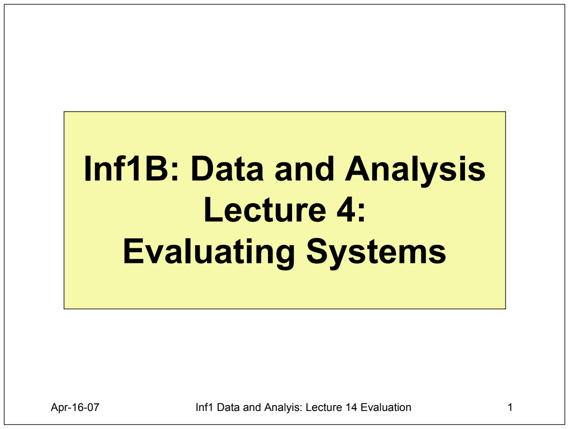
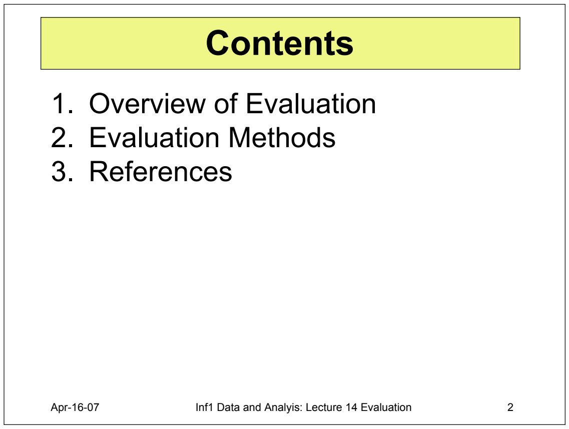
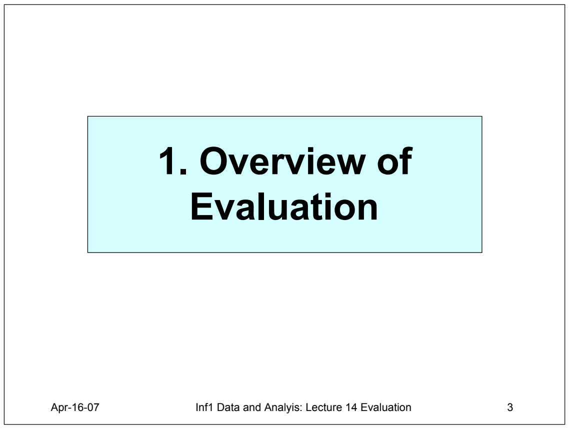


Inf1B: Data and Analysis Lecture 4: Evaluating Systems

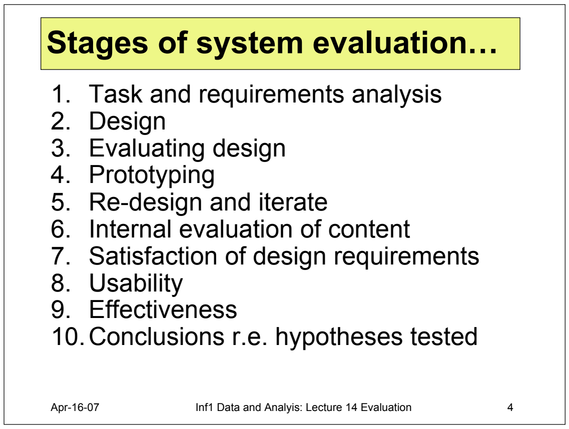


Contents

1. Overview of Evaluation
2. Evaluation Methods
3. References

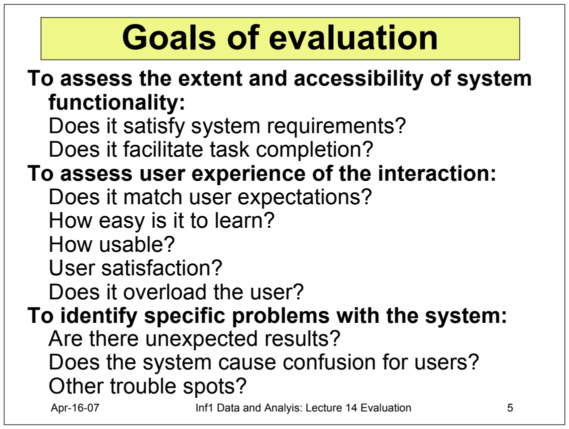


1. Overview of Evaluation



Stages of system evaluation…

1. Task and requirements analysis
2. Design
3. Evaluating design
4. Prototyping
5. Re-design and iterate
6. Internal evaluation of content
7. Satisfaction of design requirements
8. Usability
9. Effectiveness
10. Conclusions r.e. hypotheses tested

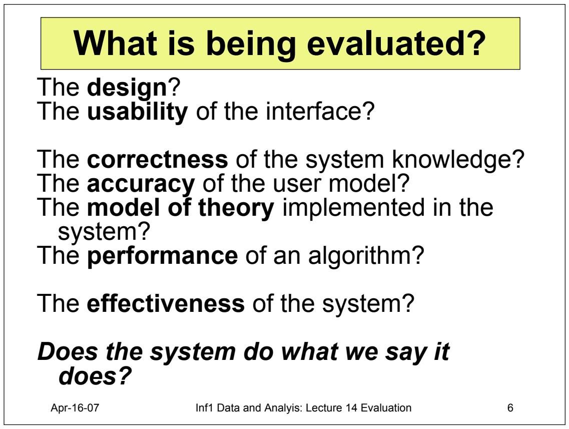


Goals of evaluation

To assess the extent and accessibility of system functionality:
Does it satisfy system requirements?
Does it facilitate task completion?

To assess user experience of the interaction:
Does it match user expectations?
How easy is it to learn?
How usable?
User satisfaction?
Does it overload the user?

To identify specific problems with the system:
Are there unexpected results?
Does the system cause confusion for users?
Other trouble spots?



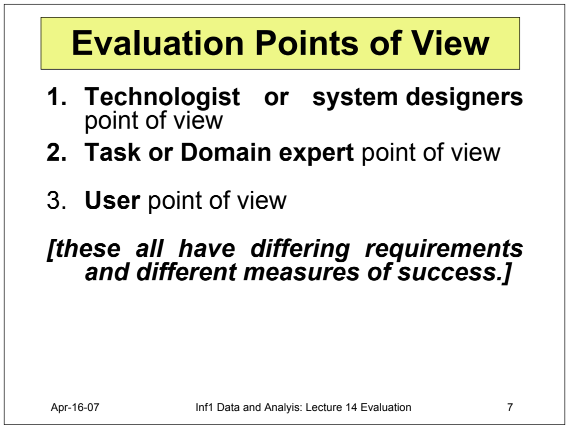
What is being evaluated?

