Draft ETSI EG 201 472 V1.1.1 (1999-12)

ETSI Guide

Human Factors (HF); Usability evaluation for the design of telecommunication systems, services and terminals

Reference

DEG/HF-00006

Keywords

service, system, terminal

ETSI

Postal address

F-06921 Sophia Antipolis Cedex - FRANCE

Office address

650 Route des Lucioles - Sophia Antipolis
Valbonne - FRANCE
Tel.: +33 4 92 94 42 00 Fax: +33 4 93 65 47 16
Siret N° 348 623 562 00017 - NAF 742 C
Association à but non lucratif enregistrée à la
Sous-Préfecture de Grasse (06) N° 7803/88

Internet

secretariat@etsi.fr
Individual copies of this ETSI deliverable
can be downloaded from
http://www.etsi.org
If you find errors in the present document, send your
comment to: editor@etsi.fr

Important notice

This ETSI deliverable may be made available in more than one electronic version or in print. In any case of existing or perceived difference in contents between such versions, the reference version is the Portable Document Format (PDF). In case of dispute, the reference should be the printing on ETSI printers of the PDF version kept on a specific network drive within ETSI Secretariat.

Contents

Intellectual Property Rights

IPRs essential or potentially essential to the present document may have been declared to ETSI. The information pertaining to these essential IPRs, if any, is publicly available for **ETSI members and non-members**, and can be found in SR 000 314: *"Intellectual Property Rights (IPRs); Essential, or potentially Essential, IPRs notified to ETSI in respect of ETSI standards"*, which is available from the ETSI Secretariat. Latest updates are available on the ETSI Web server (http://www.etsi.org/ipr).

Pursuant to the ETSI IPR Policy, no investigation, including IPR searches, has been carried out by ETSI. No guarantee can be given as to the existence of other IPRs not referenced in SR 000 314 (or the updates on the ETSI Web server) which are, or may be, or may become, essential to the present document.

Foreword

This ETSI Guide (EG) has been produced by ETSI Technical Committee Human Factors (HF), and is now submitted for the Membership Approval Procedure.

Intended users of the present document are:

- usability experts who have to conduct usability evaluations and want to consult about methods to carry them out;
- designers who want to know basic principles of usability evaluation, and perhaps want to choose between the most commonly used methods;
- any person interested in usability, and more particularly in usability testing methods.

Introduction

Since the publication of ETSI ETR 095 [1] in 1993, usability evaluation methods have evolved greatly. TC HF, as the Technical Body in charge of usability issues within ETSI, has considered important to review the evaluation methods and procedures currently used in either the telecommunications or in the software design areas, so that usability experts, service and terminal designers, and people interested in these issues have the best information and can make the best choice amongst the range of available techniques.

It is also important to note the appearance of important standards or draft standards regarding usability since the publication of ETR 095 [1] : ISO 9241 [2], which specifies usability requirements for office work with computers, and the recently approved standard ISO 13407 [3], on the Human-Centred Design Process.

Research in any area, either scientific or in applied settings, requires the use of a commonly accepted and proven methodology. Furthermore, it comprises both a "philosophy", or a standard process, and the use of common methods and procedures. Using a standard research methodology across different projects allows for comparisons and the possibility of obtaining conclusions which would be impossible if different methods are used.

Different techniques and procedures are more appropriate depending which purpose and which moment in the development process. The aim of the present document is to provide guidance to make the best choice of the method, and to provide the necessary information to carry it out in the practice.

The purpose of the present document is to provide the basis for the use of a common methodology when performing usability evaluations. Another important objective of the present document is providing guidance on the Human-Centred Design (HCD) Process, and particularly, fighting against the idea of usability evaluation as a “final test". In the HCD approach, usability evaluation, in a form or other, is a step in several stages of the proposed iterative process. The present document presents criteria for choosing the applicable method taking into account the design phase in the Human-Centred Design Approach.

On the other hand, within the 4th Research Framework supported by European Commission, and more especifically within the ACTS program (Advanced Communication Technologies and Services), numerous advanced technology trials with real users have been carried out. One of these projects, USINACTS (Usability Support in ACTS) has reviewed the different approaches and procedures used in these trials, and one of the first conclusions obtained is that information about usability evaluation methods is not so easy to find, even for usability experts. Another important output of this project is a study of the application of the HCD in real projects. USINACTS [4] has made an extensive review of these procedures, and prepared a tutorial on usability assessment methodology, which is the basis for the present document.

Previous EU-funded Research Programs results, as those in RACE-ISSUE [5] or TELEMATICS-INUSE [6] have also been reviewed for the purposes of the present document.

1 Scope

The present document gives guidance on usability evaluation methods and procedures, with special emphasis on its application within the framework of the Human-Centred Design Process.

The present document is applicable to the usability evaluation of telecommunication systems, services and terminals, taking into account their intended users from the initial phases of the design process. A closer definition of systems, services and terminals is required here: with systems, services and terminals we mean complete products for its use, either finished or in a prototype state, but not particular components. For these, there may exist particular testing methods, which are not covered in the present document. For instance, refer to current ETSI and ITU standards for icon testing, video or audio quality, etc.

The approach of the present document is applied instead of theoretical: the choice criteria for a person in charge of the evaluation process to choose which method in which design phase are exposed first, and then the different methods and procedures are reviewed, and examples are provided whenever considered appropriate. Directly applicable material, as, e.g., response format for questionnaires, will be provided wherever appropriate.

Particular issues presented in the present document are:

- A review of the current standards on Human-Centred Design Process (ISO 13407 [3]) and their application for telecommunications systems, terminals and services design.
- The state of the art in usability assessment techniques, including new methods appeared since the publication of ETR 095 [1] either in telecommunications or in software design areas.

2 References

The following documents contain provisions which, through reference in this text, constitute provisions of the present document.

- References are either specific (identified by date of publication, edition number, version number, etc.) or non-specific.
- For a specific reference, subsequent revisions do not apply.
- For a non-specific reference, the latest version applies.
- A non-specific reference to an ETS shall also be taken to refer to later versions published as an EN with the same number.

[1] ETR 095: "Human Factors; Guide for Usability Evaluations of Telecommunications Systems and Services".

[2] ISO 9241 (1988): "Ergonomic requirements for office work with visual display terminals (VDT’s). Geneva: International Standards Organization".

[3] ISO 13407 (1999): "Human-centred design processes for interactive systems".

[4] Concejero, P., Clarke, A. M., Kaasinen, A. (1999): "The USINACTS Usability Assessment Tutorial. Proceedings of 17th International Symposium on Human Factors in Telecommunications". Copenhagen, May 4-7. 1999. USINACTS Usability Assessment Tutorial is available in Internet: http://atwww.hhi.de/usinacts.html

[5] RACE 1065-ISSUE (1992): "ISSUE Usability Evaluation Guidelines. Brussels: Commission of the European Communities".

[6] INUSE project document (1996): "User Centred Design. TELEMATICS APPLICATION PROGRAMME, Brussels Commission of the European Communities".

[7] EEC Directive 90/270/EEC: "Council Directive of 29 may (1990): On minimum safety and health requirements for work with display screen equipment. Brussels: European Union".

[8] Nielsen, J. (1993): "Usability Engineering. New York: Academic Press".

[9] Nielsen, J. and Landauer, T. K. (1993): "A mathematical model of the finding of usability problems". Proc. ACM INTERCHI '93 Conf (Amsterdam, The Netherlands, 24-29 April) 206-213

[10] Adam, N., Y. Yesha et al. (1996): "Strategic Directions in Electronic Commerce and Digital Libraries". ACM Computing Surveys, vol. 28, no. 4, Dec. 1996.

