Evaluation through user participation

Evaluation Techniques

Evaluation through user participation

- Think aloud
- Cooperative evaluation
- Controlled experiment

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
  - act of describing may alter task performance

Protocol Analisys

- 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

Before Running a Think Aloud Session

- Develop a prototype
- Develop tasks that represent typical user goals
- Schedule sessions with users that match your Personas
- Organize yourself – get video camera, batteries, audio camera, tapes, pens, etc.

While running a think aloud session

Explain to the user:

- .. who you are & what you are doing
- .. that you are testing your interface, and not testing them
- .. that they can quit at any time
- .. that you won't be able to help them
- .. that you require them to continue talking, and you will remind them to "please keep talking" if they fall silent

- To simply verbalize what it is they are doing
- Verify that the user understands the tasks (have them read the tasks aloud too, and ask if there are any questions)

While the session is running

- Take good notes! Don’t rely on your video or audio tape
- If the user falls silent for more than three seconds,prompt them “please keep talking”
- Do not help the user complete a task (if the user asks for help, explain that you cannot help, and prompt them to try what they think is correct)
- Don’t defend your designs! This is not a critique of your design skills; don’t even mention that they are your designs
- Watch for signs of frustration; recommend a break if you notice the user getting particularly upset
- Remember that the user can quit at any time

After the session

Analyzing and presenting the Findings

#	Related incidents List the #s of any incidents that are related	Priority of the incident 1 = highest priority (huge usability flaw) 4 = lowest priority (minor usability flaw)	Description of the incident A summary of the incident; include quotes from the user	How the incident was found Detail the steps the user took to create this incident	Good or Bad	Potential solution to the incident, if Bad Hypothesize several solutions to the problem
E X A M P L E	none	1	User could not log in after trying four or five different things: "*Well, I really just don't see any way to log in; I give up. I feel so stupid.*" (User did not notice the log in icon)	On page #12B, User clicked on the image of a computer, but that took them to the statistics area of the site; they tried logging in to the administrative section, but didn't see the icon for regular-user login	Bad	Change the icon to a word or phrase ("Click here to login") or simply move the log in information to the first page

Example

The site: https://ec.europa.eu/info/index_en

The task: find a job posting for IT officer on https://ec.europa.eu/info/index_en

Remember

Explain to the users:

- who you are & what you are doing
- that you are testing your interface, and not testing them
- that they can quit at any time
- that you won’t be able to help them
- that you require them to continue talking, and you will remind them to “please keep talking” if they fall silent

Video

After the session

Analyze and present the Findings

Cooperative evaluation

Cooperative evaluation is a variation of think aloud in which the user is encouraged to see himself as a collaborator in the evaluation and not simply as an experimental participant.

In this case the user can ask the evaluator for clarification if a problem arises

Post-task walkthroughs

Transcript played back to participant for comment

Example

During the think aloud the participant may say ‘and now I’m selecting the undo menu’, but not tell us what was wrong to make undo necessary.

A post-talk walkthrough attempts to alleviate these problems, by reflecting the participants’ actions back to them after the event.

Evaluation through user participation

Controlled experiment

Controlled experiment

A controlled experiment is an **experiment** in which **all the variable factors** in an experimental group and a comparison control group **are kept the same except for one variable factor** in the experimental group that is changed or altered*

*https://www.merriam-webster.com/dictionary/controlled%20experiment

Example – Icon design

Problem

Two styles of icon design (naturalistic vs abstract images): which icon style is easier to remember?

Exercise: Propose a solution

Example – Icon design

- ✓ Two alternative systems (using natural and abstract)
- ✓ …

Example – Icon design

- ✓ Two alternative systems (using natural and abstract)
- ✓ Groups
- ✓ Tasks
- ✓ Measures
- ✓ …

Example – Icon design

- **Systems** -> Two interface composed of blocks of icons (natural vs abstract)
- **Groups** -> Two groups
- **Task** -> User task: “*delete a document*” using the appropriate icon (set of presentations)
- **Measures** -> For each user the time or number of mistakes

Example – Icon design

	A	B
1	Interface Style 1	Interface Style 2
2	59	67
3	61	62
4	57	64
5	67	63
6	59	72
7	55	68
8	64	70
9		
10		
11		

