

"How to conduct experiments with human test subjects"

30. November 2006

Lars Bo Larsen

These slides available at:

http://kom.aau.dk/~lbl/ieee_sb_user_test.pdf

Overview

Topics:

1. When and why do you need human test persons?
2. Recruiting test persons
3. Preparing the test:
 - Formulating the test objectives, preparing materials
 - Recruitment of test persons
4. Carrying out the test:
 - Who is involved: Experimenter, logger, etc
 - Test facilities: usability labs
 - Recording the test
5. Analysis and interpretation of test results
 - Questionnaires
 - Logging
 - Statistical analysis
6. Tools and references

When and why do you need human test persons?

The situations in which you will need to extract information from human users can be divided into three classes:

1. When you need to establish some statistical evidence of users opinions. Experience, knowledge, etc. about some technology or concept:
 - Example: You need to find out in what context users typically use their mobile phone for e.g. downloading music, how often they do it and whether they find the download speed (and price) acceptable
 - For this you will typically need to interview a larger number (e.g. more than 50 users) of people to get a reasonable statistical evidence of the usage.
 - For this, the obvious solution will be to carry out a *survey* i.e. addressing potential users via email and ask them to fill out a web-based questionnaire

When and why do you need human test persons?

2. When you need to investigate end users perform some task or experience a system (you built), record and study their reactions and interview them about their experiences and attitudes towards the system
 - Example: To investigate the users perceived quality of e.g. a video encoding scheme for various transmission rates and to compare it to the quality of existing encoders.
 - For this, you need to expose representative end users to a number of video sequences encoded at various bitrates and extract their opinions about their perceived quality in a face-to-face experiment

When and why do you need human test persons?

3. When you need to develop a user interface for some system or service in an iterative manner and you want to make sure that users understand how to interact with the system in an efficient and error-free manner.
 - For this, you need to study how a number of representative end users carry out some predefined tasks with the system and interview them about their opinions and experiences in a controlled environment

Recruiting test persons

In many situation, a user test is not very valuable, unless it can be generalised from the selected test persons to the target population.

- This requires a *sufficient number of representative test persons* to make statistically valid conclusions
- Otherwise we can only identify “trends”, “indications” or “preliminary results” – but that may often be sufficient

Two problems:

- How do we ensure our test persons are representative?
 - Identify target users and screen prospective test persons accordingly. Ensure demographic (gender, age, education, occupation, region, etc) even distribution
- And how many do we need?

Recruiting test persons

How many test persons do we need?

- It depends on the purpose. If you want statistically valid data, you need at least about *8-10 persons per test condition* (dependent variable). This quickly adds up!!
 - *Use factorial design to minimise number of test persons*
 - *Use within-subject design, where possible*
- If you want an indication (or trend) that e.g. your user interface does not contain major design errors, you can do with much less – often less than 20, or even below 10 test persons

Recruiting test persons

Method Name	Lifecycle Stage	Users Needed	Main Advantage	Main Disadvantage
Heuristic evaluation	Early design, "inner cycle" of iterative design	None	Finds individual usability problems. Can address expert user issues.	Does not involve real users, so does not find "surprises" relating to their needs.
Performance measures	Competitive analysis, final testing	At least 10	Hard numbers. Results easy to compare.	Does not find individual usability problems.
Thinking aloud	Iterative design, formative evaluation	3-5	Pinpoints user misconceptions. Cheap test.	Unnatural for users. Hard for expert users to verbalize.
Observation	Task analysis follow-up studies	3 or more	Ecological validity; reveals users' real tasks. Suggests functions and features.	Appointments hard to set up. No experimenter control.

Recruiting test persons

Questionnaires	Task analysis, follow-up studies	At least 30	Finds subjective user preferences. Easy to repeat.	Pilot work needed (to prevent misunderstandings).
Interviews	Task analysis	5	Flexible, in-depth attitude and experience probing.	Time consuming. Hard to analyze and compare.
Focus groups	Task analysis, user involvement	6-9 per group	Spontaneous reactions and group dynamics.	Hard to analyze. Low validity
Logging actual use	Final testing, follow-up studies	At least 20	Finds highly used (or unused) features. Can run continuously.	Analysis programs needed for huge mass of data. Violation of users' privacy.
User feedback	Follow-up studies	Hundreds	Tracks changes in user requirements and views.	Special organization needed to handle replies.

Preparing the Test

Before the test:

- A script must be prepared describing the purpose to the test person, how long the test is supposed to take, the equipment, etc, so s/he knows what is expected of him/her
 - *It is very important that all test persons gets exactly the same instructions*
- Most often a *pre-test questionnaire* or *interview* is carried out to record e.g. demographic information from the test person
- In many cases, the test persons are required to carry out some predefined tasks with the system.
 - These must be carefully described (on paper) so the test persons can refer to them during the test
 - Often, the *order* the users carry out the tasks must be varied to avoid bias from training effects
- In other cases, e.g. audio or video recordings or other software or materials must be prepared and ready

Preparing the Test

The most important thing is to be well-prepared. It can be very time-consuming to do user test, and they can't easily be redone, if some mistake is made. Therefore:

Test the test and all materials thoroughly **before** the actual test e.g. with friends or close colleagues

The Test Team

The Test Team/Tasks:

- Test Monitor / Experimenter
 - Overall responsible, direct contact with test participants
- Data Logger
 - Logs important events, actions, keeps track of timing and other tasks such as video Recordings
- Other: Tech. Expert, Test Observers

It is very important that only the persons essential to running the test are present (or visible) to the test person.