[11] Kaasinen, E., Clarke, A., Concejero, P. (1999): "Usability Training for Project Managers and Designers". Human-Computer Interaction HCI '99. Munich, August 1999.

[12] USINACTS Deliverable 7: "ACTS Usability Evaluation Guideline". Brussels: European Commission.

[13] USINACTS Deliverable 14: "Matrix of usability issues in USINACTS User Group". Brussels: European Commission.

[14] Campbell, D. T., and Stanley, J. C. (1966): "Experimental and Quasi-Experimental Designs for Research". Chicago: Rand Mc. Nally.

[15] Cook, T.D., and Campbell, D. T. (1979): "Quasi-experimentation: Design and Analysis Issues for Field Settings". Chicago: Rand Mc. Nally.

[16] ACTS SII Guideline G4 (1996): "Organization of advanced communication services trials with residential users". Brussels: European Commission. ACTS Research Program.

[17] Miller, S. (1975): "Experiment Design and Statistics". Essential Psychology Series A8, Methuen.

[18] Wilkinson, J. (1995): "Direct Observation". In Breakwell, G. M., Hammond, S. and Fife-Schaw, C. (eds) Research Methods in Psychology. London, Sage.

[19] Cronbach L J, Gleser G C, Nanda H and Rajaratnam N, (1972): "The dependability of behavioural measurements". New York, Wiley

[20] Mack, R. L. and Nielsen, J. (1993): "Usability inspection methods". ACM SIGCHI Bulletin 25, 1 (January), 28-33.

[21] Molich, R. and Nielsen, J. (1990): "Improving a human-computer dialogue". Communications of the ACM 33, 3 (March), 338-348

[22] Millward, F. (1995): Focus Groups. In Breakwell, G. M., Hammond, S. and Fife-Schaw, C. (eds) Research Methods in Psychology. London, Sage.

[23] Tryfos, P. (1996): "Sampling methods for Applied Research". New York. Wiley

[24] Fife-Schaw, C. (1995): "Surveys and Sampling Issues". In Breakwell, G. M., Hammond S. and Fife-Schaw C. (eds): Research Methods in Psychology. London. Sage

[25] ITU-R BT.500-7 (1995): "Methodology for the subjective assessment of the quality of television pictures". International Telecommunications Union, Geneva.

[26] Rubin, T. (1988): "User Interface design for computer systems". Ellis Horwood, Chichester.

[27] Lewis, C. (1982): "Using the "thinking-aloud" method in cognitive interface design". Research Report RC 9265. IBM T. J. Watson Research Center, Yorktown Heights, New York.

[28] Ericsson, K. A. and Simon, H. A. (1984): "Protocol Analysis": Verbal Reports as Data. The MIT Press, Cambridge, MA, USA.

[29] O'Malley, C. E. et al. (1984): "Constructive interaction: A method for studying human-computer-human interaction". PROC IFIP INTERACT '84 First Intl Conf Human-computer Interaction (London UK, 4-7 September) 269-274.

[30] Hewett, T. T. and Scott, S. (1987): "The use of thinking-out-loud and protocol analysis in development of a process model of interactive database searching". Proc IFIP INTERACT '87 second Intl Conf Human-Computer Interaction (Stuuttgart, Germany 1-4 September) 51-56.

[31] Mack, R. L., and Burdett, J. M. (1992): "When novices elicit knowledge: Question-asking in designing, evaluating and learning to use software". In Hoffman, R. (ed.): The Psychology of Expertise: Cognitive Research and Empirical AI. Springer-Verlag, New York, pp. 245-268.

[32] Tukey, J. W. (1977): "Exploratory Data Analysis". Cambridge: Addison Wesley.

[33] American Psychological Association (1994): Publication Manual, 4th ed. Washington: APA.

[34] Hays, W. L. (1988): Statistics. 4th. edition. Chicago: Holt, Rinehart and Winston.

[35] Kirk, R. B. (1982): "Experimental Design Procedures for the Behavioral Sciences". Monterey, California: Brooks Cole.

[36] Maxwell, S. E. and H. D. Delaney (1990): "Designing Experiments and Analyzing Data: A Model Comparison Perspective". Belmont, California: Wadsworth.

[37] Montgomery, D. C. (1991): "Design and Analysis of Experiments". 3rd. edition. New York: Wiley.

[38] Hoaglin, D. C., F. Mosteller and J. W. Tukey, eds. (1991): "Fundamentals of Exploratory Analysis of Variance". New York: Wiley.

[39] Gorsuch, R. L. (1983): "Factor analysis". Hillsdale, NJ: Lawrence Erlbaum.

[40] Lawley, D. N., and Maxwell, A. E. (1971): "Factor analysis as statistical method". New York: Elsevier.

[41] Afifi, A. A., and S. P. Azen. (1979): "Statistical Analysis: A Computer Oriented Approach. 2d ed. New York: Academic press.

[42] Harman, H. H. (1967): "Modern Factor Analysis". 2d ed. Chicago: University of Chicago Press.

[43] Jöreskog, K. G. and D. Sörbom (1979): "Advances in Factor Analysis and Structural Equation Models". Cambridge, Massachussetts: Abt associates.

[44] Saris, W. and H. Stronkhorst (1984): "Causal Modelling in Nonexperimental Research: An Introduction to the LISREL Approach". Amsterdam: Sociometric Research Foundation.

[45] Kruskal, J.B. and Wish, M. (1978): "Multidimensional Scaling. Sage University Paper series on Quantitative Applications in the Social Sciences", 07-011. Beverly Hills and London: Sage Publications.

[46] Arabie, P., Carroll, J.D., and DeSarbo, W.S. (1987): "Three-Way Scaling and Clustering. Sage University Paper series on Quantitative Applications in the Social Sciences", 07-065. Beverly Hills and London: Sage Publications.

[47] Tabachnick, B. G., and L. S. Fidell (1989): "Using Multivariate Statistics". New York: Harper and Row.

[48] Afifi, A. A., and V. Clark (1984): "Computer-aided Multivariate Analysis". Belmont, CA: Lifetime Learning Publications.

[49] Hartigan, J. A. (1975): "Clustering Algorithms". New York: Wiley.

[50] Aldenderfer, M. S. (1984): "Cluster Analysis". Beverly Hills: Sage Publications.

[51] Davidson, M. L. (1980): "The Multivariate Approach to Repeated Measures". BMDP Technical Report #75. Los Angeles:. BMDP Satatistical Software, Inc.

[52] Bishop, Y. M., S.E. Fienberg and P. W. Holland (1975): Discrete Multivariate Analysis: Theory and Practice. Cambridge, Massachussetts: MIT Press.

[53] Hosmer, R. J. and S. B. Lemeshow (1989): "Applied Logistic Regression". New York: Wiley.

[54] British Psychologists Association: "The Code of Conduct". Available in Internet in: http://www.bps.org.uk/charter/codofcon.htm

[55] American Psychological Association (1992a): "Ethical principles of psychologist and code of conduct". American Psychologist, 47, 1957-1611. Available in Internet in: http://www.apa.org/ethics/code.html

[55] Measurement Theory: "Frequently Asked Questions". Available in Internet in: ftp://ftp.sas.com/pub/neural/measurement.html

[56] ISO/DIS 9241-11: "Ergonomic requirements for office work with visual display terminals (VDTs) -- Part 11: Guidance on usability".