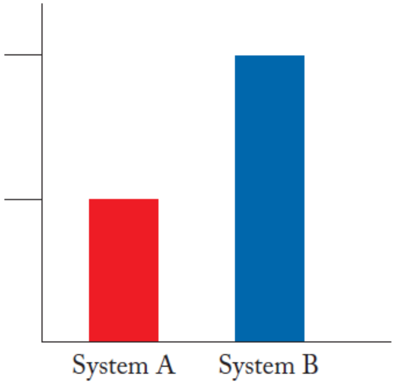
Example – Icon design

- ✓ Two alternative systems (using natural and abstract)
- ✓ Groups
- ✓ Tasks
- ✓ Measures

-> **Statistics**

Why we need statistics

Suppose you have performed a survey comparing two alternative systems and asked users which system they prefer



User preferences comparing two systems.

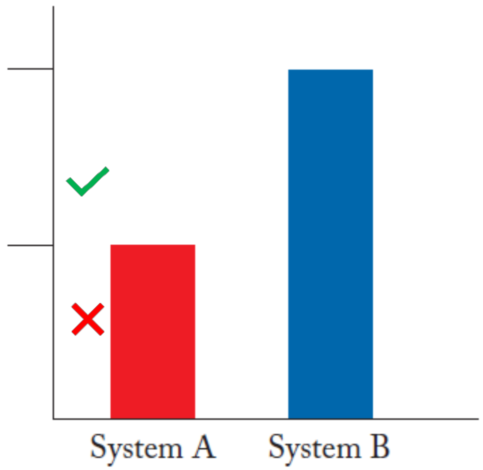
Is System B better than System A?

Why we need statistics

The left-hand scale has two notches but no values.

What if:

1. the notches are at 1000 and 2000
   the results of surveying 3000 people.
2. the notches are at 1 and 2
   the results of surveying 3 people.
3. the notches are at 10 and 20
   the results of surveying 30 people.



User preferences comparing two systems.

The job of statistics is to help with judgements such as these.

Statistics

- Statistics is a collection of methods which help us to describe, summarize, interpret and analyse data

- In other words, statistics is about trying to learn dependable things about the real world based on measurements of it

Controlled experiment

- The evaluator chooses a hypothesis to test
- Some attributes of user behaviour are measured
- A number of experimental conditions are considered differing only in the values of certain controlled variables
- Any changes in the behavioural measures are attributed to the different conditions
- Subjects, variables & hypotheses should be carefully considered

Participantes

- They should match the expected user population
- Ideally, subjects should be the real users
- If not:
  - Similar experience with computers
  - Similar knowledge of task domain
- Sample size chosen
  - Large enough to be representative
  - Minimum of 10 subjects

Variables

- Experiments consider variables under controlled conditions:
  1. Manipulated – independent variables
  2. Measured – dependent variables

Examples
Interface style, level of help, n° of menu items, icon design
Different values may be given, each value is a level of the variable
Checking a menu list, means to measure search speed-up for 10, 8 or 6 items within the menu

Hypotheses

- The hypothesis is a prediction of the outcome of the experiment
- The experiment aims to prove the hypothesis disproving the null hypothesis (no difference in the dependent variables caused by changes in the independent variables)
- Statistical analysis provides

Experimental design

- Firstly one must choose the hypothesis: what one is trying to demonstrate
- Clarification between indep. & dep. variables
- How many participants are available?
- Choice of experimental method to be used:
  - Between-groups

    each subject performs experiment under each condition
  - Within-groups

    each subject performs under only one condition

Experimental design

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

Example – Icon design

Problem

Two styles of icon design (naturalistic vs abstract images): which icon style is easier to remember?

Icon design: Step 1

Hypothesis

Users will remember the natural icons more easily than the abstract ones

The null hypothesis is that there will be no difference (in how users will remember icons) between natural ad abstract icon style

Icon design: Step 2

Independent variables

The style of icon. We have two alternatives: natural and abstract

Dependent varialbles

How can we measure the idea of “remember more easily”?