The **design**?
The **usability** of the interface?

The **correctness** of the system knowledge?
The **accuracy** of the user model?
The **model of theory** implemented in the system?
The **performance** of an algorithm?

The **effectiveness** of the system?

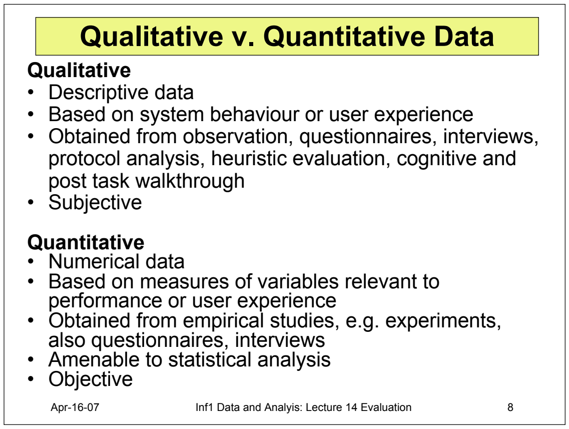
Does the system do what we say it does?



Evaluation Points of View

1. **Technologist or system designers** point of view
2. **Task or Domain expert** point of view
3. **User** point of view

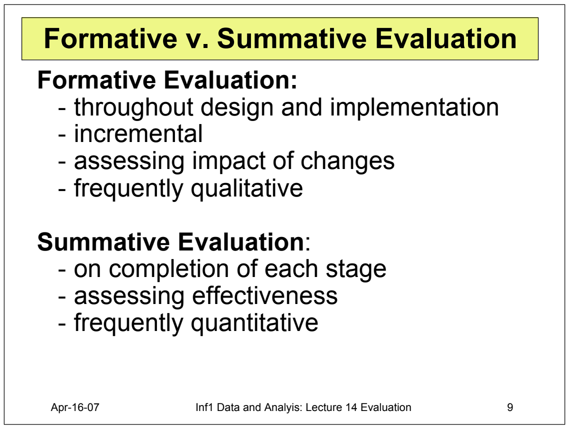
[these all have differing requirements and different measures of success.]



Qualitative v. Quantitative Data

Qualitative
- Descriptive data
- Based on system behaviour or user experience
- Obtained from observation, questionnaires, interviews, protocol analysis, heuristic evaluation, cognitive and post task walkthrough
- Subjective

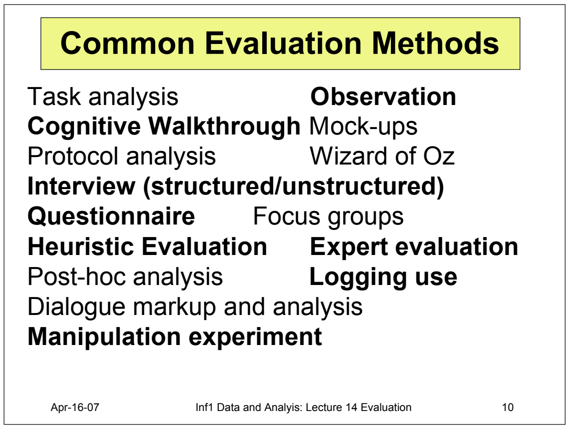
Quantitative
- Numerical data
- Based on measures of variables relevant to performance or user experience
- Obtained from empirical studies, e.g. experiments, also questionnaires, interviews
- Amenable to statistical analysis
- Objective



Formative v. Summative Evaluation

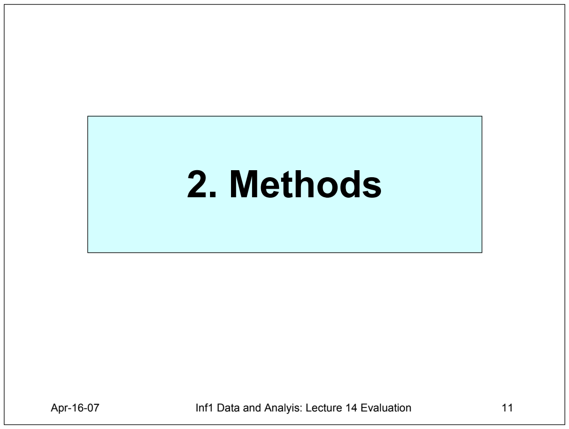
Formative Evaluation:
- throughout design and implementation
- incremental
- assessing impact of changes
- frequently qualitative

Summative Evaluation:
- on completion of each stage
- assessing effectiveness
- frequently quantitative

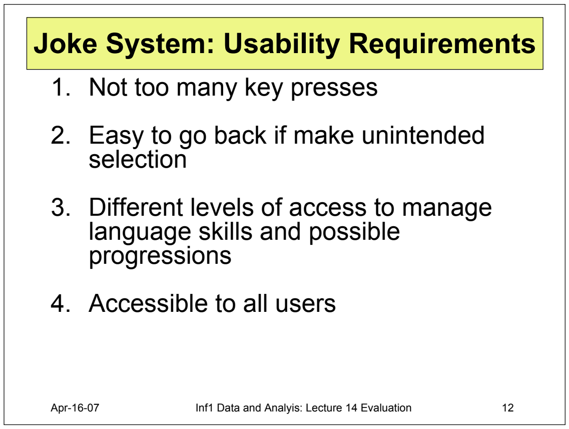


Common Evaluation Methods

Task analysis **Observation**
Cognitive Walkthrough Mock-ups
Protocol analysis Wizard of Oz
Interview (structured/unstructured)
Questionnaire Focus groups
Heuristic Evaluation **Expert evaluation**
Post-hoc analysis **Logging use**
Dialogue markup and analysis
Manipulation experiment

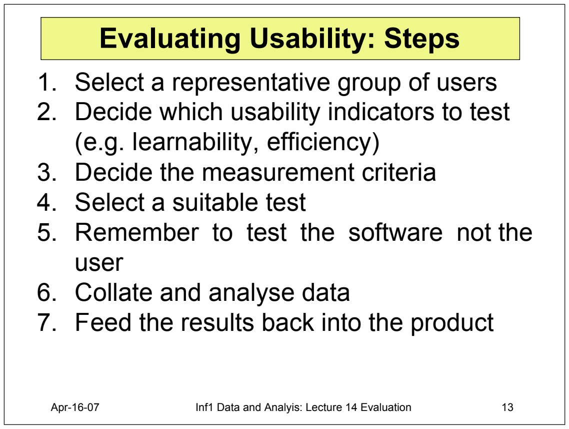


2. Methods



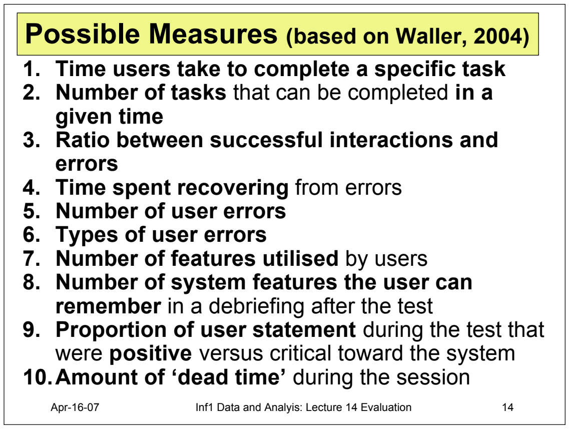
Joke System: Usability Requirements

1. Not too many key presses
2. Easy to go back if make unintended selection
3. Different levels of access to manage language skills and possible progressions
4. Accessible to all users



