# Usability Testing

## Introduction

We can try to make software and web-sites usable by adhering to guidelines, testing against metrics, formal modelling, etc..

However, the only way to find out if we have succeeded is to test the software or web-site on potential users.

Testing can be carried out at various stages during design and development, e.g.:

- At a preliminary stage, to determine requirements, expectations, etc.
- During design, as a means of testing general concepts or individual elements of a proposed system.
- At the prototype stage, to find out if the design meets expectations
- etc.

The best approach is *iterative* testing, i.e., testing at each stage of the design and development cycle.

## Designing Controlled Studies

For example, suppose that we have designed an interactive training system.

We believe this system will offer improved learning outcomes compared with existing systems, and we wish to discover if this is true.

Therefore, we plan to conduct a controlled study in which we compare our system with another system.

In order to carry out a controlled study we need:

- Two (or more) *conditions* to compare, e.g. task performance on two different systems.
- A *task* which can be performed on both systems
- A *prediction* that can be tested
- A set of *variables*, including an *independent variable* and one or more *dependent variables*.
- A number of *subjects* who may need to be divided into *groups*
- An *experimental procedure*

## Conditions

A common approach in Usability Engineering is to compare a new system with an existing system. Thus there are two conditions under which subjects are tested:

- 1. performing a task on the new system.
- 2. performing the same task, under the same conditions, on an existing system.

The existing system might be:

- an earlier, comparable system (e.g., for in-house development)
- a recognised standard, if one exists
- a well-known system, so that others have a reference against which they can judge our findings.

If we conduct a controlled study in which we compare a new system against a reference system, we use the following terminology:

- The condition in which the new system is used is known as the *experimental condition*.
- The condition in which the reference system is used is known as the *control condition*.

If there are more than two conditions in a study, or if the conditions have equal status, we may refer to them simply as *condition 1*, *condition 2*, etc.

In the interactive training system example, we are comparing two conditions, one of which represents a kind of standard.

Therefore we will have:

- A *control condition* in which the comparison system is used, and
- An *experimental condition* in which our system is used.

# Task

The task must:

- Be relevant to the systems being compared.
- Be testable against some criterion
- Not place some of the subjects at an unfair advantage/disadvantage

For example, when comparing interactive training systems, the task might be to learn about a topic using either the *experimental* or *control* system and then take a test.

The task should also be chosen so as not to place one of the systems at an unfair advantage/disadvantage

However, there are exceptions to this.

For example, our system might incorporate a feature that we believe will improve learning of a particular type of material.

In this case we would want to include the relevant type of knowledge in the task, even though we expect this to favour our system over the comparison system.

The task should not favour either system in any other respect.

# Prediction

The prediction must be framed so that it is *testable*.

Our prediction might be:

- *Subjects who use the experimental system will perform better in the post-test assessment than subjects who use the control system.*

# Variables

We need:

- an *independent variable* which we deliberately manipulate in some way
- one or more *dependent variables* which we measure to see if they vary as a result of the change in the independent variable.

Our *independent variable* might be the system used - our *experimental system* or the *control system*.

The *dependent variable* is the one we will measure in order to determine if changing the independent variable has produced an effect.

The *dependent variable* will be some measure of performance on the two systems, e.g.:

- task-completion time
- level of knowledge/skills acquired
- user satisfaction
- etc.

For example, if testing an interactive training system, we might test subjects' knowledge after performing a task on one or other system

The scores from this test will form the dependent variable.

Ideally there should only be *one* independent variable in a controlled study. All other factors should be held stable.

If this is true, we can assume that any change in the *dependent variable(s)* results from the change we have engineered in the *independent variable*

# Subjects and Groups

The subjects should be chosen to suit the system under test, e.g.:

- potential customers, if testing an eCommerce system
- students, if testing an eLearning application
- people with a relevant special need, if testing an accessible system

Having chosen the subjects, we also have to decide how to assign them to the conditions.

The options are:

- *Independent measures*: divide the subjects randomly into groups, and test each group under a different condition.
- *Matched subjects*: as above, but match the groups according to relevant criteria (e.g., the average IQ score is the same for each group).
- *Repeated measures*: all subjects are tested under all conditions.

The *Independent Measures* design usually requires least effort in testing.

However, since different groups are tested under each condition, we cannot be sure if any changes observed are due to differences between the conditions or differences between the groups of subjects.

To overcome this problem, we need a large group of subjects.

The *Matched Subjects* design requires more effort since we must first test our subjects in order to assign them to groups.