We will assume to measure:

- the number of mistakes in selection and
- the time to select the icon

Icon design: Step 3

Experimental method

We chose within-subject (each user performs under each different condition)

To reduce the learning effect we addressed order (half the subjects A–B and half B–A). Icons were randomly placed in the blocks

Icon design: Step 4

Experiment details

- Two interface composed of blocks of icons (natural vs abstract)
- User task: "*delete a document*" using the appropriate icon (set of presentations)
- Random placing of icons in the block
- Each user performs the task under each condition
- Users in two groups with different starting condition
- For each user the time and the n. of mistakes

Icon design: Step 5

Analyze our results

…ANOVA

ANOVA

Analysis of variance, ANOVA, is a technique from statistics that allows us to deal with several populations

In its simplest form, ANOVA provides a statistical test of whether or not the means of several groups are equal, and therefore generalizes the t-test to more than two groups

The F-test

The F-test is used for comparing the factors of the total deviation. For example, in one-way ANOVA, statistical significance is tested for by comparing the F test statistic

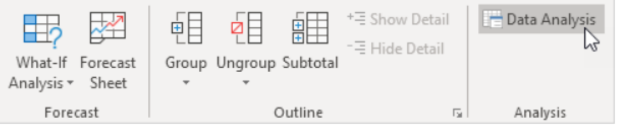
$$F = \frac{SS_B \,/(I-1)}{SS_W \,/[I(J-1)]}$$

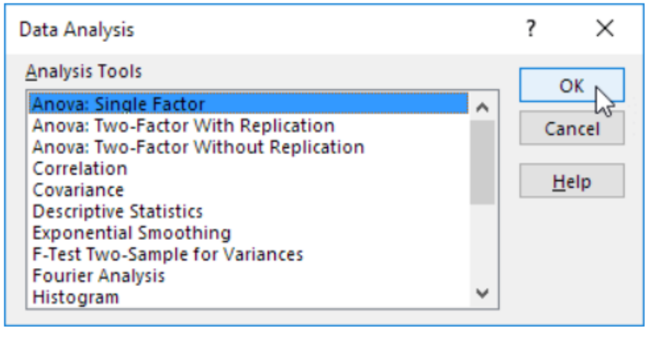
https://www.excel-easy.com/examples/anova.html

ANOVA in Excel

	A	B
1	Interface Style 1	Interface Style 2
2	59	67
3	61	62
4	57	64
5	67	63
6	59	72
7	55	68
8	64	70
9		
10		
11		

ANOVA in Excel





NOTE: if the Data Analysis command is not available in your version of Excel, you need to load the Analysis ToolPak add-in program.

ANOVA in Excel

	A	B	C	D	E	F	G	H
1	Interface Style 1	Interface Style 2						
2	59	67						
3	61	62						
4	57	64						
5	67	63						
6	59	72						
7	55	68						
8	64	70						
9								
10								
11								
12	**ANOVA - Single Factor**							
13	Alpha	0,05						
14								
15	Groups	Count	Sum	Mean	Variance			
16	Column 1	7	422	60,2857142857143	16,9047619047619			
17	Column 2	7	466	66,5714285714286	13,952380952381			
18								
19	Source of Variation	SS	df	MS	F	P-value	F critical	
20	Between Groups	138,285714285732	1	138,285714285732	8,96296296296411	0,01119451	4,74722535	
21	Within Groups	185,142857142857	12	15,4285714285714				
22	Total	323,428571428571	13					
23								
24								

If F > F crit, we reject the null hypothesis. This is the case, 8.962962 > 4,74722. Therefore, we reject the null hypothesis

Probing the unknown

Probing the unknown

Statistics aims to probe something in the world you don't know. In order to do this you can use:

- Hypothesis testing (ANOVA)
- Confidence intervals
- Bayesian statistics

The first two use essentially the same theoretical approach

Types of statistics

- **Hypothesis testing**
- Confidence intervals
- Bayesian statistics

Traditional statistics

Hypothesis testing

Analyze experimental results considering the p-value

P-value represents the likelihood that the data we have measured, or the results we have obtained, could have arisen by chance

Hypothesis testing

It uses various methods and measures to come up with the common $p < 5\%$ or $p < 1\%$ result.

Perhaps what does it mean this 5% (or 1%)?

What can and cannot be concluded from a non-significant result?

Hypothesis testing

A core term is null hypothesis, also written H_0, which is usually what you want to disprove

Example

H_0 might be that your new software design has made no difference to error rates.