The test monitor/experimenter

The most important person:

During the test, the monitor is responsible for all aspects of conducting the test, including greeting the participant, collecting data, assisting and probing, and debriefing the participant

After the test, collate the information, debrief the other team members, note results, etc.

The test monitor/experimenter

Who?

NOT the designer

- it is very hard to remain objective when testing your own design/product
- e.g. you might explain away user problems rather than acknowledging them as real issues

This is a tough requirement! - Especially here at the university it is nearly always the designer who tests

The test monitor/experimenter

Test Monitor Skills

- Good Memory
 - need to remember events, comments made earlier during test
- Good at concentrating
 - need to be observing test subjects for maybe one or two hours throughout the day
- Good organiser/coordinator
- Flexible

The test monitor/experimenter

Emphatic / Good communicating skills

- must establish a relationship “make friends” with test participants

Good Listener

- Must understand the content and implications of the participants comments - helps understand the underlying rationale behind the participants actions

Test Monitor Pitfalls

1. Leading rather than enabling
2. Too involved in data collection/timing/logging tasks
3. Acting too knowledgeable
4. Not Flexible
5. Not relating well to test participants
6. Jumping to conclusions / Impatient

Other Persons/Functions

Data logger (might actually be more than one person):

- “Logs data” - ie. responsible for acquiring the data required for the test.
 - Timing events
 - Noting special events (e.g. errors, user comments/expressions, etc.)
- When planning the test, it is good idea to make a table and assign codes to expected activities, to make sure that these actions are easily identifiable during the test (and can be noted down quickly)
- Operates the video recorder, responsible for getting everything documented (including sound). Often more cameras.
- Must plan recordings carefully before test (trials)

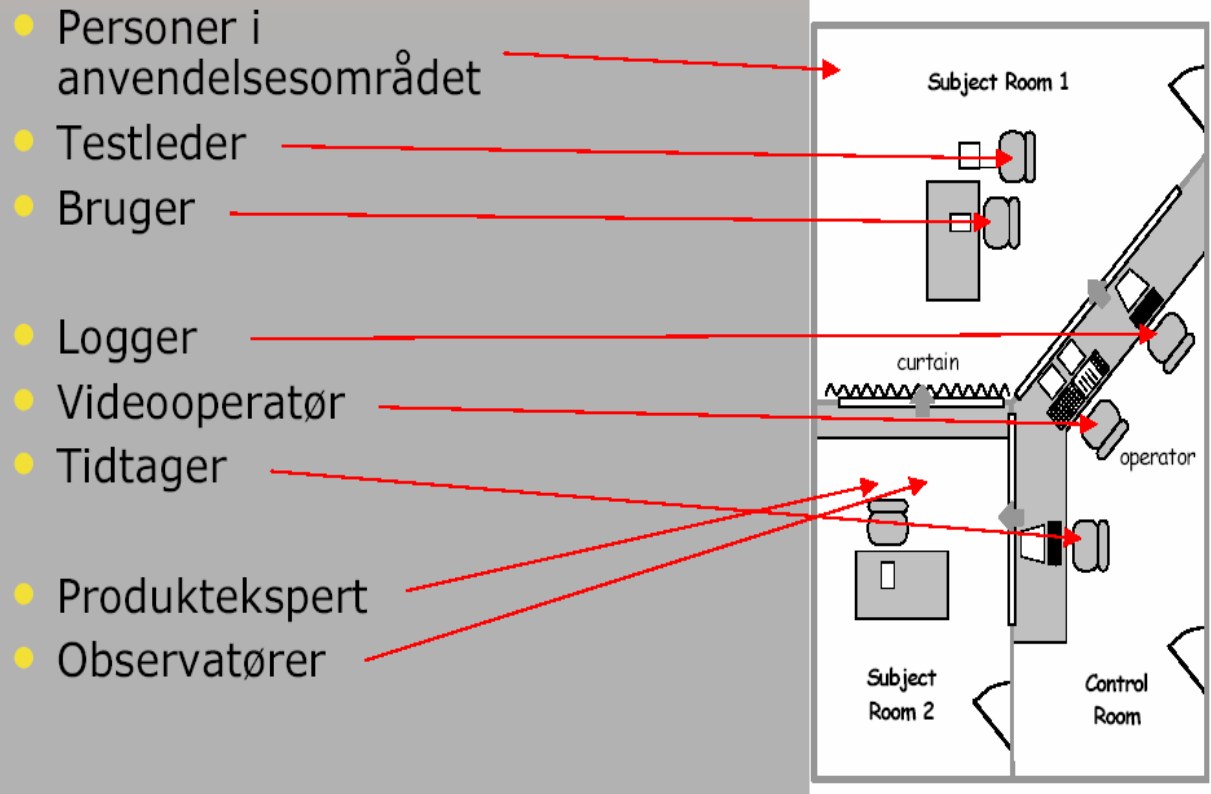
Tools

The most important tools available for recording and documenting the tests are:

- Usability test lab or other controlled facility
- Video Recording Equipment (at least for backup or reference)
- Pencil and Paper, stopwatch (scripts, forms, questionnaires)
- Software (automated logging, tracking keyboard, mouse, screen, etc)