3 Abbreviations

3.1 Abbreviations

For the purposes of the present document, the following abbreviations apply:

ACTS	Advanced Communications Technologies and Services
ANOVA	ANalysis Of VAriance
APA	American Psychological Association
BPS	British Psychological Society
DV	Dependent Variable
EG	ETSI Guide
EG	ETSI Guide
ETR	ETSI Technical Report
ETSI	European Telecommunications Standards Institute
EU	European Union
FAQ	Frequently Asked Questions
HCD	Human Centred Design
ISO	International Standards Organization
ISSUE	IBC Systems and Services Usability Engineering (a RACE Project)
ITU	International Telecommunications Union
IV	Independent Variable
LISREL	LInear Structural RELationships (A multimariate statistical technique)
MANOVA	Multivariate ANalysis Of VAriance (A multivariate statistical technique)
MDS	Multi-Dimensional Scaling (A multivariate statistical technique)
MOS	Mean Opinion Score
PCA	Principal Components Analysis (A multivariate statistical technique)
RACE	Research on Advanced Communications in Europe (an EU-funded Research Program)
RT	Reaction Time
TC HF	Technical Committee Human Factors
USINACTS	USability IN ACTS (an ACTS project)

4 What is usability?

This clause presents current definitions of usability, including the social and economic impact of the application of the techniques for improving usability in product development.

4.1 The ISO definition of usability

ETR 095 [1] attempted to make a definition of usability, based on the distinction between performance measures and attitudes towards a system. Since its publication, the most widely accepted standard on usability is ISO 9241 [2] , on which the definitions of usability in the present document are based.

The ISO 9241 [2] standard describes ergonomic requirements for office work with visual display terminals. Part 2 of the standard is Guidance on usability aspects. This standard defines how to specify and measure the usability of products. and the factors which have an effect on usability.

The standard states that when specifying or measuring usability, the following information is needed:

1. A description of the intended goals.
2. A description of the components of the context of use including users, tasks, equipment and environments. This may be a description of an existing context, or a specification of intended contexts. The relevant aspects of the context and the level of detail required will depend on the scope of the issues being addressed. The description of the context needs to be sufficiently detailed so that those aspects of the context which may have a significant influence on usability could be reproduced.
3. Target or actual values of effectiveness, efficiency, and satisfaction for the intended contexts.

In order to specify or measure usability it is necessary to identify the goals and to decompose effectiveness, efficiency and satisfaction and the components of the context of use into sub components with measurable and verifiable attributes.

- **Effectiveness** is the accuracy and completeness which specified users can achieve specified goals in particular environments.
- **Efficiency**: the resources expended in relation to the accuracy and completeness of goals achieved.
- **Satisfaction**: the comfort and acceptability of the work system to its users and other people affected by its use.

The components and the relationships between them are illustrated in figure 1.

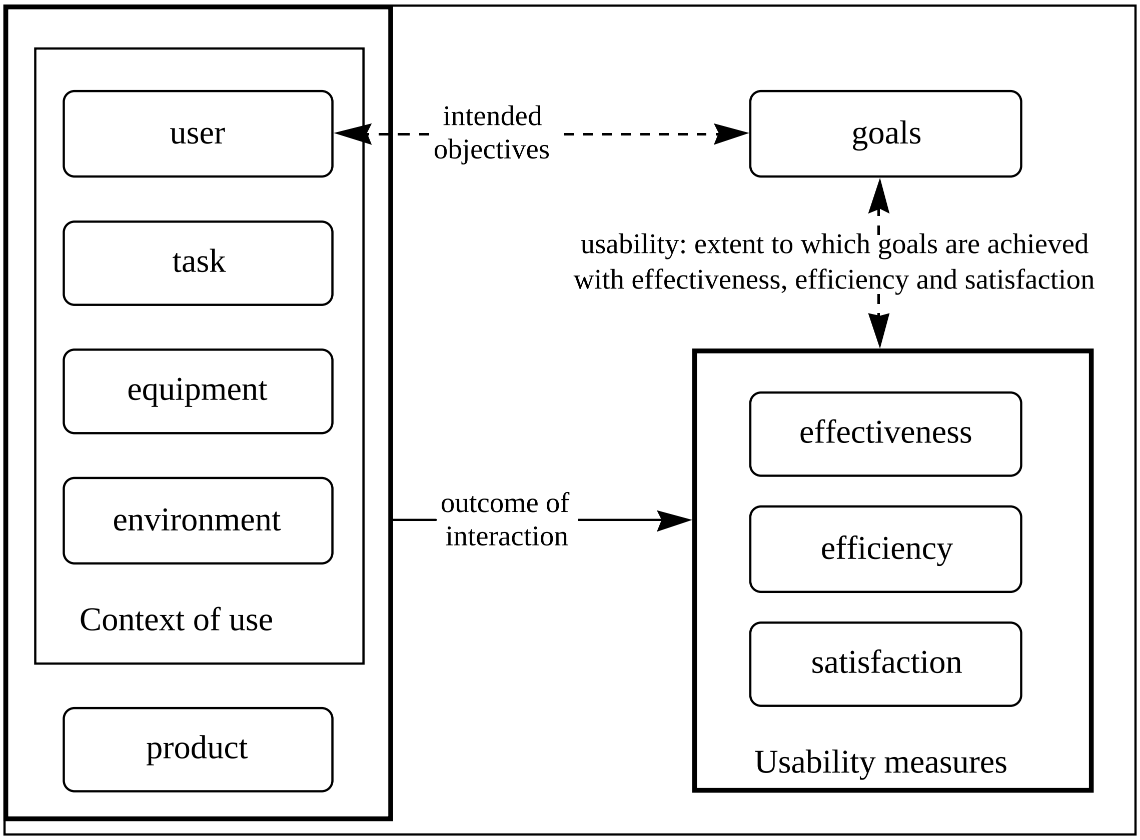


Figure 1: Usability framework according to ISO/DIS 9241-11 [56]

The context of use defined by the standard includes the following factors:

Description of users

Characteristics of the users need to be described. These can include knowledge, skill, experience, education, training, physical attributes, and motor and sensory capabilities. It may be necessary to define the characteristics of different types of user, for example users having different levels of experience or performing different roles.

Description of tasks

Tasks are the activities undertaken to achieve a goal. Characteristics of tasks which may influence usability should be described, e.g. the frequency and the duration of the task.

Detailed descriptions of the activities and processes may be required if the description of the context is to be used as a basis for the design or evaluation of details of interaction with the product. This may include description of the allocation of activities and steps between the human and technological resources. Tasks should not be described solely in terms of the functions or features provided by a product or system. Any description of the activities and steps involved in performing the task should be related to the goals which are to be achieved.

For the purposes of evaluating usability, a set of key tasks will typically be selected to represent the significant aspects of the overall task. User tasks and sub-tasks can be identified by task analysis.

Description of equipment

Relevant characteristics of the equipment need to be described. The description of the hardware, software and materials may be in terms of a set of products, one or more of which may be the focus of usability specification or evaluation, or it maybe in terms of a set of attributes or performance characteristics of the hardware, software and other materials.

Description of environments

Relevant characteristics of the physical and social environment need to be described. Aspects which may need to be described include attributes of the wider technical environment (e.g. the local area network), the physical environment (e.g. workplace, furniture), the ambient environment (e.g. temperature, humidity) and the social and cultural environment (e.g. work practices, organizational structure and attitudes).