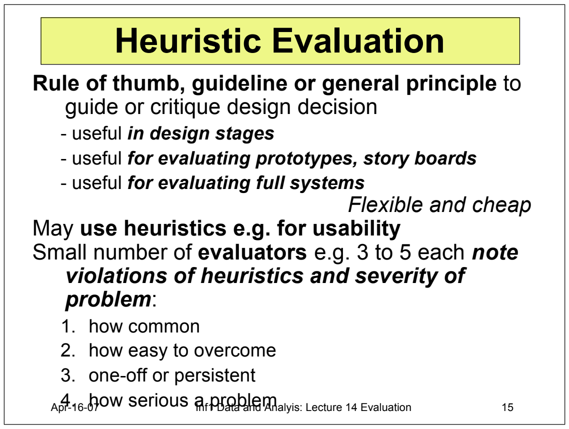
Evaluating Usability: Steps

1. Select a representative group of users
2. Decide which usability indicators to test (e.g. learnability, efficiency)
3. Decide the measurement criteria
4. Select a suitable test
5. Remember to test the software not the user
6. Collate and analyse data
7. Feed the results back into the product



Possible Measures (based on Waller, 2004)

1. **Time users take to complete a specific task**
2. **Number of tasks** that can be completed **in a given time**
3. **Ratio between successful interactions and errors**
4. **Time spent recovering** from errors
5. **Number of user errors**
6. **Types of user errors**
7. **Number of features utilised** by users
8. **Number of system features the user can remember** in a debriefing after the test
9. **Proportion of user statement** during the test that were **positive** versus critical toward the system
10. **Amount of 'dead time'** during the session



Heuristic Evaluation

Rule of thumb, guideline or general principle to guide or critique design decision

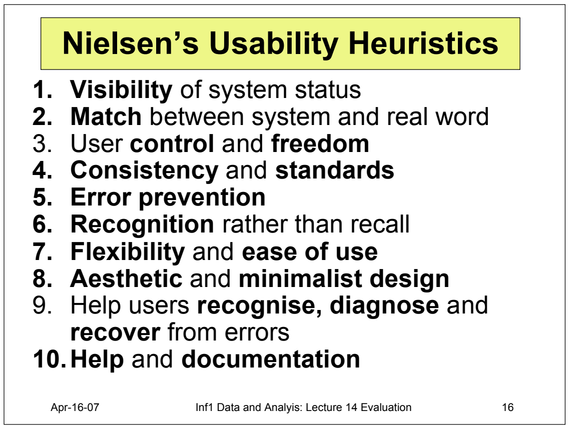
- useful ***in design stages***
- useful ***for evaluating prototypes, story boards***
- useful ***for evaluating full systems***

Flexible and cheap

May **use heuristics e.g. for usability**

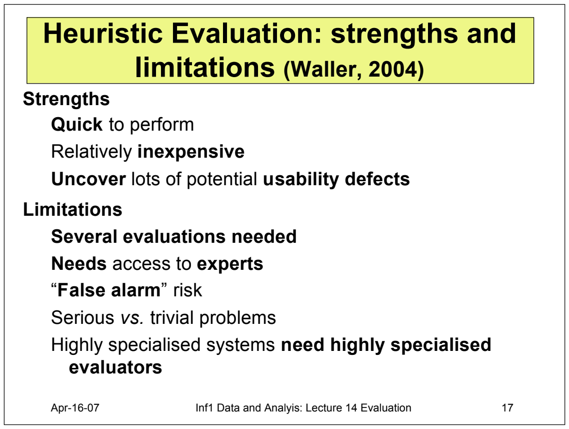
Small number of **evaluators** e.g. 3 to 5 each ***note violations of heuristics and severity of problem***:

1. how common
2. how easy to overcome
3. one-off or persistent
4. how serious a problem



Nielsen's Usability Heuristics

1. **Visibility** of system status
2. **Match** between system and real word
3. User **control** and **freedom**
4. **Consistency** and **standards**
5. **Error prevention**
6. **Recognition** rather than recall
7. **Flexibility** and **ease of use**
8. **Aesthetic** and **minimalist design**
9. Help users **recognise, diagnose** and **recover** from errors
10. **Help** and **documentation**



Heuristic Evaluation: strengths and limitations (Waller, 2004)

Strengths

Quick to perform

Relatively **inexpensive**

Uncover lots of potential **usability defects**

Limitations

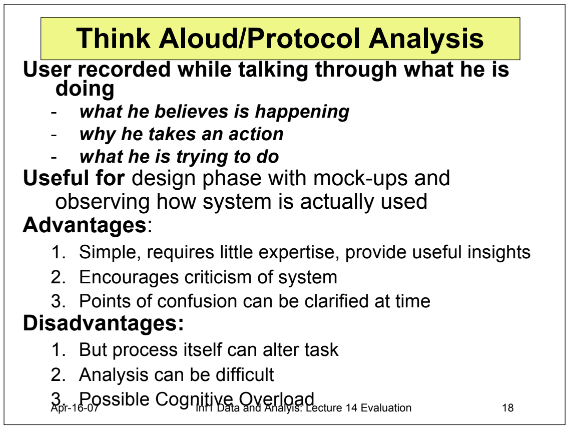
Several evaluations needed

Needs access to **experts**

"**False alarm**" risk

Serious *vs.* trivial problems

Highly specialised systems **need highly specialised evaluators**



Think Aloud/Protocol Analysis

User recorded while talking through what he is doing

- ***what he believes is happening***
- ***why he takes an action***
- ***what he is trying to do***

Useful for design phase with mock-ups and observing how system is actually used

Advantages:

1. Simple, requires little expertise, provide useful insights
2. Encourages criticism of system
3. Points of confusion can be clarified at time

Disadvantages:

1. But process itself can alter task
2. Analysis can be difficult
3. Possible Cognitive Overload

Logging Use

Automatic recording of user actions can be built into software for later analysis

- Enables replay of full interaction
- Keystroke and mouse movement
- Errors
- Timing and duration of tasks and sub-tasks

Advantages:

1. Objective data
2. Can identify frequent use of features
3. Automatic, and unobtrusive

Disadvantages:

1. Actions logged need to be interpreted
2. Technical problem and file storage
3. Privacy issues

Cognitive Walkthrough

User is asked to reflect on actions and decisions taken in performing a task, **post-task**

1. Re-enact task, replay session or use session transcript
2. User is asked questions at particular points of interest

Timing:

- ***immediately post-task*** (easier for user to remember)
- ***later*** (more time for evaluator to identify points of interest)

Useful when talk aloud would be too intrusive

Physiological Responses:Eye Tracking

Measure **how users feel** as well as what they do

Eye Tracking: now less invasive (not previously suitable for usability testing)

- Reflect amount of cognitive processing required for tasks
- Patterns of movement may suggest areas of screen that are easy/difficult to process

Can measure:

1. Number of fixations
2. Fixation duration
3. Scan path

Need more work on how to interpret*, e.g. if looking at text is user reading it?*

Becoming standard equipment

Physiological Responses: other measures

Emotional response may be measured through:

- **Heart activity** - may indicate stress, anger
- Sweat via **Galvanic skin response** (GSR) - higher emotional state, effort
- **Electrical activity in muscles** (EMG) - task involvement
- **Electrical activity in brain** (ECG) - decision making, motivation, attention
- Other **stress measures**, e.g. pressure on mouse/keys

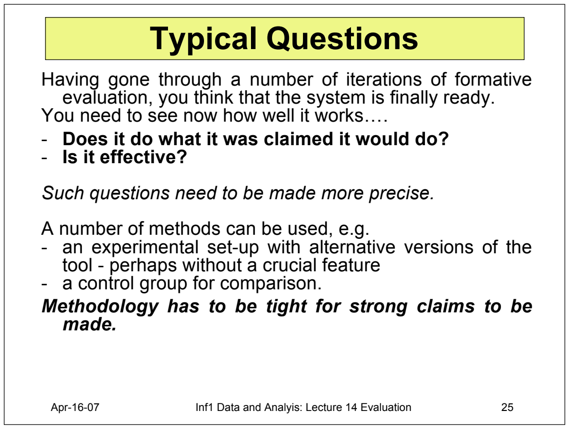
Exact relation between events and measures is not always clear

Offers possibly objective information in particular to inform affective state of user

Methods for collecting maths errors

Task analysis **Observation**
Cognitive Walkthrough Mock-ups
Protocol analysis **Wizard of Oz**
Video Recording **Interview**
Questionnaire Focus groups
Sensitivity Analysis Expert evaluation
Post-hoc analysis **Logging use**
Dialogue mark-up and analysis
Manipulation experiment
Self Report Sentient analysis

3. Experimental Design



Typical Questions

Having gone through a number of iterations of formative evaluation, you think that the system is finally ready.
You need to see now how well it works….

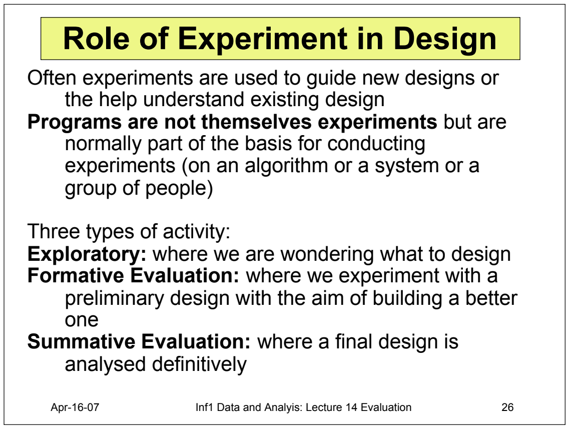
- **Does it do what it was claimed it would do?**
- **Is it effective?**

Such questions need to be made more precise.

A number of methods can be used, e.g.

- an experimental set-up with alternative versions of the tool - perhaps without a crucial feature
- a control group for comparison.

Methodology has to be tight for strong claims to be made.



Role of Experiment in Design

Often experiments are used to guide new designs or the help understand existing design

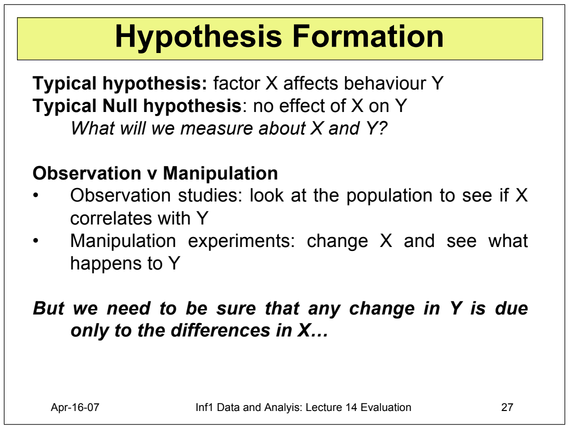
Programs are not themselves experiments but are normally part of the basis for conducting experiments (on an algorithm or a system or a group of people)

Three types of activity:

Exploratory: where we are wondering what to design

Formative Evaluation: where we experiment with a preliminary design with the aim of building a better one

Summative Evaluation: where a final design is analysed definitively



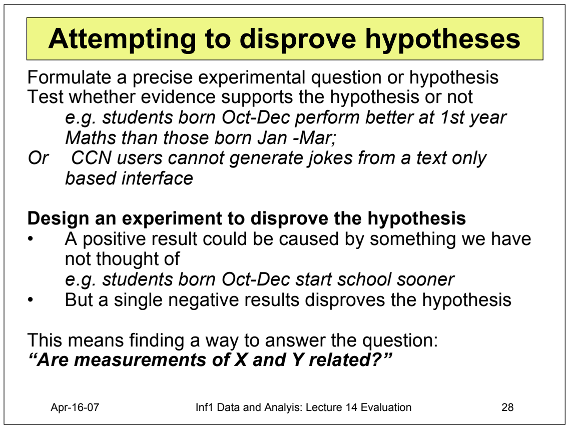
Hypothesis Formation

Typical hypothesis: factor X affects behaviour Y
Typical Null hypothesis: no effect of X on Y
What will we measure about X and Y?