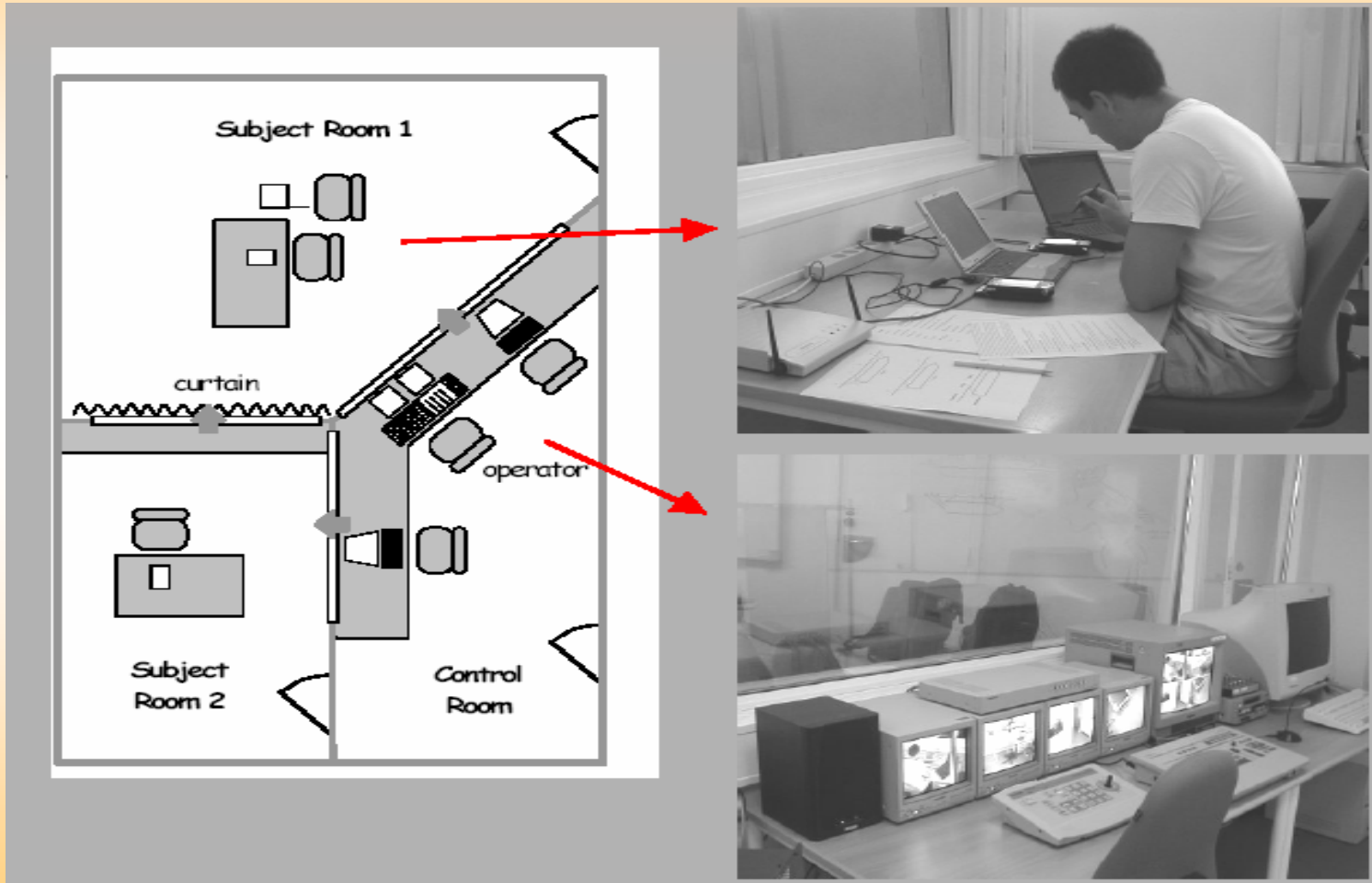
Classical Usability Lab

A usability lab is divided into a number of rooms, as a minimum a control and a test room