Usability measures

Usability measures include effectiveness, efficiency and satisfaction. These are measured in user trials of the product.

4.2 Social and economic impact of considering usability in the design process

What benefits can be achieved through an appropriate application of the Human-Centred Design approach?

There are many reasons, and one of them is the recent legal regulations for designing safe systems which do not harm the health or the well being of their users EEC Directive 90/270/EEC [7].

But usability engineering has proved that application of the principles of Human-Centred design have high payoffs (INUSE project document: User Centred Design, 1996) [6]:

- Reduced production costs: the overall development times and costs can be reduced by avoiding over design and reducing the number of changes required late in design.

- Reduced support costs: systems which are easier to use require less training, less user support and less subsequent maintenance.

- Reduced costs in use: systems better matched to user needs improve productivity and the quality of actions and decisions. Easier to use systems reduce stress and enable workers to handle a wider variety of tasks. Difficult to use systems reduce health and well being, motivation and may increase staff turnover and absenteeism. Difficult to use systems are time consuming to use, and are not exploited to full advantage, as the user may be discouraged from using advanced features. In some circumstances, they may never be used. An ineffective system may be a major financial liability for the user organization.

- Improved product quality: Human-centred design results in products which have a higher quality of use and are more competitive in a market which is demanding easier to use systems.

All these benefits are obtained taking on account the total life-cycle costs of the product, not only those of the development, but also the set-up phase and the maintenance phase. The initial costs of human-centred design activities are compensated by all the benefits that produce at the end. In spite of being so important benefits, they are seldom expressed in economic terms.

It is not so easy to estimate the benefits produced by applying Human-Centred Design. This is usually a difficult task, which has only recently found some solutions. For instance, Jakob Nielsen has reported estimates of the benefits produced by the application of usability techniques in software development (Nielsen, 1993 [8]). Nielsen and Landauer (1993) [9] have also reported benefits in several projects which they claim can be up to 5000 times the cost.

On the other hand, social impact of these technologies is experimenting spectacular growth. As the global information infrastructure expands in exponential fashion (e.g., during last years, Internet population doubled every few months), there are important changes in how people work and in some other aspects of daily life (cf. Adam et al. 1996) [10] . The information is now more global, and more easily accessed. Virtually all entities, from large companies to individuals, through small and medium enterprises, are engaged in activities that increasingly involve accessing remote databases, and the competitiveness depend heavily on the effectiveness and efficiency of that access.

As a consequence, the potential user community for advanced technologies is becoming very large, and rather non-technical, incorporating people not previously used to computers or I.T., or telecommunication technologies. It is therefore urgently necessary and strategically critical to put in practice all kinds of techniques and approaches that enable these people to become used with advanced technologies. Among the most urgent tasks are to develop user interfaces that require minimal technical expertise by the users, and support a wide variety of information-intensive tasks.

5 Usability evaluation and the Human-Centred Design Process

Which are the principles of Human-Centred Design?

The essential principles of Human-Centred Design are:

- an appropriate allocation of functions between user and system should be carried out;
- the design solutions should be iterative, and the decisions should be made depending on the results of usability;
- evaluations, thus providing feedback from each phase to the others;
- it requires that users actively involve in the design;
- the design teams should be multidisciplinary, thus taking on account all knowledge required to produce a usable system.

To what products can the HCD be applied?

The procedures outlined in this clause can be applied to any system component the user may have to interact with. This includes hardware and software components, as well as user manuals.

5.1 The Human-Centred Design Process

ISO standard ISO 13407 [3], Human centred design processes for interactive systems, provides guidance on human centred design activities throughout the life cycle of computer-based interactive systems. The standard is targeted to people who manage design processes. According to ISO 13407 [3] , human centred design consists of four different types of design activities:

- understand and specify the context of use;
- specify the user and organizational requirements;
- produce design solutions;

- evaluate design against requirements.

According to ISO 13407 [3] , a development project must specify the procedures used, the information collected and use made of the results.

The main problem with the currently available HCD process found by the USINACTS project [11] [12] [13] is that it cannot be clearly connected to the software design process. It is essential to be able to talk to the designers in their own language, i.e. to describe the phases of software design process and integrate the activities of human centred design to this familiar process. ISO 13407 [3] is a concrete step to this direction.

The results of user involvement are most effective in the early phases of the projects. That is why HCD process makes emphasis in the application of methods for user requirements definition and software requirements analysis. User requirements cannot be fixed in the beginning of the project but the project must be prepared to refine user requirements throughout the design process. New ideas during the development should be registered even if the project does not plan to implement them at that moment. Each requirement shall be equipped with information about where the requirement came from and how the project decided to handle the requirement. If the requirement was rejected, the project shall record why.

The usability activities in the design process can be described as illustrated in figure 2.

1) define initial user requirements;

2) visualise design decisions in the software requirements phase:

   - use case descriptions;

   - screen views, user interface prototypes;

   - collect feedback and new ideas from the users;

3) prototype during the design phase and evaluate with the users;

4) organise field trials in the end, but it is recommended not to plan field trials without having laboratory tests with the users first. In the laboratory trials the main usability problems can be identified and they can be solved before the field trials. In the field trials you can get feedback of the usability of the system in real and continuous use.

