# 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.

Usability testing may take different forms, depending upon the stage at which it is carried out and the type of data required:

* **Surveys**

Subjects use a system (either informally or under supervision) then fill-in a questionnaire or are interviewed. Surveys can be either:

* + **Qualitative** - the questionnaire contains 'open' questions that may elicit a wide range of responses.

For example, 'What did you like most about the web-site?'.

* + **Quantitative** - the questionnaire contains questions or statements that require a 'yes/no' or numerical response. The results can be analysed statistically if required.

For example, 'The performance was too slow', to which the user should indicate agreement or disagreement on a numerical scale.

* **Observation**

Users are observed (or videoed) using a system, and data is gathered on (e.g.) time taken to perform tasks, number of errors made, etc.

* **Controlled tests**

Can be used at various stages of design to test (e.g.) particular features of a planned interface, novel methods of interaction, complete prototypes, etc..

Can be carried out using 'lo-fi', limited-functionality or fully-functional systems.

Usually involve comparing a new system with a reference system.

The comparison is based on measurable/observable factors such as time to complete task, number of errors made, etc..

The results would normally be analysed statistically.

## Designing Controlled Studies

Let us 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. performance of a task on:
  + a new/experimental system
  + an existing system which serves as a reference
* A **task** which can be performed on both systems
* A **prediction** that can be tested
* A set of **variables**, including:
  + 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 we have made in the independent variable.
* A number of **subjects** who may need to be divided into **groups**
* An **experimental procedure**

### Conditions

A common experimental approach in HCI is to compare a new system with an existing system.

Thus there are two conditions:

* Subjects perform a task on the new system.
* Subjects perform the same task, under the same conditions, on the existing system.

The existing system might be:

* 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 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 perform a certain operation having first learned about it using either the *experimental* or *control* system.

### Prediction

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

Our prediction might be:

* *use of our system will improve learning outcomes compared with use of the control system.*

### Variables

The **independent variable** is the one we wish to control.

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, since subjects undertake the task under each condition, they may perform better under the second/subsequent conditions than on the first because they have had more practice.

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

### 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**

This is the most stringent category. To qualify as Parametric, data must be interval-scaled (see above) and in addition:

* + The scores must be drawn from a *normal population*

If we were to measure our subjects on factors which are important in the study (e.g., intelligence), the results would lie on a *normal distribution* (sometimes known as a bell-curve).

* + The scores must be drawn from a population that has *normal variance*

If we were measure our subjects as described above, the spread of scores would be the same as that found in the general population.

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

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

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.

## Data Analysis

Let us assume that we have conducted a study to evaluate our interactive training system.

The design is as follows:

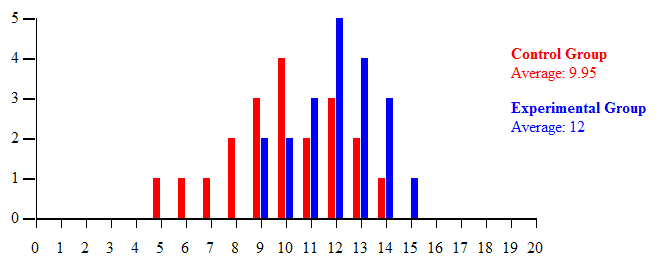
* Two groups of 20 subjects each
* Subjects randomly assigned to groups (independent measures).
* One group performs a set task using our system (*experimental condition*) while the other group performs the same task using the comparison system (*control condition*).
* At the end of the experiment, subjects take a test. The results from this test represent the dependent variable.

|  |  |  |
| --- | --- | --- |
| Assume that we have obtained the data shown opposite:  The left column shows the scores obtained by each subject in the control group.  The right column shows the scores obtained by each subject in the experimental group. | 5 | 15 |
| 12 | 11 |
| 13 | 12 |
| 10 | 13 |
| 7 | 10 |
| 9 | 14 |
| 10 | 12 |
| 12 | 13 |
| 8 | 9 |
| 6 | 11 |
| 10 | 13 |
| 9 | 14 |
| 14 | 12 |
| 8 | 12 |
| 11 | 10 |
| 9 | 11 |
| 11 | 13 |
| 13 | 9 |
| 10 | 12 |
| 12 | 14 |

## The Frequency Distribution

As a first attempt at visualising the result, we might create a *frequency distribution*.

This is a graph showing the frequency with which each score occurs under each condition:



The frequency distribution shows us that the scores for the experimental group appear to be higher than the scores for the control group.

This is a commonly-used *descriptive* method.

It presents the data, without loss, in a form that allows the characteristics of the data to be understood more easily than is possible using just the raw data.

## The Average

Descriptive measures are useful but have limitations. Often we need to summarise the data in some way.

One of the simplest ways to summarise data is by calculating the averages.

The average scores obtained by the two groups in this experiment were:

* Control group: 9.95
* Experimental group: 12

However, this tells us very little about the data.

For small groups of subjects, a single very low or very high score (an *outlier*) can significantly affect the average.

This would be obvious in a frequency distribution but not in an average value.

Therefore the average, while useful, does not capture all the features of the data.

## The Variance

A more useful way of summarising data is to state the *variance*.

This is obtained as follows:

* take the average of the set of scores
* take each score in turn and
  + subtract it from the average to yield the difference
  + square the difference (thus removing the sign)
* sum the squared differences
* divide the total by the number of scores

This is expressed more formally as:

|  |  |
| --- | --- |
|  | Σ ( *X* - *X* )2 |
| variance = |  |
|  | *N* |

Where:

* *X* indicates any score in the set
* *X* indicates the average of the scores
* Σ represents the operation of summing the scores
* *N* represents the number of scores

The variance indicates the amount of *dispersion* in the scores.

By quoting just two values - the variance and the average - we can summarise a set of scores in considerable detail.

Applying this to the set of scores, we obtain the following values for variance:

* Control group: 5.45
* Experimental group: 2.7

Note that there is no obvious relation between variance and score.

For example, a variance of 2.7 does not mean that the average variation between a score and the average is 2.7.

## Standard Deviation

Another widely-used measure of dispersion is the *standard deviation*.

The standard deviation is simply the square-root of the variance.

Standard deviation is often represented as *S*.

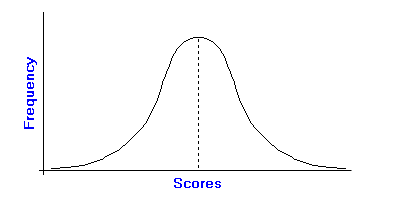
Therefore, variance is often represented as *S*2.

Applying this to the set of scores, we obtain the following values for standard deviation:

* Control group: 2.33
* Experimental group: 1.64

The frequency distribution graph obtained earlier shows marked differences between the two sets of scores, not only in their average values but also in their 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*.



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

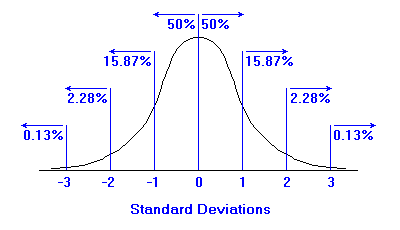
Examples include intelligence, height, and the life of electric lamps.

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 *standard deviation*.

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.

A particularly useful feature of the normal distribution is that the percentage of scores falling above or below a given value of standard deviation is fixed.



Thus it's possible to determine the percentage of scores occurring in any portion of the curve simply by specifying the required portion in terms of its distance from the mean in standard deviations.