Observation v Manipulation

- Observation studies: look at the population to see if X correlates with Y
- Manipulation experiments: change X and see what happens to Y

But we need to be sure that any change in Y is due only to the differences in X…



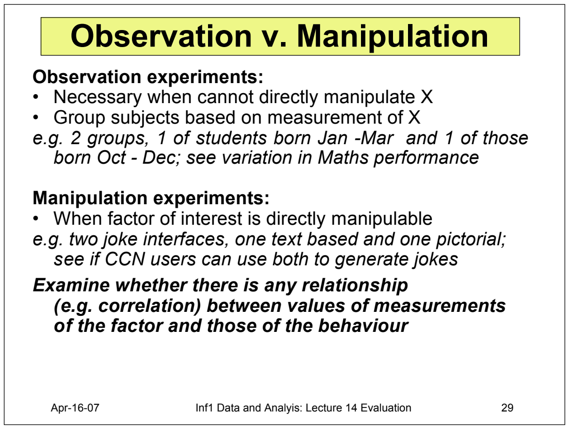
Attempting to disprove hypotheses

Formulate a precise experimental question or hypothesis
Test whether evidence supports the hypothesis or not
e.g. students born Oct-Dec perform better at 1st year Maths than those born Jan -Mar;
Or CCN users cannot generate jokes from a text only based interface

Design an experiment to disprove the hypothesis

- A positive result could be caused by something we have not thought of
  e.g. students born Oct-Dec start school sooner
- But a single negative results disproves the hypothesis

This means finding a way to answer the question:
"Are measurements of X and Y related?"



Observation v. Manipulation

Observation experiments:

- Necessary when cannot directly manipulate X
- Group subjects based on measurement of X

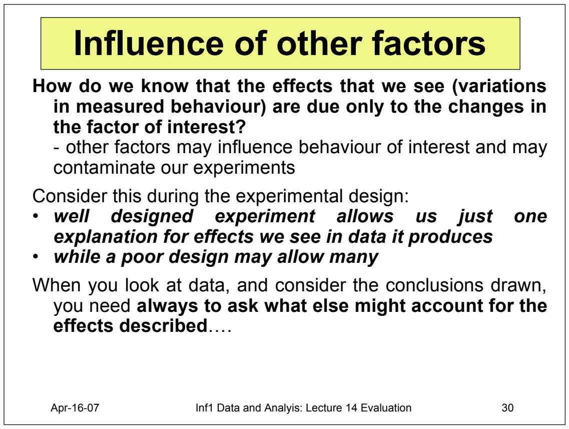
e.g. 2 groups, 1 of students born Jan -Mar and 1 of those born Oct - Dec; see variation in Maths performance

Manipulation experiments:

- When factor of interest is directly manipulable

e.g. two joke interfaces, one text based and one pictorial; see if CCN users can use both to generate jokes

Examine whether there is any relationship (e.g. correlation) between values of measurements of the factor and those of the behaviour



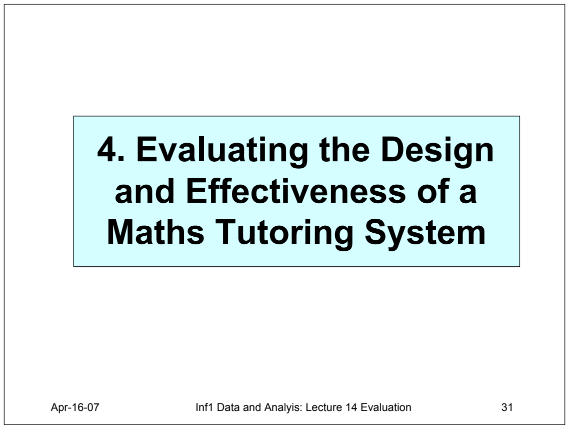
Influence of other factors

How do we know that the effects that we see (variations in measured behaviour) are due only to the changes in the factor of interest?
- other factors may influence behaviour of interest and may contaminate our experiments

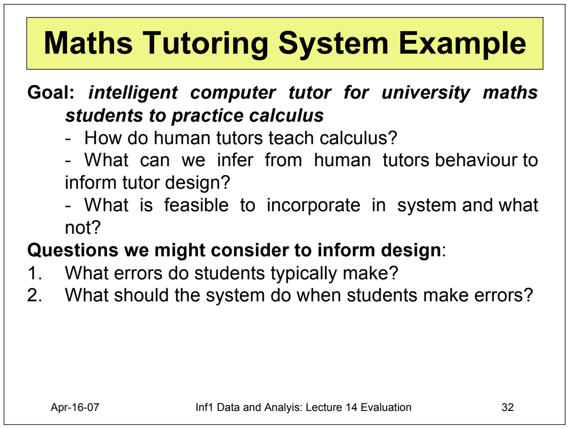
Consider this during the experimental design:

- ***well designed experiment allows us just one explanation for effects we see in data it produces***
- ***while a poor design may allow many***

When you look at data, and consider the conclusions drawn, you need **always to ask what else might account for the effects described**….



4. Evaluating the Design and Effectiveness of a Maths Tutoring System



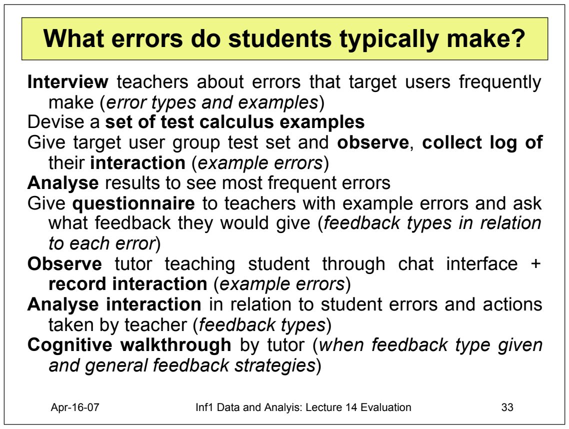
Maths Tutoring System Example

Goal: ***intelligent computer tutor for university maths students to practice calculus***

- How do human tutors teach calculus?
- What can we infer from human tutors behaviour to inform tutor design?
- What is feasible to incorporate in system and what not?

Questions we might consider to inform design:

1. What errors do students typically make?
2. What should the system do when students make errors?



What errors do students typically make?

Interview teachers about errors that target users frequently make (*error types and examples*)

Devise a **set of test calculus examples**

Give target user group test set and **observe**, **collect log of** their **interaction** (*example errors*)

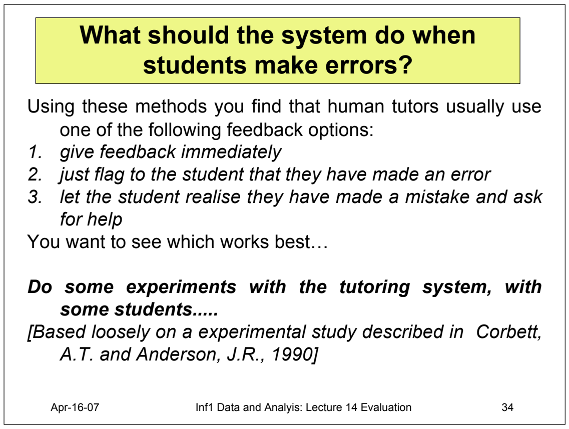
Analyse results to see most frequent errors

Give **questionnaire** to teachers with example errors and ask what feedback they would give (*feedback types in relation to each error*)

Observe tutor teaching student through chat interface + **record interaction** (*example errors*)

Analyse interaction in relation to student errors and actions taken by teacher (*feedback types*)

Cognitive walkthrough by tutor (*when feedback type given and general feedback strategies*)



What should the system do when students make errors?