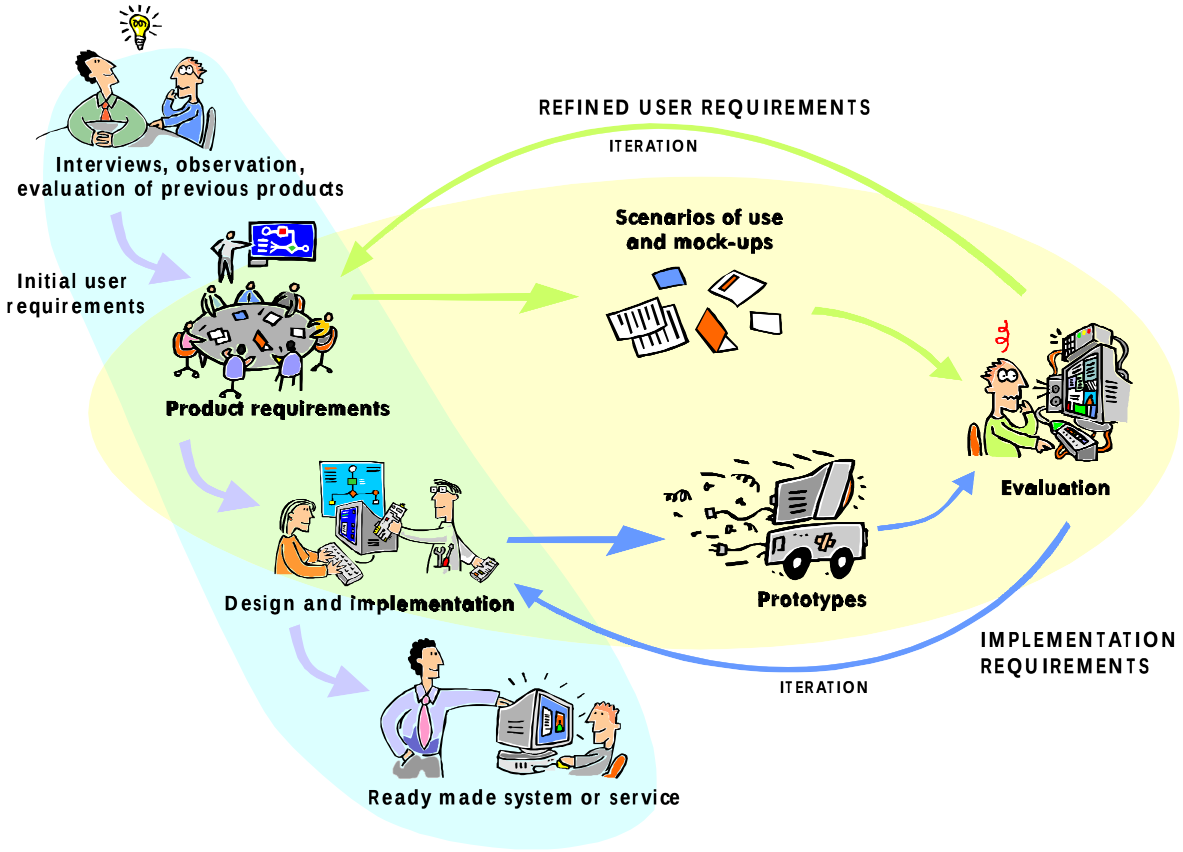

REFINED USER REQUIREMENTS
ITERATION
Interviews, observation, evaluation of previous products
Initial user requirements
Scenarios of use and mock-ups
Product requirements
Evaluation
Design and implementation
Prototypes
IMPLEMENTATION REQUIREMENTS
ITERATION
Ready made system or service

Figure 2: Usability activities during the design process as recommended by the USINACTS Project

The application of HCD in development projects has proven to provide a great benefit. Below you will find answer to typical questions regarding this application:

What can a Human-Centred Design do?

A Human-Centred Design can be an important contribution to guarantee the success of the product with a variety of different users and to avoid the failure of the product. This could happen if a technologically sound product does not take into account the requirements of the users, one of which is usability.

What a Human-Centred Design cannot do?

Neither Human Factors, nor usability engineering can do any kind of magic for solving usability problems at the end of projects. This is a very common mistake. For this reason, it is important to stress the participation of the usability specialists in all the steps of the design process, to avoid surprises when the system is fully specified and working.

Usability engineering, even when carefully planned and carried out, cannot guarantee to sell a product per-se. Although an increasingly important attribute for people's purchasing decisions, marketing activities must complement and stress the most important features of the product.

5.2 Usability evaluation within the Human-Centred Design Process

Very often, usability techniques are only considered at the end of the development of a product, which is a mistake, since it is most useful in all steps of the development process. ETSI ETR 095 [1] specifies the application of usability techniques in the different phases:

Definition of usability goals.

The usability goals are the desired end states which should be met for the system or service be judged as usable. The dimensions of usability have already been commented: effectiveness, efficiency, satisfaction, learnability and flexibility.

These goals should be particularized in measurable criteria, either absolute (e.g., the minimum level the system or service should meet), or relative (in comparison with previous systems or prototypes). It is important to specify how the validity of these measures going to be analyzed, this is, up to what extent the measures being taken will relate to the global measure of the system. E.g., how much error rates relate to the user's final consideration of the system.

Identification of user's population, task and environmental characteristics.

This step specifies the target group for the evaluation. It is essential to perform sampling or the selection of the users who will participate in the evaluation process. Tasks are usually many, and very different. Appropriate sampling or identification of the critical tasks for evaluation are the base of further steps of the evaluation process. Environmental characteristics should also be defined, since they may have influence on the measures being undertaken. These variables should be identified during this phase and addressed in the design phase.

Specification of usability criteria.

Usability criteria specify how any one particular goal is to be achieved. This includes operationalization of the variables, this is, specifying how they will be measured, the precision with which it can be achieved, etc.

Scenario building.

A scenario represents a sequence or flow of user actions required to achieve a specific task. There are usually many different ways to achieve the goal, and then different users may follow different steps while interacting with the system or service.

Building the scenarios for each task ensures that all users follow the same process, and that measures taken are comparable (e.g., time and errors measures, and subjective data).

The scenarios should be based on realistic user tasks, and only after the whole set of them have been assembled it will be possible to perform the usability testing.

Usability testing.

This step should only be performed after all the previous steps have been accomplished, and if the design (see below for each technique) has been completed. This is by far the most expensive and time-consuming phase, so its success or failure critically depends on its appropriate planning.

6 Choosing the evaluation method

This clause is intended to give a summary table of the properties of each method reviewed in the present document, so that evaluators can compare them and choose the most appropriate for their requirements and the moment of the evaluation.

The table is divided into two different parts: first part describes testing and evaluation methods, and the second part includes data collection and measurement techniques. This means that you can prepare one of the methods in the former category with any (although not always!) of the data collection methods in the later category. For instance, you can plan experiments gathering data by means of audio-video recording and questionnaires at the same time. But, for example, questionnaires are usually not made at the same time that observation.

Of course, the usability expert is always free to choose according to some other needs. Here we try to put the most common techniques, and also its most common applications.

Table 1: Testing and evaluation methods

Method Name	Design cycle Stage	Users Needed	Main Advantage	Main Disadvantage
Experiments	Components (hardware or software) design. Establishing generic principles for system design, basic HF research.	Depends on complexity. At least 10 for design cell.	It allows to test design hypotheses or competing alternatives in an optimal way.	Complex techniques involved, which requires expert knowledge for maximum benefit. Usually made in the usability laboratory, and not in the real use environment.
Field Observation	Final testing. Task analysis.	3 or more	It is made in real use environment: provides first-hand feedback on the user's interactions in the context of the real task, and it is flexible to circumstances.	Very costly. Difficult to analyze, and to know the reasons for behaviour. Different observers may differ in interpretations
Heuristic evaluation	Early design, "inner cycle" of iterative design	None (it is made by experts)	Finds out individual usability problems. Can address expert user issues.	Does not involve real users, so does not find "surprises" relating to their needs.
Focus groups	Task analysis, user requirements	6-9 per group	Spontaneous reactions and group dynamics. Allows to find out opinions or factors to be incorporated in other methods (e.g., surveys)	Hard to analyze. Low validity.
Input logging	Final testing, follow-up studies	At least 20	Finds highly used or unused features. Can run continuously, and may be felt non-intrusive. Data gathering is automatic, and a permanent record of the interaction can be kept.	Analysis programs needed for huge mass of data. Data is at a very fine level, requiring time-consuming data consolidation and reduction. Violation of users' privacy.
Surveys	Follow-up studies, User feedback. User requirements.	Hundreds	Tracks changes in user requirements. Analysis of user's opinion for the working system in its real environment.	Sampling procedures and field tests organization require a lot of work, thus costly.