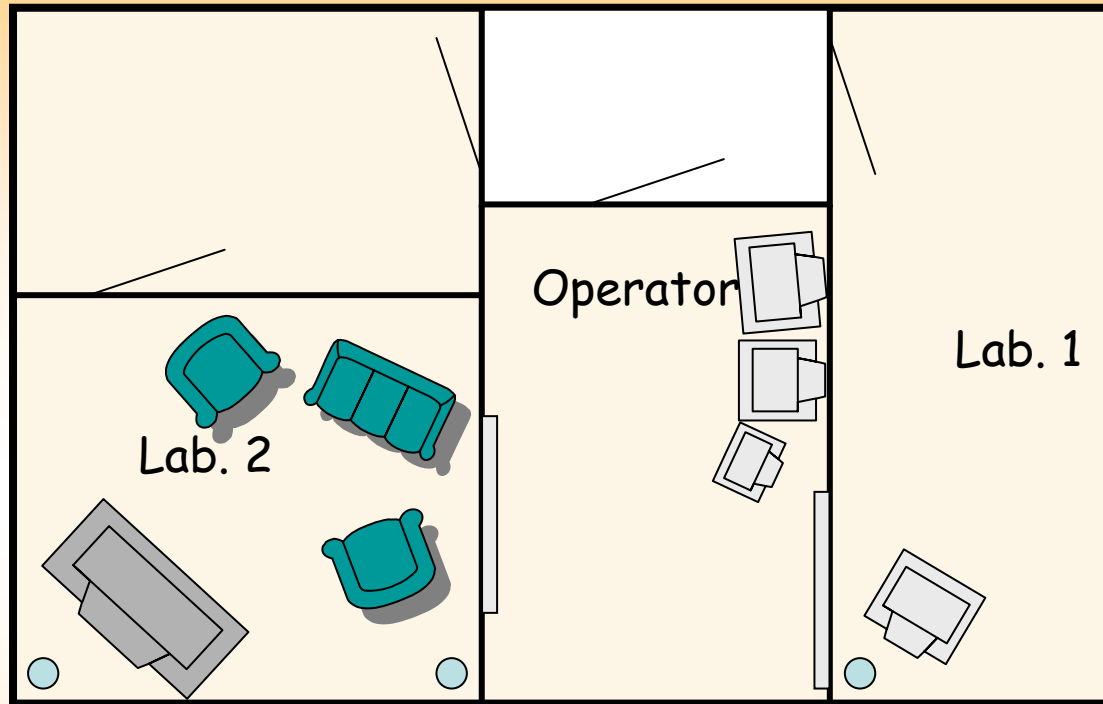


The figure shows the usability lab at the CS dept. (slides and pictures courtesy of Jan Stage, cs-dept).

CS dept. Usability Lab



InDiMedia usability lab at NJv14



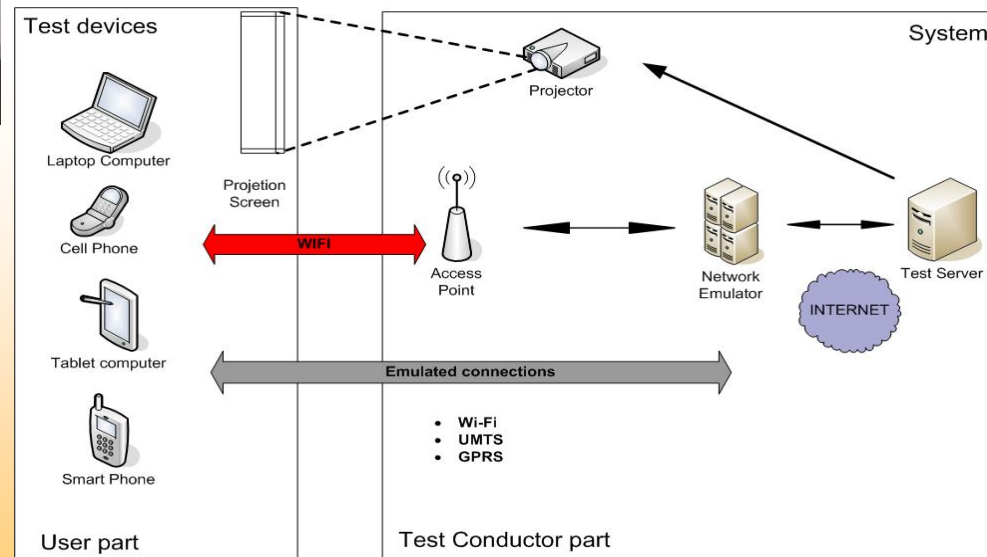
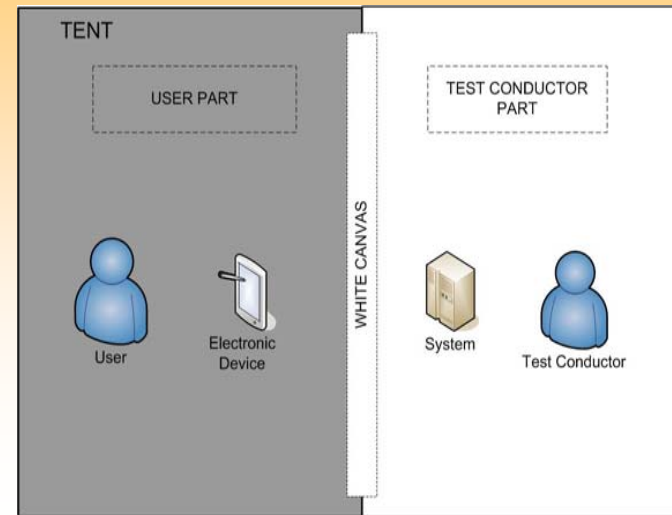
One test room simulates an ordinary living room, the other test room has an office-like setup

POSH – Usability Test Facility - Physical surroundings



“The tent”: The purpose is to enable user tests in a simulated, mobile environment.

- Two parts inside.
- Black canvas.
- White separation canvas for the projection.



Screen Capture Tool

Camtasia:

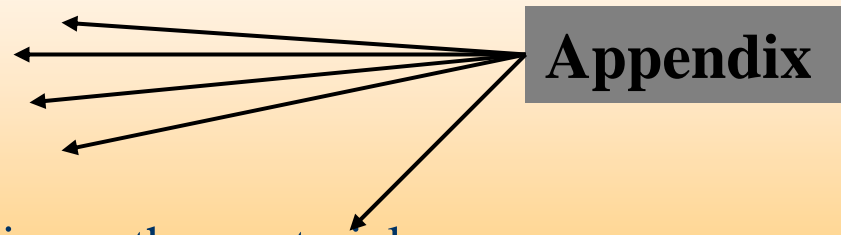
- Excellent tool!
- Captures (PC) screen and records all mouse events
- Can use the microphone to simultaneously capture users' speech