However, since we know that the groups are matched, we can be more confident that any changes observed are due to differences between the conditions, not differences between the groups of subjects.

This is a more *sensitive* experimental design than Independent Measures, allowing us to obtain valid results using fewer subjects.

However, we have to ensure that the subjects are matched on all criteria relevant to the study, and it is not always easy to define these criteria.

The *Repeated Measures* design is very sensitive.

Since the same subjects perform under all conditions, any changes observed must be due to differences between the conditions, not differences between the subjects.

This allows us to obtain valid results using very few subjects.

However, when using a repeated measures design, steps must be taken to avoid *learning effects*:

- The same subjects undertake the same task under each condition.
- They are likely to become better at the task with practice.
- Therefore, they may perform less well under the first condition than under the second/subsequent conditions.

This can be overcome in various ways, e.g.:

- half the subjects undertake the conditions in one order, while the other half undertake them in the reverse order.
- two different tasks are used:
    - half the subjects perform Task A on the experimental system and Task B on the control system
    - the other half perform Task B on the experimental system and Task A on the control system

| | | ***Group 1*** | ***Group 2*** |
|---|---|---|---|
| Option 1 | *Trial 1* | Our System | Control System |
| | *Trial 2* | Control System | Our System |
| **Option 2** | *Trial 1* | Our System<br>Task A | Control System<br>Task A |
| | *Trial 2* | Control System<br>Task B | Our System<br>Task B |

# Experimental Procedure

We must also specify the procedure for the study. This should cover every aspect of the conduct of the study including:

- How much should subjects be told before the experiment begins?
- How much help should subjects receive during the task?
- How much time should be allowed for completion of the task?

# Quality of Data

When designing a test or questionnaire, careful thought should be given to the kind of data it will generate.

If our aim (for example) is merely to gather ideas on how to improve a system, then a qualitative questionnaire will be suitable.

However, if we hope to demonstrate that our system is better than existing systems in some way(s), we may want to use a statistical test to prove this.

In this latter case, we will need to design our test or questionnaire carefully to ensure it yields testable data.

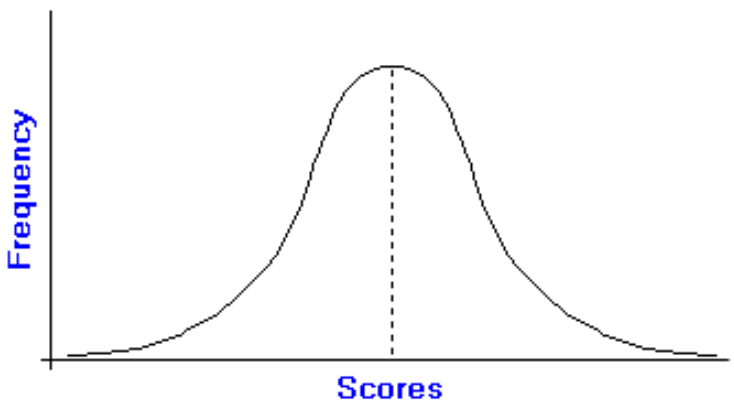
Statisticians classify data under the following headings:

- *Nominal-scaled data*
    - There is no numerical relationship between scores
    - e.g., a score of 2 is not necessarily higher than a score of 1.
- *Ordinal-scaled data*
    - A score of 2 is higher than a score of 1, but not necessarily twice as high.
    - Data obtained from questionnaires is usually ordinal-scaled.
- *Interval-scaled data*
    - A score of 2 is exactly twice as high as a score of 1.
    - Timing data is usually interval-scaled.
- *Parametric data*
    - The data must be interval-scaled (see above)
    - The scores must be drawn from a population that has
        - a *normal distribution*
        - *normal variance*

    Parametric is the most stringent category of data.

# Normal Distribution

If we were to take samples from an *infinite* number of subjects and then chart the frequency distribution, we would probably find that the results show a *normal distribution* (sometimes known as a *bell-curve*).




Research has shown that many psychological, biological and physical variables have normal - or nearly normal - distributions.

Examples include intelligence, height, and the life of many electrical and mechanical systems.

The normal distribution has the following features:

- It is symmetrical, with most of the scores falling in the central region.
- Because it is symmetrical, all measures of central tendency (mean, mode, median) have the same value.
- It can be defined using only the *mean* and the *Variance* (or *standard deviation* - see below).
    - Therefore, once it is known that a set of scores conforms to a normal distribution, and the mean and standard deviation are known, it's very easy to obtain a wide range of information about the data.