Table 2: Data collection and measurement techniques

Technique	Method in which it is used	Main Advantage	Main Disadvantage
Questionnaires	Surveys. Experiments. Structured interviews.	Easily elaborated and compared, once a validated instrument is developed. Written interchange inherently more “formal" and less natural than a spoken interchange May be self-completed by users , and thus easy and cheap to repeat. Usually appropriate to find out subjective user preferences and attitudes.	Pilot work needed to validate the instrument can be costly and complex. May require prompting to users for stimulating completion. Less effective communication: questions and answers may be interpreted differently or not be well understood. Contradictions may be overlooked or require second round of questions to check.
Interviews	User requirements. Task analysis	Flexible, in-depth attitude, experience probing and spontaneous information. Effective communication: ability to explain questions better and to interpret answers better. Contradictions may be pointed out and explained right away.	Time consuming. Hard to analyze and compare. Open answers data must be consolidated and structured for comparison. Requires considerable manpower. Reactions may be influenced by interviewer.
Performance measures (e.g., reaction time, error rates)	Performance evaluation Experiments	Objective measures. Results easy to compare. They can be also taken by experts (e.g. mistakes) and include expert judgements	Does not find subjective constructs (opinion, attitudes, satisfaction).
Thinking aloud	Experiments. Interviews.	Points out cognitive processes implied in the use of the system. It highlights users misconceptions and conceptual models.	Unnatural for users. Hard for expert users to verbalise. Information is difficult to analyze.
Audio-video recording	Observation Experiments	Records all behaviours and can be kept for analysis in the future. Wealth of data also on “body language" reactions Possibility of multiple analysis by different observers increasing reliability of results	Ethical and legal requirements (see clause 9). Behaviour has to be categorized. Very costly and lengthy (10 hours to analyse satisfactorily 1 hour of videotape)

7 Usability evaluation techniques in detail

7.1 Reliability and validity of usability evaluation

Reliability and validity are two important principles in any scientific research, and also in usability evaluation. Both concepts are traditional in Psychometrics and related fields (e.g., survey research), but also apply to any usability testing.

Reliability is defined as the property of a measure of being stable on time and across different conditions, i.e., a reliable measure should not be affected by changes in the interviewer, language, or environmental conditions. For instance, if a questionnaire is reliable it should yield the same results if used by different researcher on different occasions.

Validity is defined as the degree to which the research procedure has measured whatever it was intended to measure. The validity of a measurement (e.g. that obtained by means of a survey) is a technical concept which refers to the degree in which the empiric evidence and the theoretical basis support the interpretations and inferences made from the measurements. Validity also comprises the degree of adequation of the measurement instrument (and therefore of the measurements made from it) for a specific usage.

A thorough review of validity testing is clearly beyond the scope of the present document, but it is important to state the different types of validity. The best known approach to validity in social research was commenced by Campbell y Stanley (1966) [14], and updated in 1979 by Cook and Campbell [15]. Best known classification of validity is in two concepts:

- **Internal Validity**: the degree to which the observed relationship between variables is obtained from the hypotheses handled by the researcher, or there are other hypotheses that can explain the same results. A classical example of lack of internal validity is the presence of spurious correlation's.

This includes Content and/or construct validity: this is a concept which refers to the degree which the items or questions of the test or survey represent the totality of the possible items for the construct or behaviour being measured. On the other hand, usability testing procedures usually measure non-observable constructs, in some cases very complex ones, e.g., attitudes. Construct validity refers to the degree which the questions or items in the test or questionnaire appropriately represent the construct under measurement.

- **External Validity**: the degree to which the conclusions from a particular research, made in particular conditions, can be generalized to any other kind of conditions or contexts (e.g., different samples, places or moments in the time). A typical problem for external validity is the representativeness of the sample in the research. As part of the HCD process, as outlined above, the identification of the user population is a crucial step. To achieve the goal of having a representative group of users of that population, sampling is an essential process in many usability testing procedures. Good advice about this particular issue can be found in ACTS SII guideline G4, "Organization of advanced communication services trials" [16].

Campbell and Stanley distinguish between four types of external validity:

Criterion-related or predictive validity

The criterion is a variable or characteristic of real interest whose performance we are interested to predict from the measure (e.g. purchasing a particular device from an attitude towards it). The criterion is a direct and independent measurement which is tried to predict by the measure. Criterion validity can be tested by computing the correspondence between the classification done with the measure and the classification made with the external criterion.

Historic validity:

This is the degree to which the research is dependent on the particular historic moment in which the research is made. This of course greatly depends on the particular research area.

Ecologic validity:

This is the degree to which the research results can be generalized to real-world situations (in practice, all research is made in more or less artificial situations).

Validity of statistical conclusions:

it comprises answering the question "are results due to any statistical artifact or not?". Many modern statistical procedures are so complex that they can produce results which have no real relationship with the real world.

7.2 Experiments

7.2.1 When should experiments be used in usability evaluation?

Experiments are still the quintessential scientific research method. There are many reasons for this, and the first one is that experiments allow to obtain "strong" conclusions about research hypotheses, and to make optimum decisions between competing alternatives. Another very important reason is that using appropriately the experimental framework the obtained knowledge is accumulative, the continuous replications and variations of the conditions in a particular series of experiments allow the researcher to obtain more data to confirm or not his or her hypotheses. This sense of "research framework" and of replication and variation of particular conditions has been used as a kind of philosophy, as well as taking advantage of the available techniques, for applied research areas, like in quality control. Furthermore, it is perhaps the only practical way for usability experts to develop theories or models to explain how and why people prefer some things against others, for example. In this sense, experiments are essential to develop guidelines that can be generalized to a broader scope than the case that has been tested.

However all these advantages, its use in applied settings, as in usability evaluation, may become really complex, and we think it is important to clearly point out its requirements so that the usability expert can obtain the maximum benefit of these powerful techniques. Some well known usability researchers, as Nielsen (1993) [8], do not cover traditional experiments in their handbooks, and they prefer other quicker and less complex techniques. Actually, if you get an experimental design handbook, you will find out how complex this issue can be, and at the same time, how it has become an area for specialists itself, with very few references or examples of real applications.

Our aim then with this part of the present document is to provide a summary of the most common experimental methods, especially those applicable to usability evaluation, and to clearly specify their requirements to take full advantage of its application. Other points in this chapter provide guidance about choosing the particular design and controlling the experimental factors and other variables, another characteristic feature of this method.

7.2.2 Requirements of experiments

- Experiments are the best method to test between competing hypotheses or alternatives. This also means that, for really making an experiment, there must be alternative systems or situations to be compared. In system development this can be complex, difficult, or impossible. However, they are usually very easy to do in the prototype phase. The usability expert should make clear to the development group that following the experimental approach, these methods provide optimum results. The hypothesis being tested should be very clearly made before the experiment, and the optimum way is to have them formalized in a mathematical way.

- Experiments can be carried out even if the system being tested does not exist. For instance, Wizard of Oz techniques are a tool to produce alternative menu procedures in speech recognition services. Basically, in the "Wizard of Oz" technique, the user interacts normally with the system, but the user input is transmitted to the "wizard" who is the experimenter, and who simulates the functioning of the non-existing service [8].

- The basis of any experiment is control. The more variables are controlled, the better. However, this does not mean that for controlling the variables they should be included in the experimental design. In this sense, the less complex the better. Therefore, all intervening variables must be controlled, or randomized, and only those of interest to the evaluator included in the experimental design.

- Experiments are never made isolated. It is a complex and costly technique, which only gives its full potential when made in a most planned and structured way. Therefore, think on any particular experiment not as the final part of a research program, but as the beginning of a series of experiments to iteratively find out the solution to the hypotheses. The basic concept of experiments is replication.

- Detailed design of all aspects of the experiment is essential. For complex research situations (e.g., many interacting factors), avoid complex designs as those including many experimental factors, and try to solve the complex problems in an iterative way, eliminating and controlling more factors in each step. Experiments are not the best method to handle the complexity of many applied usability evaluations, but used appropriately can provide response to surprisingly complex questions.

- Sampling of the participants in the experiments (usually called subjects) is an essential phase of the experimental method. Although it is very often neglected, a rigorous sampling is the best guarantee to have generalizable results from the experiments. In this sense, we do not consider these methods as different from those as surveys or observation.