<http://www.techsmith.com/download/trials.asp>
(30 days free trial version)

Documentation – the Test Report

Test report sections:

- (Executive) summary
- Background
- Goals
- Methodology
- Platform and test bed
- Facility
- Test persons
- Analysis
- Recommendations/conclusions
- User data
- Product images
- Test user screener
- Transcripts of interviews
- Instruction sheet, questionnaires, other materials



Compile and Summarize Data

It is important to start compiling data during the experiments

- *While information is still fresh*

An important aspect of compiling the data is to put it into a form that allows you to see *patterns*

The most obvious patterns are *summaries*

Performance data:

- Mean and standard deviations, confidence intervals
- (Median and range)
- E.G. For completion times and task accuracy

Data Collection

Performance Data

- Time to: complete tasks, achieve competence, training time, error recovery, etc
- Counts and rates (# errors, use of help, steps to complete task, comprehension test)

Preference Data

- Prefers A over B
- Suggestions for improvements
- rationales for performance
- ranking/ratings

Performance Data

Online collection

- generate timestamped logfiles (e.g. keystrokes, mouse clicks, etc)
- online questionnaire

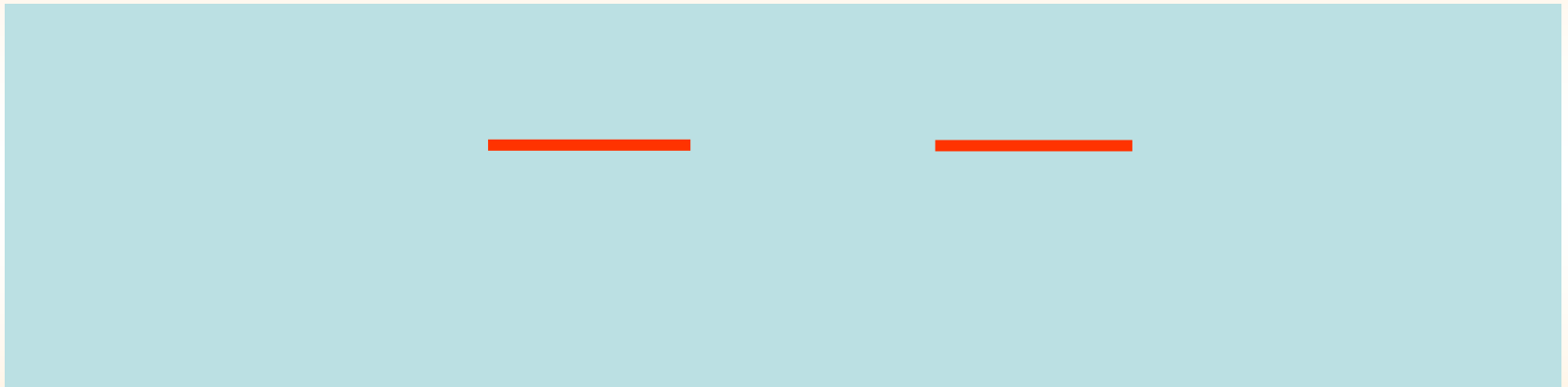
Manual collection

- schemas and tables
- video analysis

Example of performance data report

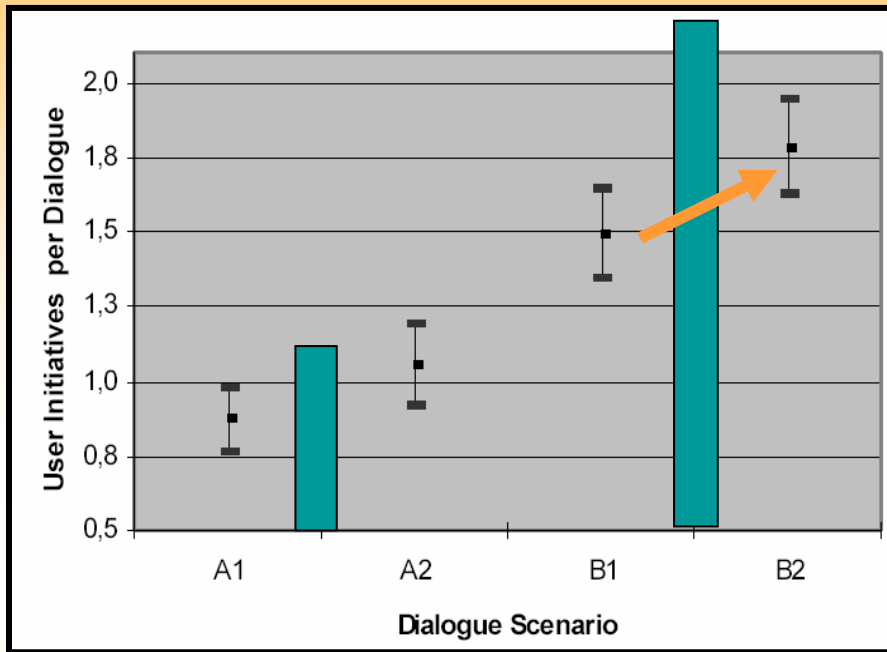
All users were required to carry out two scenarios, A and B

The table shows the average time spent in the login subtasks for the first (A1,B1) and second (A2,B2) dialogues



A paired, two tailed t-test revealed a significant reduction of the time spent in the “Id_number” sub task when comparing the first to the second dialogue ($p = 0.03$)

Example of performance data report



Average number of user initiatives per dialogue for the "A" and "B" scenarios, for the first and second dialogues.

