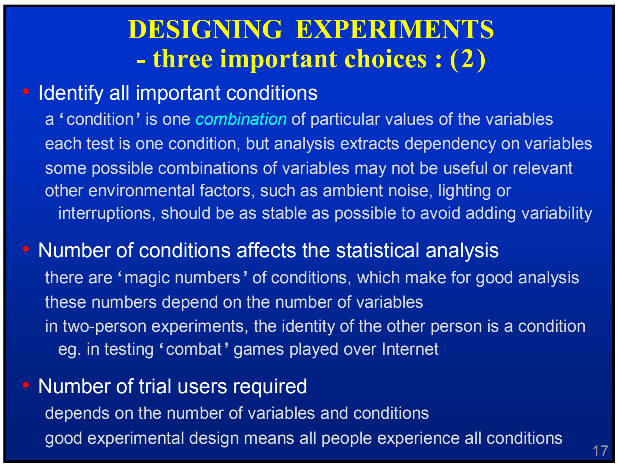
EVALUATION METHOD
- experiments

- Types of experiment
  - comparative - measures performance *relative* to another system
  - absolute - measures if system meets specified requirements
- Parameters to measure
  - number of errors made by user, but detected and corrected by user
  - number of errors made by user that are not detected by user
  - time taken to complete a representative task
  - user satisfaction or opinion - can use a 'well-described' numerical scale
- Measurement techniques
  - computer or instrumentation-based logging of events or timings
  - use experts to analyse user behaviour, by observation or video recording
  - debriefing (by interview or questionnaire) or manual response from user
- Problem
  - the artificiality of the environment or context of the experiment



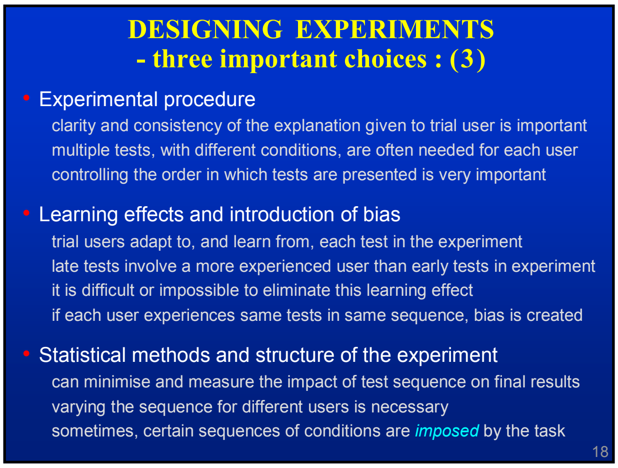
DESIGNING EXPERIMENTS
- three important choices : (1)

- Identify important variables - and choose which to vary
  - a variable is a factor that is likely to significantly influence results
  - variables should not interact, but be as independent as possible
  - some possible variables may have little effect eg. shape of mouse
  - it may be difficult to alter some variables eg. layout of keyboard
- Construct a framework for statistical analysis
  - each independent variable is one term in a statistical model of the result
  - using fewer variables makes analysis easier but decreases relevance
- Other objectives in choosing variables
  - minimise the total time taken for each test in the experiment
  - make context of the task in each test representative and meaningful
  - make environment as realistic as possible, subject to other objectives
  - eliminate 'nuisance' variables where possible eg. distractions



DESIGNING EXPERIMENTS
- three important choices : (2)

- Identify all important conditions
  - a 'condition' is one *combination* of particular values of the variables
  - each test is one condition, but analysis extracts dependency on variables
  - some possible combinations of variables may not be useful or relevant
  - other environmental factors, such as ambient noise, lighting or interruptions, should be as stable as possible to avoid adding variability
- Number of conditions affects the statistical analysis
  - there are 'magic numbers' of conditions, which make for good analysis
  - these numbers depend on the number of variables
  - in two-person experiments, the identity of the other person is a condition eg. in testing 'combat' games played over Internet
- Number of trial users required
  - depends on the number of variables and conditions
  - good experimental design means all people experience all conditions



DESIGNING EXPERIMENTS
- three important choices : (3)

- Experimental procedure
  - clarity and consistency of the explanation given to trial user is important
  - multiple tests, with different conditions, are often needed for each user
  - controlling the order in which tests are presented is very important
- Learning effects and introduction of bias
  - trial users adapt to, and learn from, each test in the experiment
  - late tests involve a more experienced user than early tests in experiment
  - it is difficult or impossible to eliminate this learning effect
  - if each user experiences same tests in same sequence, bias is created
- Statistical methods and structure of the experiment
  - can minimise and measure the impact of test sequence on final results
  - varying the sequence for different users is necessary
  - sometimes, certain sequences of conditions are *imposed* by the task

18