7.2.3 Usual experimental designs

Experimental design is a very complex issue, for its great variety of procedures in very different areas, and a detailed description of even a few of them is beyond the scope of the present document. Therefore, we will only deal with the most basic ways of designing an experiment in our context.

As it has been said, the aim of experimental design is to highlight the effect of the experimental factor and avoid undesired effects by strictly controlling the influence of irrelevant variables. Underlying most experimental designs is the idea of comparing the performance of groups of subjects who have experienced different levels of the experimental factor. The influence of irrelevant subject-related variables can be controlled by randomly allocating the subjects or experimental units to the various conditions of the experiment.

Repeated Measures

The most direct way of controlling subject variables is to ensure that all groups of experimental subjects have identical characteristics. This seemingly impossible goal is achieved by simply using the same subjects in each experimental group. This is called a repeated measures design, i.e. one where repeated measures are made on the same subjects under all the experimental conditions. These are the most usual experimental designs in usability research, as in Psychology or Medical Research, because, all other things being equal, the researcher needs fewer subjects for controlling the effects of the various variables than with independent groups designs. However, there is also a disadvantage, and is that, in general, statistical analysis of data obtained with this type of designs is more cumbersome, and can make difficult to distinguish between hypotheses and alternatives.

The arrangement is illustrated in the figure below. For instance, subject1 experiences condition 1, 2, 3, etc. This is repeated for each subject and the data are then compared. Any differences between the two conditions should only result from experimenting the different conditions.

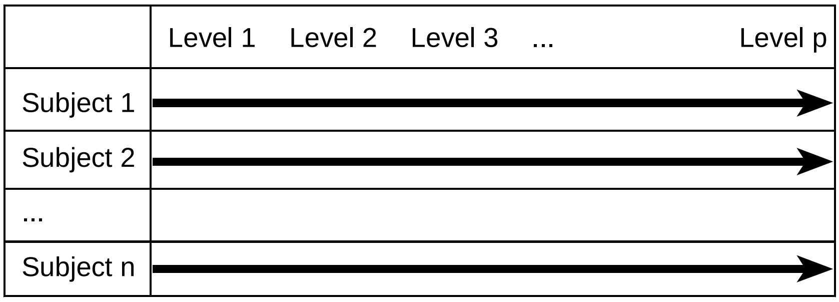


Figure 2: Illustration of repeated measures experimental design

The problem with the repeated measures design is that it is vulnerable to another irrelevant variable - the order effect. This effect refers to the influence on the response produced by the order in which the subjects experience each experimental condition. If for instance, the experimental conditions involve similar tasks (which they often do), then the subjects may learn from the first condition and achieve consistently higher scores during the second condition. This type of order effect is called a practice effect. If the experimental condition is quite long, for example an hour or more, then the subjects may become tired leading to consistently lower scores in the second experimental condition. This is called a fatigue effect. In order to control for order effects when using a repeated measures design a procedure known as counterbalancing is employed. This requires that half the subjects perform the experimental tasks in one order (i.e. A-B) and the other half perform the tasks in the reverse order (B-A). The advantages of any practice effects and the disadvantage of any fatigue effects are thereby allocated evenly to the two experimental conditions (ISSUE Usability Guidelines) [5].

Independent Groups

In the independent groups designs, subjects are allocated randomly to the separate experimental conditions. The random allocation of subjects to the experimental groups is the essential condition to reduce the risk of systematic bias. All things being equal, and in comparison with repeated measures designs, these designs require more subjects. Therefore, the more groups, the more people needed. This is one reason why these designs are usually kept very simple (few levels, few groups). Results obtained with these designs depend on the variability in each group: if this is high, then finding effects can be quite difficult; if it is low, then it is very possible to find them. Provided that the sample sizes are the same, independent group designs are known to be less powerful than repeated measures designs.

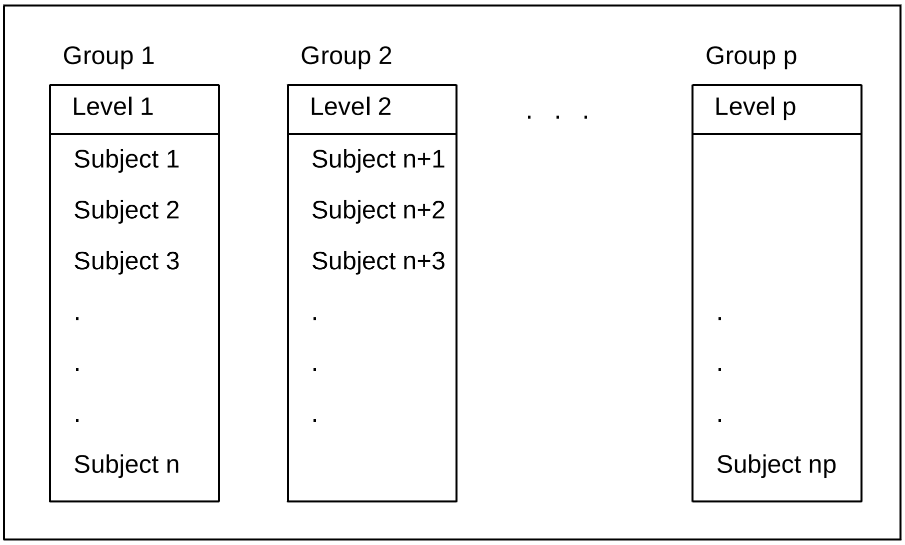


Figure 3: Image representing independent groups design

Choosing the appropriate experimental design

A major aim of all three experimental designs dealt with below is to eliminate systematic differences that are introduced by the subjects themselves. Knowing which experimental design is the most appropriate for a given experiment is largely a matter of experience on the part of the experimenter; there are no fixed rules. However in choosing an appropriate design, the experimenter has to make a number of key decisions. This process is summarized in the form of the flowchart that appears in Figure below.

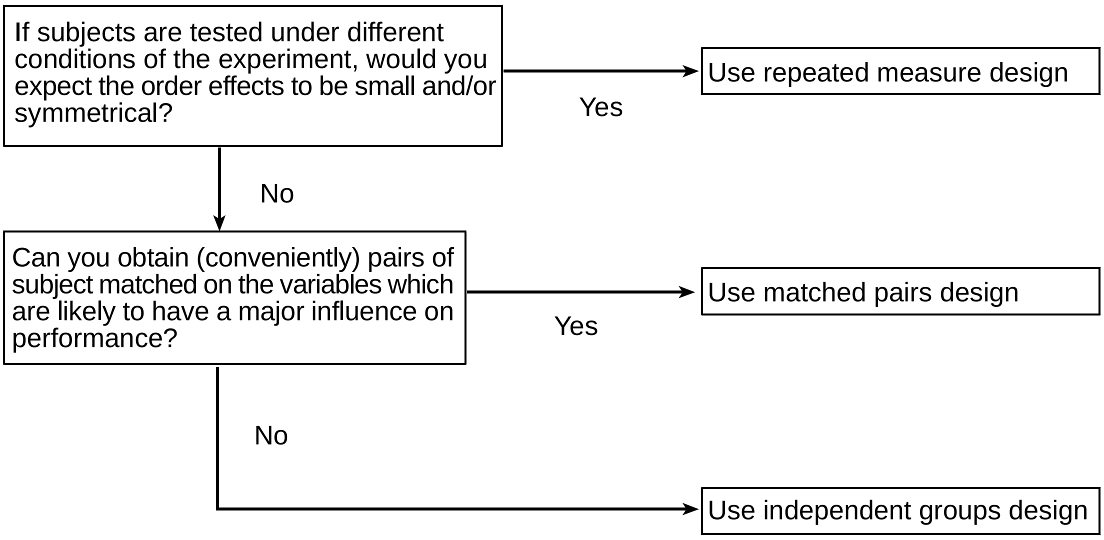

If subjects are tested under different conditions of the experiment, would you expect the order effects to be small and/or symmetrical?
Yes
Use repeated measure design
No
Can you obtain (conveniently) pairs of subject matched on the variables which are likely to have a major influence on performance?
Yes
Use matched pairs design
No
Use independent groups design