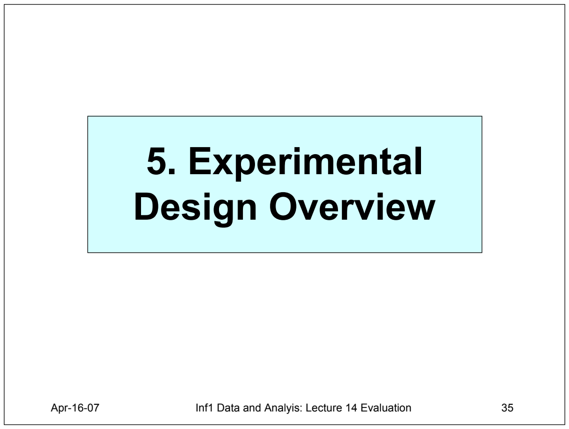
Using these methods you find that human tutors usually use one of the following feedback options:

1. *give feedback immediately*
2. *just flag to the student that they have made an error*
3. *let the student realise they have made a mistake and ask for help*

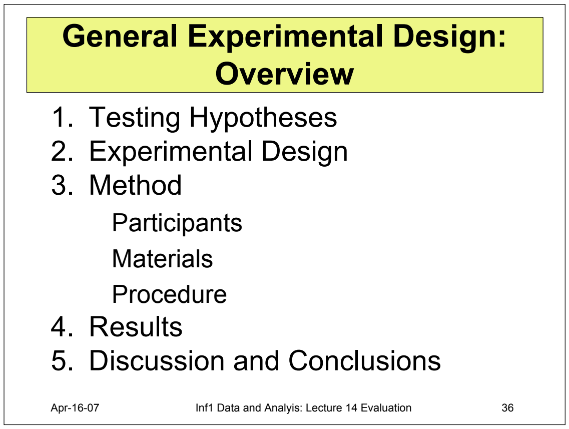
You want to see which works best…

Do some experiments with the tutoring system, with some students.....

[Based loosely on a experimental study described in Corbett, A.T. and Anderson, J.R., 1990]

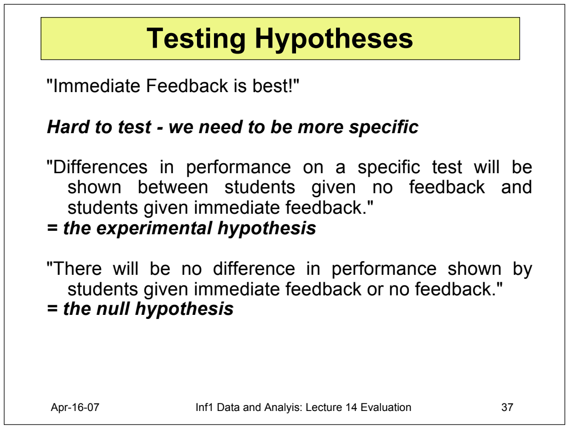


5. Experimental Design Overview



General Experimental Design: Overview

1. Testing Hypotheses
2. Experimental Design
3. Method
   - Participants
   - Materials
   - Procedure
4. Results
5. Discussion and Conclusions



Testing Hypotheses

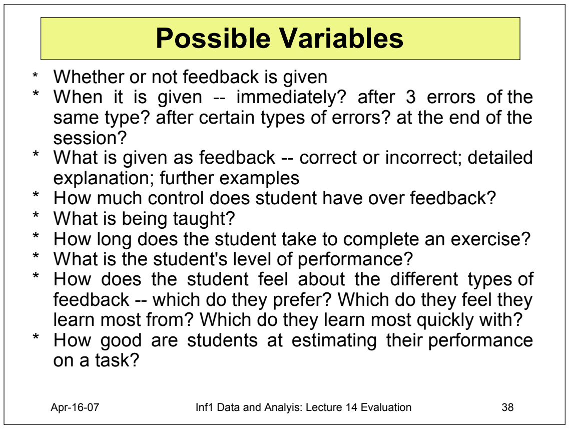
"Immediate Feedback is best!"

Hard to test - we need to be more specific

"Differences in performance on a specific test will be shown between students given no feedback and students given immediate feedback."
= the experimental hypothesis

"There will be no difference in performance shown by students given immediate feedback or no feedback."
= the null hypothesis

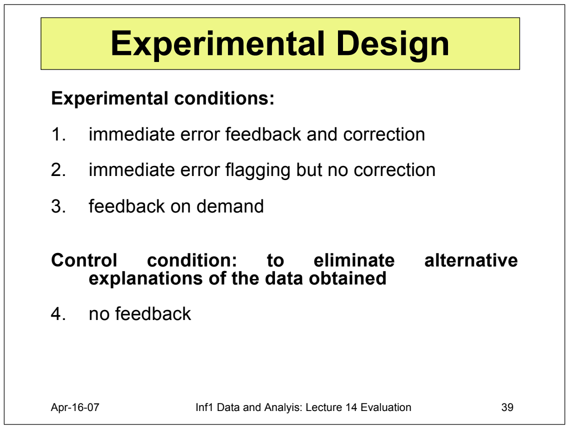
Apr-16-07 Inf1 Data and Analyis: Lecture 14 Evaluation 37



Possible Variables

- Whether or not feedback is given
- When it is given -- immediately? after 3 errors of the same type? after certain types of errors? at the end of the session?
- What is given as feedback -- correct or incorrect; detailed explanation; further examples
- How much control does student have over feedback?
- What is being taught?
- How long does the student take to complete an exercise?
- What is the student's level of performance?
- How does the student feel about the different types of feedback -- which do they prefer? Which do they feel they learn most from? Which do they learn most quickly with?
- How good are students at estimating their performance on a task?