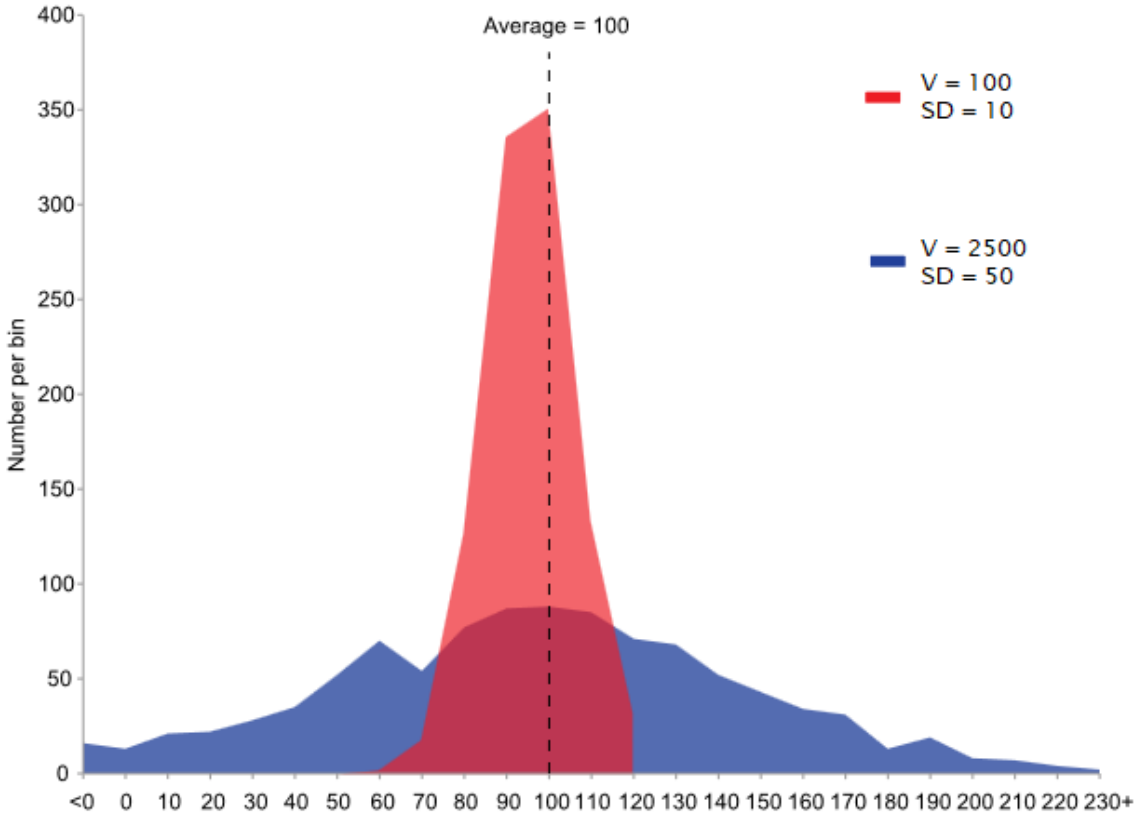
# Variance

*Variance* measures how far a set of (random) numbers are spread out from their average value.

Two sets of results may both be normally distributed but have different spreads, or variances.

The spread of a set of results may also be expressed as a *Standard Deviation* (SD).

Standard Deviation is the square root of Variance




# Sensitivity

The higher the quality of the data we gather, the fewer samples we will need in order to draw valid conclusions.

- If we design a test so that it yields (e.g.) *parametric data* from *related samples*, we can carry out the test using relatively few subjects and still obtain valid results.
- If we design a test so that it yields (e.g.) *nominal-scaled* data from *independent samples*, we will have to test far more subjects in order to obtain valid results.

There are statistical tests to suit each combination of these factors.

| Related Samples Tests | Independent Samples Tests |
|---|---|
| **t-test (related-samples)**<br>*parametric data* | **t-test (independent-samples)**<br>*parametric data* |
| **Wilcoxon**<br>*interval-scaled data* | |
| **Sign test**<br>*ordinal-scaled data* | **Mann-Whitney**<br>*ordinal-scaled data* |
| | **$X^2$ test**<br>*nominal-scaled data* |

In some cases it is simply not possible to design a test that yields parametric or interval-scaled data.

In such cases there is no option but to gather nominal or ordinal-scaled data and take a large number of samples.

However, where it is possible to gather higher-quality data it is advisable to do so.