An unpaired two-tailed t-test shows a significant ($p = 0.02$) increase in the number of user initiatives relative to the total number of turns for scenario B2 compared to B1.

User Preference Data

The only way to find out the user's preferences, experience, etc. is to *ask* them

- This is done via (structured) *interviews* or, most often by filling out a *questionnaire*:

Types of questions:

- Limited-choice questions:
 - Sum answers to each individual question shows how many participants selected each choice
- Free-form questions and comments:
 - List questions and group answers into suitable categories

Questionnaires

Screening Questionnaire

- Used for selecting participants

Background (Pre-test) Questionnaire

- Used to confirm and elaborate info from the screening prior to the test
- Collect demographics

Post-test Questionnaire

- Used to acquire preference information from the participants
- Focus is on opinions and feelings

Alternative/supplementary: Debriefing Interviews

Post-test Questionnaire

Use the Problem Statements as a base for the questions

- questions must be brief, concrete and precise
- questions must relate to what is not directly observable
- minimize questions needing complex or elaborate responses, instead rely on close-ended questions, such as:
 - check boxes
 - Short fill-ins
 - true-false
 - Likert statements
- Conduct a pilot test to verify the questionnaire

Closed-ended questions

Likert Scales:

Overall, I found the widget easy to use.

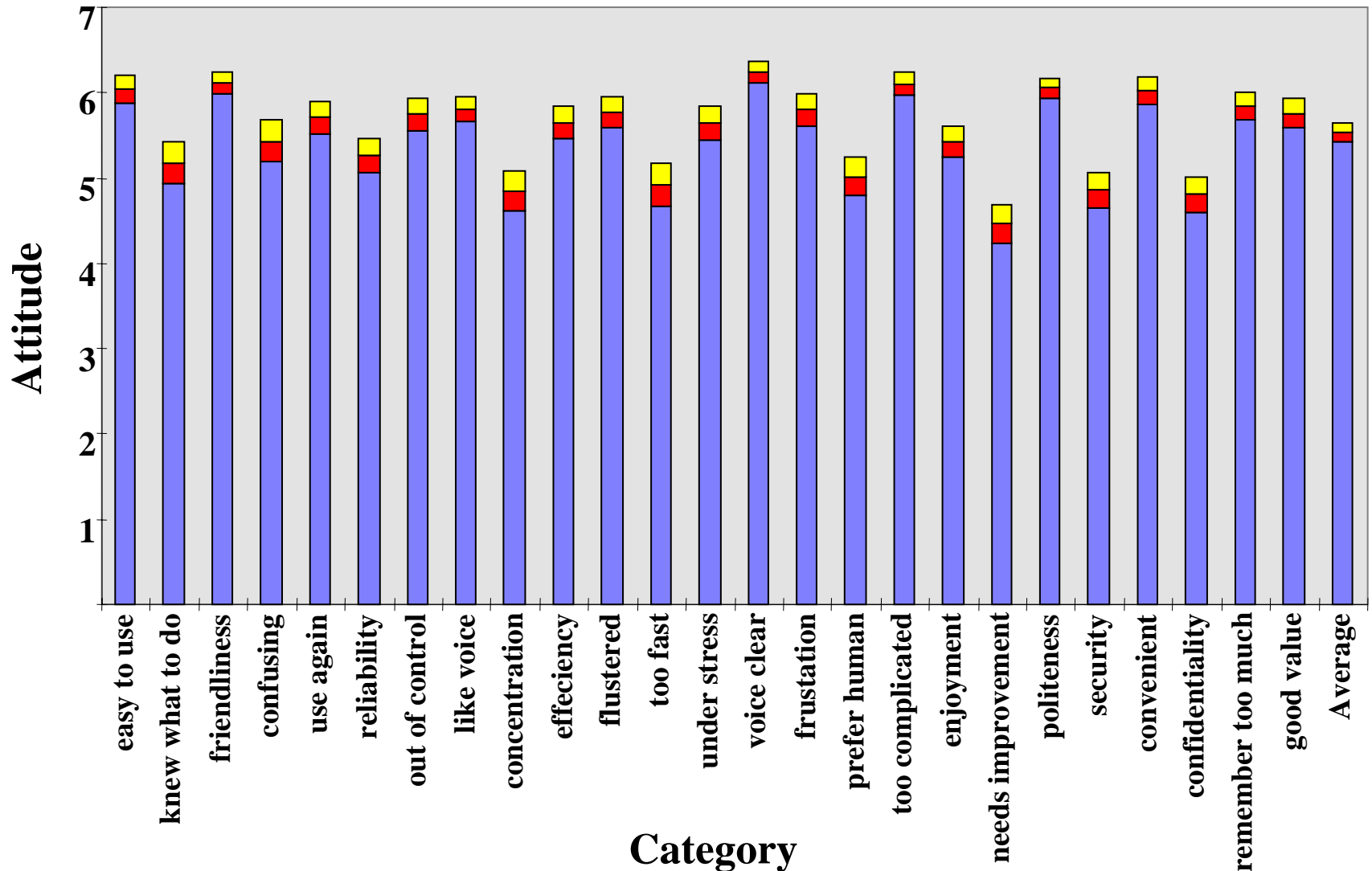
___Strongly Disagree ___Disagree ___Neutral ___Agree ___Strongly agree

Semantic Differentials

Simple	3	2	1	0	1	2	3	Complex
Hi-tech	3	2	1	0	1	2	3	Lo-Tech
Reliable	3	2	1	0	1	2	3	Unreliable
Durable	3	2	1	0	1	2	3	Breakable

Example of Questionnaire with 25 Statements

Average User Attitudes with 98% confidence intervals



Some References and Readings

Questionnaires:

Check out Gary Perlman's online questionnaires at:

<http://www.acm.org/~perlman/question.html>

Jurek Kirakowski's Usability questionnaire faq:

<http://www.ucc.ie/hfrg/resources/qfaq1.html>

The DUE – online dynamic usability evaluation questionnaire:

<http://cpk.auc.dk/education/IMM/DUE/>

Inferential Statistics

Used to obtain *proof* of statistically *significant* results.