Apr-16-07 Inf1 Data and Analyis: Lecture 14 Evaluation 38



Experimental Design

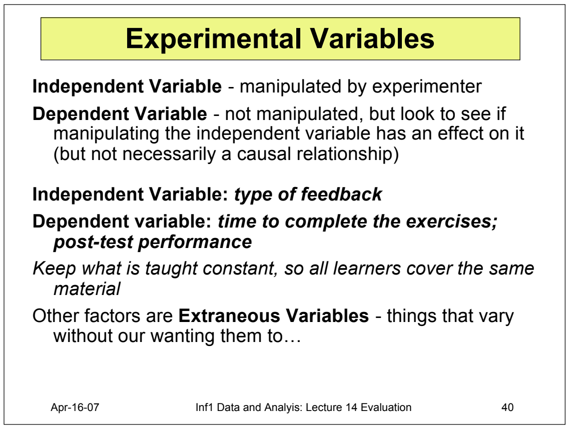
Experimental conditions:

1. immediate error feedback and correction
2. immediate error flagging but no correction
3. feedback on demand

Control condition: to eliminate alternative explanations of the data obtained

4. no feedback

Apr-16-07 Inf1 Data and Analyis: Lecture 14 Evaluation 39



Experimental Variables

Independent Variable - manipulated by experimenter

Dependent Variable - not manipulated, but look to see if manipulating the independent variable has an effect on it (but not necessarily a causal relationship)

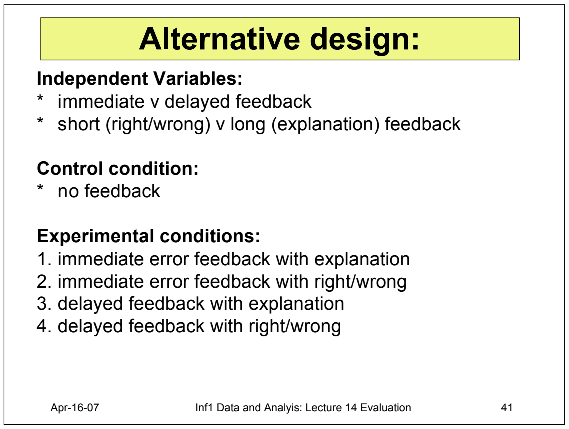
Independent Variable: *type of feedback*

Dependent variable: *time to complete the exercises; post-test performance*

Keep what is taught constant, so all learners cover the same material

Other factors are **Extraneous Variables** - things that vary without our wanting them to…

Apr-16-07 Inf1 Data and Analyis: Lecture 14 Evaluation 40



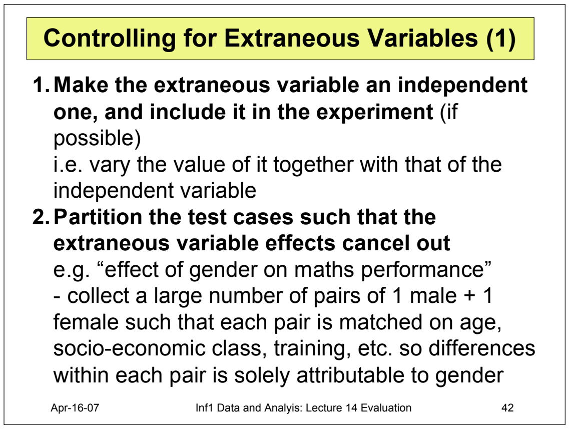
Alternative design:

Independent Variables:
* immediate v delayed feedback
* short (right/wrong) v long (explanation) feedback

Control condition:
* no feedback

Experimental conditions:
1. immediate error feedback with explanation
2. immediate error feedback with right/wrong
3. delayed feedback with explanation
4. delayed feedback with right/wrong

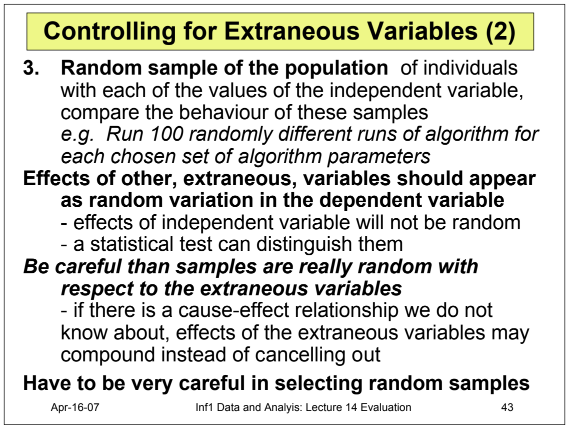
Apr-16-07 Inf1 Data and Analyis: Lecture 14 Evaluation 41



Controlling for Extraneous Variables (1)

1. **Make the extraneous variable an independent one, and include it in the experiment** (if possible)
   i.e. vary the value of it together with that of the independent variable
2. **Partition the test cases such that the extraneous variable effects cancel out**
   e.g. “effect of gender on maths performance”
   - collect a large number of pairs of 1 male + 1 female such that each pair is matched on age, socio-economic class, training, etc. so differences within each pair is solely attributable to gender

Apr-16-07 Inf1 Data and Analyis: Lecture 14 Evaluation 42



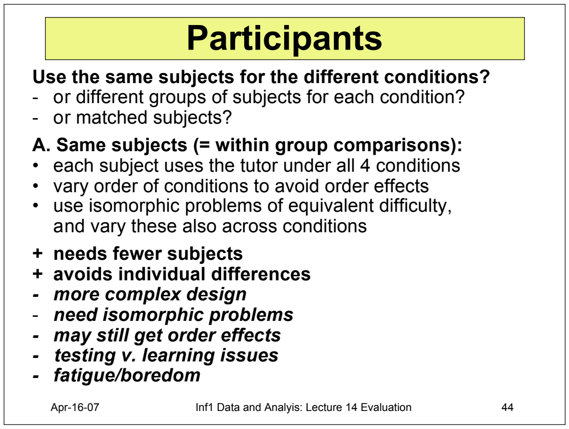
Controlling for Extraneous Variables (2)

3. **Random sample of the population** of individuals with each of the values of the independent variable, compare the behaviour of these samples
   e.g. Run 100 randomly different runs of algorithm for each chosen set of algorithm parameters

Effects of other, extraneous, variables should appear as random variation in the dependent variable
- effects of independent variable will not be random
- a statistical test can distinguish them

Be careful than samples are really random with respect to the extraneous variables
- if there is a cause-effect relationship we do not know about, effects of the extraneous variables may compound instead of cancelling out

Have to be very careful in selecting random samples

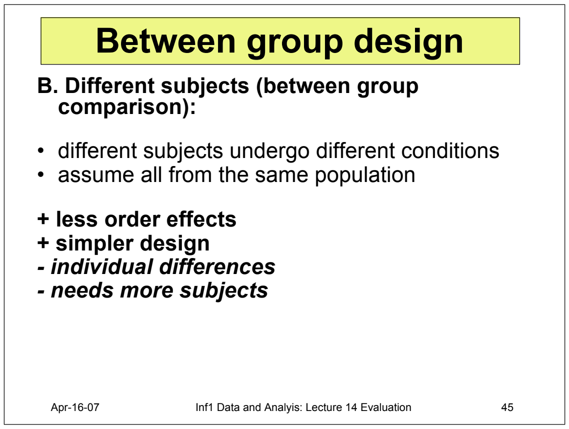


Participants

Use the same subjects for the different conditions?
- or different groups of subjects for each condition?
- or matched subjects?