In contrast, the alternative hypothesis, written H_1, is what you typically would like to be true

Hypothesis testing

The hypothesis testing reasoning goes like this:

- **if** the null hypothesis H_0 were true
  -> **then** the observations (measurements) are very unlikely
- **else if** the null hypothesis H0 is false we can use the collected data to verify if the alternative H1 is (probably) true

H_0 and the significance level

The smallest significance level that is normally regarded as reasonable evidence is 5%.

This means that if the likelihood of the null hypothesis is less than 5%, you 'reject' it and assume the alternative must be true.

NOTE: <u>this *does not mean* that there is a high probability that the alternative is true or false</u> but merely that the null hypothesis is unlikely to have given rise to the observed results

Example

You have a new design for a software

Our null hypothesis was that your new design shows no difference with the old one

Your hypothesis is that the new design is better than the old design

Example: case 1

Analyzing data you see that there is no statistically significant difference between the two (H_0 is true)

A possible explanation is that there is no difference between new and old design

But it may also <u>mean that your experiment was not good enough to detect the difference</u>

You can *never* (simply) reason: a non-significant result means H_0 is true -> H_1 is false

H_0 and the significance level

Law courts can return three verdicts: guilty, not guilty or not proven

In statistics ‘not significant’ is just the same: ‘not proven’

Example: case 2

Analyzing data you reject null hypothesis

Can we say that H_1 is true?

If your results are significant you can use collected data to decide if new design is better that old design

Focus on choosing participants

Representing the real world

- The sample
- The population

The sample

- The sample is the user you tested in a specific date/time under certain conditions
- Imagine the user made 3 errors and finished the task in 17 minutes and 23 second
- Would the same user on a different day, under different conditions have made the same errors?
- What about other users. How many samples do we need -> Population

The population

The population is a larger group of people you want to know about

Example

We may be interested in collecting information about those participating in HCI course

In some cases we don’t have the ‘real population’

Example

We may be interested in collecting information about those who will participate into new courses in the future

This hypothetical ‘real’ event may be represented mathematically as a theoretical distribution

There and back again

The job of statistics—> moving from data about the real world back to knowledge about the real world

- Given the complete past history of ten million users of a website, what does this tell us about their future behaviour or the behaviour of a new user of the site?

- Given the error rates of 20 people on an artificial task in a lab, what can you tell about the behaviour of a typical user in their everyday situation?

Focus on variables and measures

Noise and randomness

How random is the world?

We have a sample of heights of 20 randomly chosen people from an organisation

We can measure each of their heights relatively accurately, but maybe even this has some inaccuracy -> **noise**

They are randomly chosen from the far larger population of employees.

There is a degree of **randomness** in the measurements on which we base our decision making

How random is the world?

- The behaviour of random phenomena is often far more chaotic than we expect

- We are used to 'tame,' predictable phenomena in order to verify specific aspects but this may lead to misinterpreting data

- The mathematics of formal statistics attempts to see through this noise and give a clear view of robust properties of the underlying phenomenon (ANOVA)

Bias and variability

When you take a measurement two of the core things you need to know about are:

- Bias - is about systematic effects that skew your results in one way or another
  Are your results fair?
  If not -> We have a good estimate of the wrong thing

- Variability - how likely is it to be close to the real value
  Are your results reliable?
  If not -> a poor estimate of the right thing

Independence and non- independence

Independence is about whether one measurement or factor gives information about another

Non-independence may increase variability, lead to misattribution of effects, or even suggest a completely wrong effect

Kinds of independence

What can influence independence:

- Measurements
  - Order effects
  - Context or 'day' effects
  - Experimenter effects
- Factor effects
- Sample composition
  - Internal—subjects related to each other
  - External—subject choice related to topic

Independence of measurements

Kinds of independence

Kinds of independence:
- Measurements
  - Order effects
  - Context or ‘day’ effects
  - Experimenter effects

There are a number of ways in which measurements may be related to one another:

- **order effects**

Users see system A first, followed by system B

- **context or ‘day’ effects**

Bad weather often affects people’s moods

- **experimenter effects**

A contextual factor is you

Independence of factor effects

Kinds of independence

Kinds of independence:

- Factor effects

Is when there is some form of relationship or correlation between the various factors that you are measuring aspects of

Example

If you measure the death rate amongst patients in specialist hospitals it is often higher than in general hospitals

Can we say that patients do not get as good care in specialist hospitals?