Most often used in *comparative* studies

Problems:

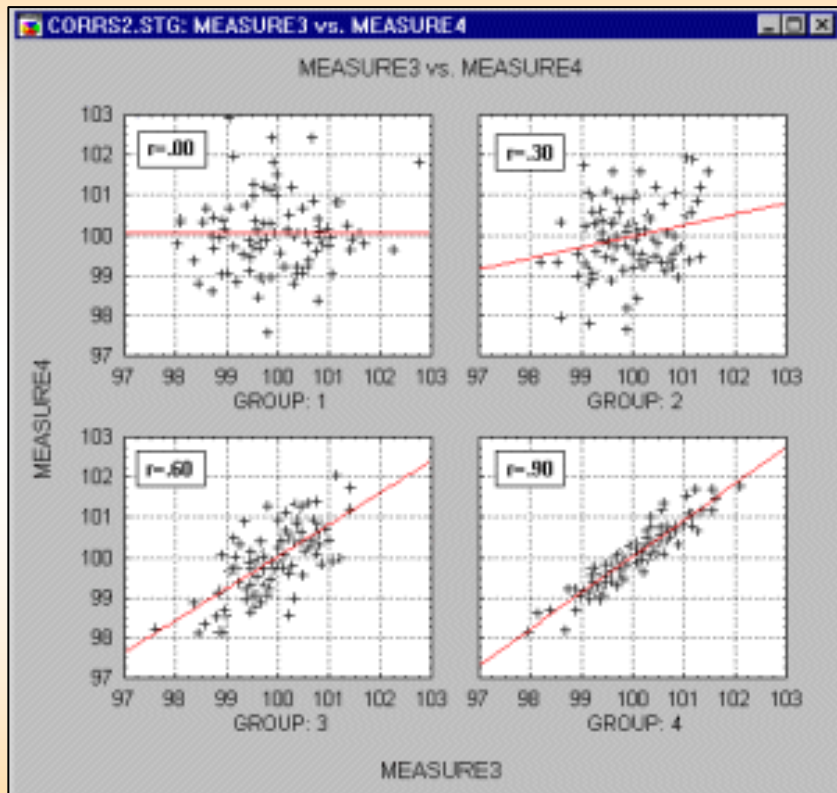
- Difficult to choose appropriate statistical method
- Needs many test (>10-12) participants per condition (random variable)
- Demands rigorous control of experimental conditions

Benefits:

- Provides a solid scientific evidence of the test
- Provides benchmarks and a basis for comparative tests

Correlation

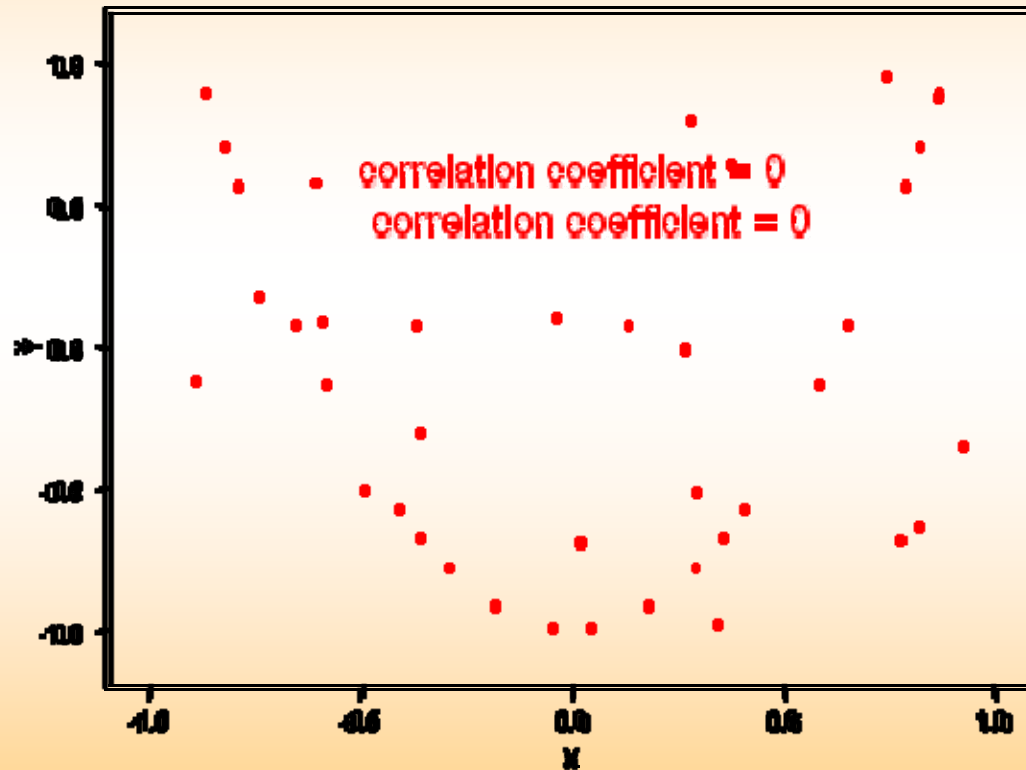
The simplest inferential measure is correlation:



- In addition to the correlation coefficient (R^2) it is often desirable to perform a linear regression (or a multiple linear regression (MLR) analysis).
- Note that correlation and regression say nothing about the causality of the relationships!!