A. Same subjects (= within group comparisons):
- each subject uses the tutor under all 4 conditions
- vary order of conditions to avoid order effects
- use isomorphic problems of equivalent difficulty, and vary these also across conditions

+ needs fewer subjects
+ avoids individual differences
- more complex design
- need isomorphic problems
- may still get order effects
- testing v. learning issues
- fatigue/boredom

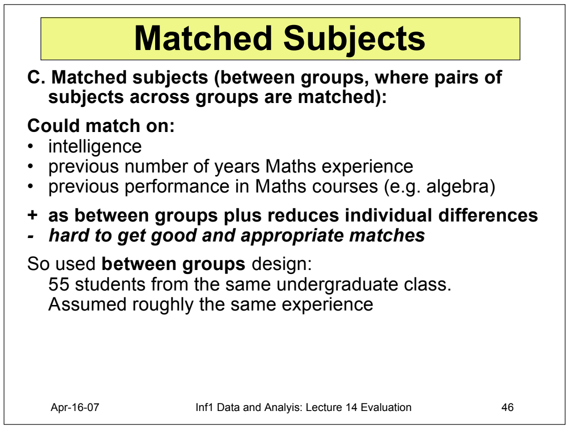


Between group design

B. Different subjects (between group comparison):

- different subjects undergo different conditions
- assume all from the same population

+ less order effects
+ simpler design
- individual differences
- needs more subjects



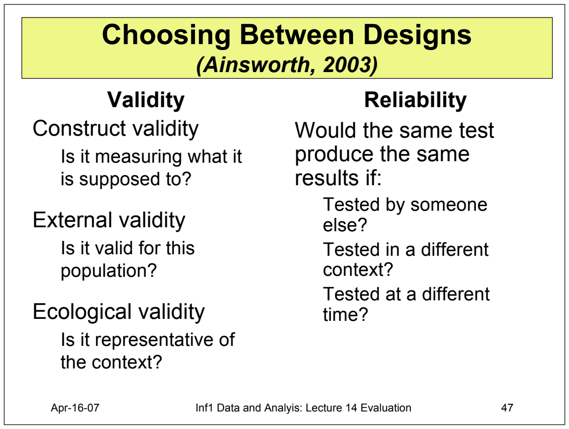
Matched Subjects

C. Matched subjects (between groups, where pairs of subjects across groups are matched):

Could match on:
- intelligence
- previous number of years Maths experience
- previous performance in Maths courses (e.g. algebra)

+ as between groups plus reduces individual differences
- hard to get good and appropriate matches

So used **between groups** design:
55 students from the same undergraduate class.
Assumed roughly the same experience



Choosing Between Designs
(Ainsworth, 2003)

Validity

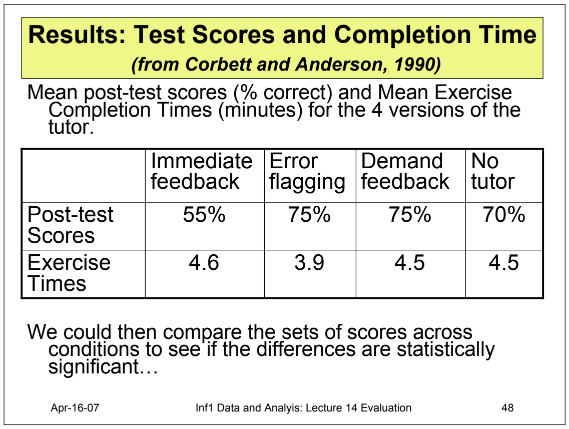
Construct validity
Is it measuring what it is supposed to?

External validity
Is it valid for this population?

Ecological validity
Is it representative of the context?

Reliability

Would the same test produce the same results if:
Tested by someone else?
Tested in a different context?
Tested at a different time?

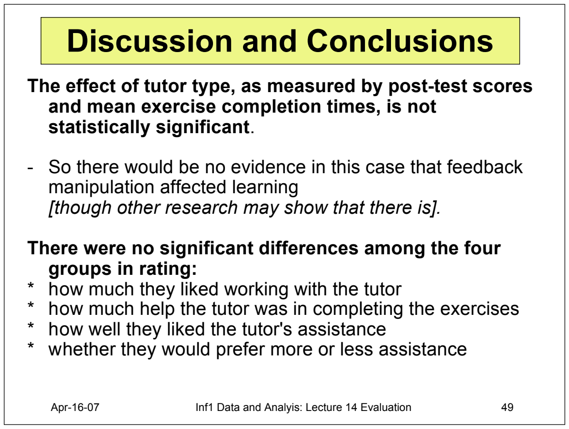


Results: Test Scores and Completion Time
(from Corbett and Anderson, 1990)

Mean post-test scores (% correct) and Mean Exercise Completion Times (minutes) for the 4 versions of the tutor.

	Immediate feedback	Error flagging	Demand feedback	No tutor
Post-test Scores	55%	75%	75%	70%
Exercise Times	4.6	3.9	4.5	4.5

We could then compare the sets of scores across conditions to see if the differences are statistically significant…



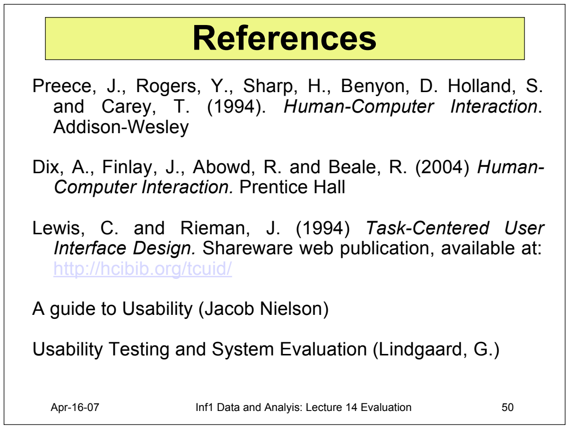
Discussion and Conclusions

The effect of tutor type, as measured by post-test scores and mean exercise completion times, is not statistically significant.

- So there would be no evidence in this case that feedback manipulation affected learning
[though other research may show that there is].

There were no significant differences among the four groups in rating:

* how much they liked working with the tutor
* how much help the tutor was in completing the exercises
* how well they liked the tutor's assistance
* whether they would prefer more or less assistance



References

Preece, J., Rogers, Y., Sharp, H., Benyon, D. Holland, S. and Carey, T. (1994). *Human-Computer Interaction*. Addison-Wesley

Dix, A., Finlay, J., Abowd, R. and Beale, R. (2004) *Human-Computer Interaction.* Prentice Hall

Lewis, C. and Rieman, J. (1994) *Task-Centered User Interface Design.* Shareware web publication, available at: http://hcibib.org/tcuid/

A guide to Usability (Jacob Nielson)

Usability Testing and System Evaluation (Lindgaard, G.)