Independence of sample composition

Kinds of independence

Kinds of independence:

- Sample composition
  - Internal—subjects related to each other
  - External—subject choice related to topic

Internal non-independence is when subjects are likely to be similar to one another but in no particular direction with regard to your question

Example

Snowball sample

Snowball samples

Kinds of independence

Kinds of independence:

- Sample composition
  - Internal—subjects related to each other
  - External—subject choice related to topic

Is when you have an initial set of contacts, often friends or colleagues, and ask them to suggest any of their own contacts who might take part in your survey

It is problematic for sampling political opinions, but may be acceptable for shoe size (except if you are dealing with basketball team)

External non-independence

Kinds of independence

Kinds of independence:

- Sample composition
  - Internal—subjects related to each other
  - External—subject choice related to topic

External non-independence is when the choice of subjects is actually connected with the topic being studied

Example

Doing a survey about preferences between MacOS and Windows (or iPhone and Android) in the Apple Store

..but also using a mobile app-based survey on a topic which is likely to be age related

Back to our example – Icon design

Problem

Two styles of icon design (naturalistic vs abstract images): which icon style is easier to remember?

Icon design: Step 3

Experimental method

We chose within-subject
Each user performs under each different condition

Sample composition

To reduce the learning effect we addressed order
Half the subjects A–B and half B–A).
Icons were randomly placed in the blocks

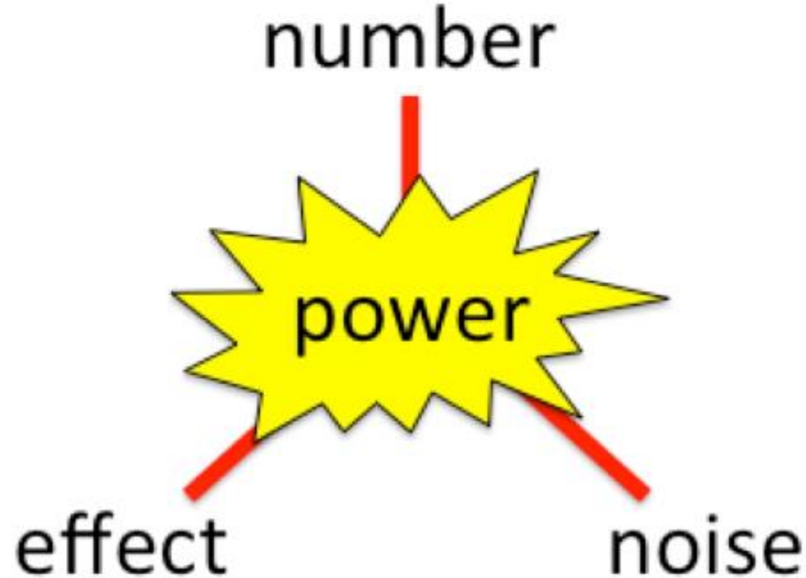
Order effects

Design and Interpretation

Statistical power

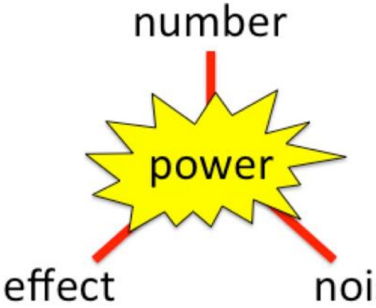
Power arises from a combination of:

- the size of the effect you are trying to detect
- the size of the study (number of trials/participants)
- the size of the 'noise' (the random or uncontrolled factors)



General strategy

- Increase number
  we often need a *very large increase in the number of subjects or trials* in order to reduce the variability of our results to an acceptable level
- Reduce noise
  Noise is about variation due to factors that you do not control or about which you have little information; we can attempt to attack either of these
- Increase effect size
  We can attempt to manipulate the sensitivity of our study.