Correlation

Beware: Correlation only measures the strength of a *linear* relationship between two variables.!!



Statistical Significance

What is "statistical significance" (p-value)?

- The statistical significance of a result is an estimated measure of the degree to which it is "true" (in the sense of "representative of the population").

The higher the p-value, the less we can believe that the observed relation between variables in the sample is a reliable indicator of the relation between the respective variables in the population.

In other words, the probability that the observed difference is due to pure chance in the sample, and not an actual difference in the entire population

Statistical Significance

Example:

- A p-value of .05 indicates that there is a 95% probability that the relation between the variables found in our sample corresponds to the values of the population.

Conventions:

- $P < 0.05$ is considered marginally significant
- $P < 0.01$ is considered significant
- $P < 0.005 - 0.001$ is considered highly significant

Analysis of Variance (ANOVA)

ANOVA is one of the most common methods to analyse the relationship between variables

- The purpose of ANOVA is to test for significant differences between means (for groups or variables) for statistical significance.
- In the case of only two groups, ANOVA becomes identical to the t-test

Recommendations and Conclusion

Be well prepared!!

- Identify clear goals
- Decide which kind of test you want or need and how many resources you have available
- Prepare test materials and “test the test”
- Often costly or impossible to redo test

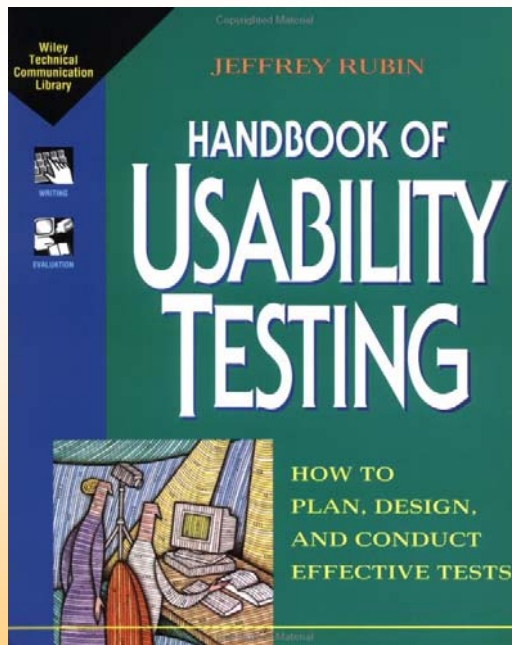
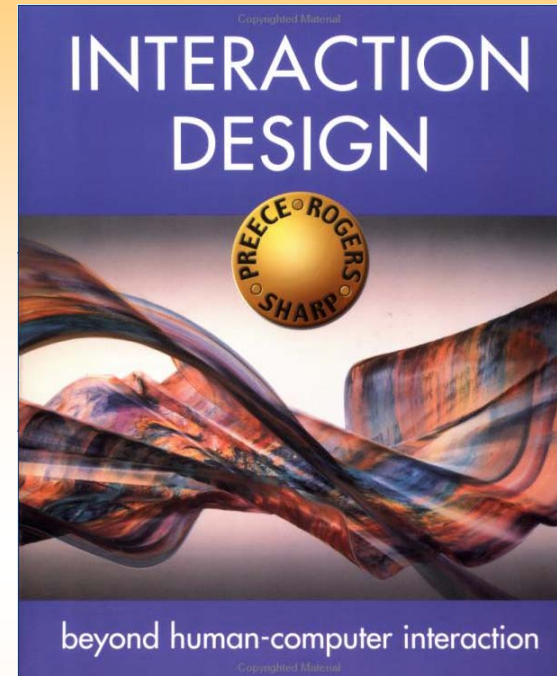
Remember, we are dealing with humans:

- There are ethical (maybe even legal) considerations
- Make sure your test users are comfortable and know what is expected from them
- Users don't make mistakes – errors are *never* their fault!

Recommended books on user testing

Books:

Jenny Preece et al: "Interaction Design", Wiley 2004
(see chapter on usability testing)
[http:// id-book.com](http://id-book.com)



Jeffrey Rubin: "Handbook of Usability Testing -
how to plan, design and conduct effective tests"
Wiley, 1994, USA.

Courses

I teach two courses on HCI and User testing on the 8th and 9th semesters of the Intelligent MultiMedia (IMM) masters' programme:

Spring Semester: 2 ECTS “Human Computer Interaction” course:

<http://www.hst.aau.dk/~ska/hci06/index.htm>

Fall Semester: 1 ECTS “Design and Evaluation of Usability Experiments” course:

http://cpk.auc.dk/education/IMM/s9-06/usability_course/index.html

These slides are available at:

http://kom.aau.dk/~lbl/ieee_sb_user_test.pdf