Subjects

Subjects

- More subjects or trials (increase number)
- Within-subjects/within-groups studies (reduce noise)
- Matched users (reduce noise)
- Targeted user group (increase effect)

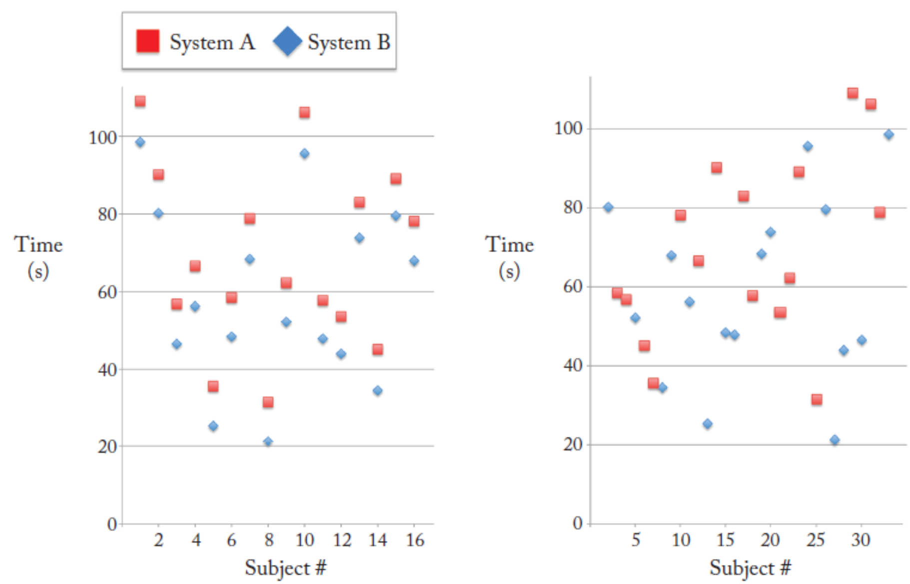
Within subjects and between subjects

- Within-subjects experiment
  Each subject perform for all conditions
- Between-subjects experiment
  Each subject perform for a single condition

Within subjects and between subjects

Example

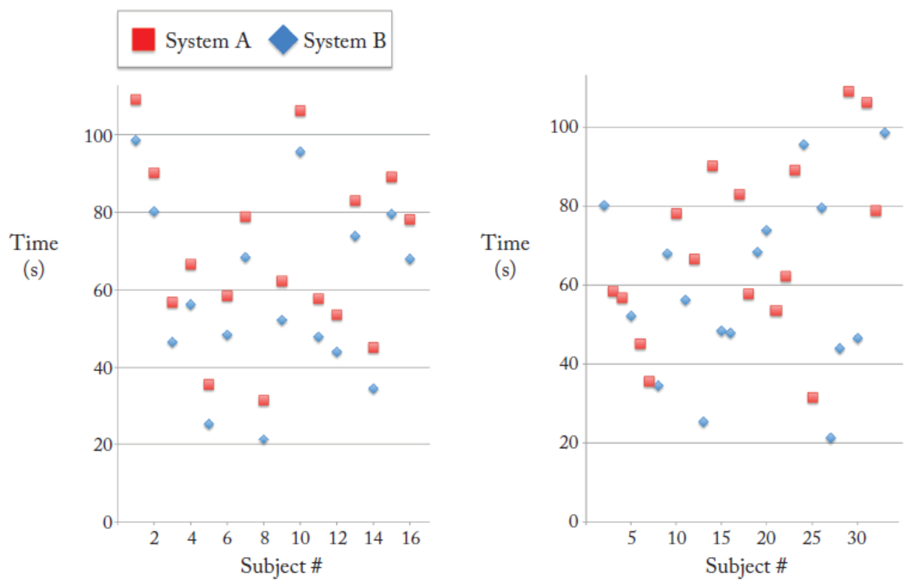
Imagine you are comparing two different experimental systems A and B, and have recorded users’ task completion time for each.



(i) Within-Subjects (ii) Between-Subjects

Within subjects and between subjects

- In the left-hand graph system A is always slower than system B
- In the right-hand graph is hard to tell the difference between the conditions: they are masked by the large differences between individuals



(i) Within-Subjects (ii) Between-Subjects

Within subjects and between subjects

- In the case of within subjects designs the main problem is order effects

- The normal way to address order effect is to randomise or balance the orders (half the subjects A–B and half B–A)

Within subjects and between subjects

- If we could clone users between subjects will be a good solution -> no learning effects and the same (cloned) user between conditions (but no sample increase)

- To emulate this we can pair subjects who are very similar, say in terms of gender, age, or skills, and allocate one from each pair to each condition (Matched users)

Targeted user group

We can go one step further and deliberately choose a group for whom we believe we will see the maximum **effect**

For example

You have designed a new menu system, which

you believe has a lower short-term memory requirement

If you test it on university students you may not see any difference

If you chose more elderly users you would be more likely to see differences

Task

Task

As well as choosing *whom*, we can manipulate *what* we ask them to do.

- Distractor tasks (increase effect)
- Targeted tasks (increase effect)
- Demonic interventions (increase effect)
- Restricted tasks (reduce noise)

Distractor tasks (increase effect)

A distractor task is an additional task which has the aim of saturating some aspect of the user's cognitive abilities, so that differences in load of the systems or conditions being studied become apparent

Example

In mobile interface design when users are tested using an interface whilst walking and avoiding obstacles

Targeted tasks (increase effect)

We choose targeted tasks that deliberately expose the effects of our interventions

Example

If you have modified a word-processor to improve the menu layout and structure, it makes sense to have a task that involves a lot of complex menu navigation rather than simply typing

Demonic interventions (increase effect)

In the extreme one can deliberately produce tasks that are plain nasty

Of course, creating extreme situations means there are problems of generalisation

Restricted tasks (reduce noise)

- The more control one has over the study, the less uncontrolled variation there is and hence the noise is smaller
- In a fully in-the-wild setting people may be affected by other people around them, weather, traffic, etc.
- However, one can still exercise a degree of control, even when conducting research in the wild

One way is to use restricted tasks

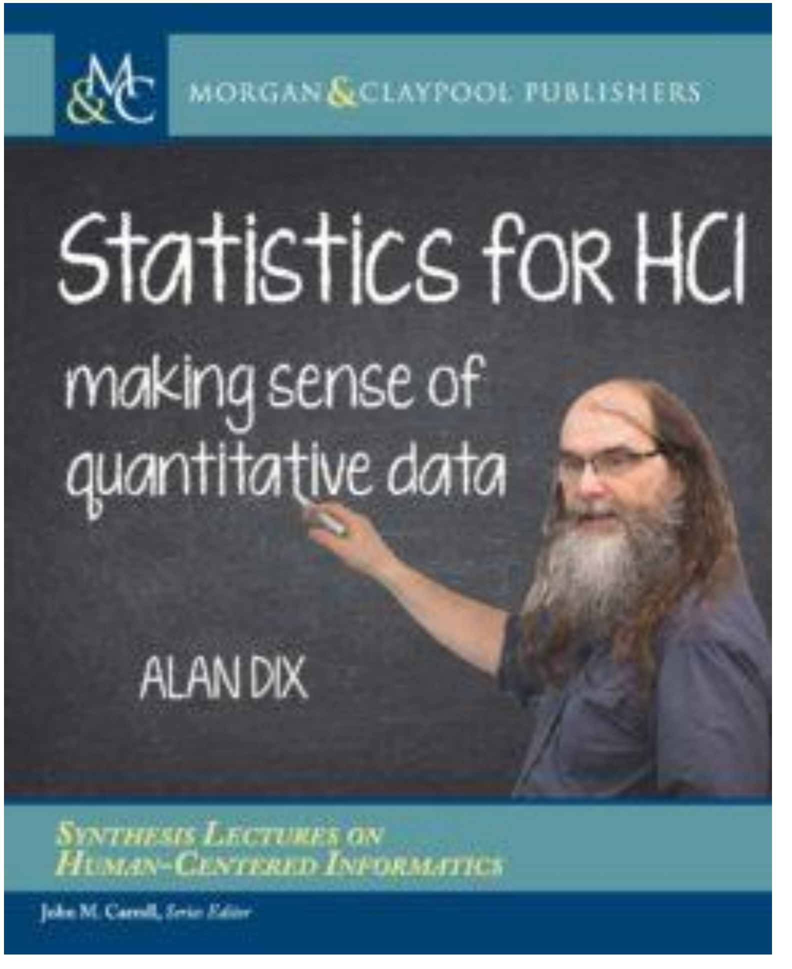
Restricted tasks (reduce noise)

- Participants are in a real situation but you give them a scripted task to perform

- Another approach is use a restricted device or system
  Example, you might lock a mobile phone so that it can only use the app being tested.

Reference

Statistics for HCI: Making Sense of Quantitative Data
Morgan & Claypool, April 2020, 181 pages



https://alandix.com/statistics/book/