Figure 4: How to control subject variables (from Miller, 1975) [17]

Controlling Situational Variables

Experimental design does not end with the control of subject variables. The influence of the experimental situation may also undermine an experiment. As stated previously, the basic technique is to hold as many variables as possible constant in the experimental scenario. The situational variables include the physical characteristics of the experimental room or laboratory such as temperature, lighting, background noise and apparatus. Clearly once these variables are held constant, they are incapable of introducing either systematic or random bias into the experimental data.

A further source of situational bias may originate from the experimenters themselves. Most psychological studies rely on a set of standardized instructions (or 'script') to ensure that subjects receive identical information and any influence that may be due to the experimenter is minimized. There is evidence that experimenter characteristics such as age, sex and in general his or her behaviour may have subtle effects on subjects' performance.

These types of situational variables are not always apparent during the planning of the experimental trials. This is one of the many reasons why it is important to perform a pilot trial before introducing 'real' subjects into the experimental scenario. A pilot trial corresponds approximately to a 'dry run' of the experimental trial where the experimenters can familiarise themselves with the experimental procedure and apparatus/software may be tested for reliability. Some experimenters find it useful to take on the role of the subject during the pilot trial phase. This is a useful technique for detecting those situational variables that may not be apparent 'on paper'.

7.3 Field observation

Observation can be considered a fundamental aspect of any science. For instance, there would have not been Physics as we know it today without researchers observing the behaviour of physical objects of any kind. The same can be said in many other research areas, including ours.

First, we have to point out some important concepts:

Observation can also be considered a data collection method. For instance, the records taken by the usability expert when a user is testing a prototype. However, we are considering observation as a research method. In this sense, observation is a method in itself, which distinguishes from others in that it requires that the system is tested in its natural use environment.

Observation is **not** simply watching. Observation requires a structured plan, and in this sense, it has the same complexity than other methods, as experiments.

Observation is structured, planned and systematic.

Therefore, the main feature of observation is that it is made in the natural environment. All behaviours of interest, in all the different moments in time, and in all environments of interest, will be considered as the population of observation units. The first phase is precisely defining this population of observational units (Wilkinson, 1995) [18]. Then, there must be a process of event and time sampling, which consists of choosing a representative subgroup of these elements, and which will be the observation plan. Sampling techniques are perfectly applicable for these purposes.

A system to record behaviour has to be devised. Audio-video recordings are the best procedure, since they allow that different observers categorize the behaviours. However, there can also be other records, as those taken by observers (field notes), checklists, or ratings.

The following step is producing behaviour categories, which will provide the data for the analyses. This step is extremely important, since quantifying behaviour as a function of these categories will be the most time-consuming task of this procedure. Therefore, ensuring the validity of the categories is an essential point.

After the categorization of behaviours by at least 2 observers, tests should be made to find out whether there are differences because of the observers or because of the behaviours. Statistical procedures to test inter-observer agreement are usually found in books, but the most complete approach for this purpose is still Generalizability Theory (Cronbach, Gleser, Nanda and Rajaratnam, 1972) [19].

There are different approaches to observation (Wilkinson, 1995) [18], in particular non-participant (the most traditional approach) and participant observation. This last category means that the observer has some intervention in the situation (e.g., when the experimenter interacts with a user of a videoconferencing system).

Advantages

- It is a direct method.
- Since it should be made in the natural environment, observation provides data impossible to get in the usability laboratory. For instance, when testing a public telephone, people's reaction will sure be different in a controlled than in a public environment.
- There is no experimental or other manipulation.
- Useful when there are behaviours or situations on which subjects are unable to report with accuracy, such as, e.g., with children, or non-verbal behaviour.

Disadvantages

- Very costly and really difficult to carry out.
- Validity depends heavily on the categorization process.
- It can be biased by using different observers, who may produce different records, thus producing "the observer effect" (there are differences as a function of observers, thus making difficult making conclusions).
- It can be difficult to know the reasons why particular behaviours are made.

7.4 Heuristic evaluation

7.4.1 What is heuristic evaluation and when can it be used?

The goal of heuristic evaluation is to find the usability problems in a user interface design so that they can be attended to as part of an iterative design process. To make the evaluations we need a few evaluators that examine and make evaluations about the interface. This evaluations are based on usability principles (Heuristics). In this paper we use the Nielsen heuristics [8] [20] [21], but we can find other principles in the interface and usability guidelines. The task of the evaluator is generate a list of usability problems found in the interface, annotated with references to those usability principles that were violated by the designing each case.

A number of advantages are claimed for heuristic evaluation. They include:

- it is cheap;
- it is intuitive and it is easy to motivate people to do it;
- it does not require advanced planning;
- it can be used early in the development process.

A disadvantage of the method is that it sometimes identifies usability problems without providing direct suggestions to solve them The method is biased by the current mindset of the evaluators and normally does not generate breakthroughs in the evaluated design.

7.4.2 Recommendations for using heuristic evaluation

In the heuristic evaluation design the experimenter has to considerate the follow issues:

- In several investigations the results showed that only one evaluator performing a heuristic evaluation found only 35% of the usability problems in the interface so that different evaluators are recommended to be used. On average over 5 evaluators (at least three) are necessary.
- The results of one evaluator can not be known by the other ones until all evaluators finish their trial.
- The results of the evaluation have to be recorded while are generated (for example, saying aloud the problems found and the experimenter writing it).

- The experimenter cannot give more information that necessary (especially if non-expert evaluators are being used), because with this information the evaluator can infer actions and solutions about problems that the final user cannot.
- The evaluation session should not be longer that two hours.
- A general recommendation would be that the user go through the interface at least twice. The first pass would be intended to get a feel for the flow of the interaction and the general scope of the system. The second pass then allows the evaluator to focus on specific interface elements.
- When the heuristic evaluation is insert in the design process, is recommended conduct a session including the evaluators, experimenters and interface designers. In this session should be discussed the problems, solutions and redesign of the interface.
- Major individual differences are found between evaluators in the problems identification task, however evaluators experienced in usability and interface have better results that non-experienced evaluators.

7.4.3 A list of heuristics

These usability principles should be followed by all user interface designers. This specific list was developed by Jakob Nielsen and Rolf Molich [21], and can be easily applied to a large range of different systems:

- Simple and natural dialogue: Dialogues should not contain information that is irrelevant or rarely needed. Every extra unit of information in a dialogue competes with the relevant units of information and diminishes their relative visibility. All information should appear in a natural and logical order.
- Speak the users' language: The dialogue should be expressed clearly in words, phrases, and concepts familiar to the user, rather than in system-oriented terms.
- Minimize the users' memory load: The user should not have to remember information from one part of the dialogue to another. Instructions for use of the system should be visible or easily retrievable whenever appropriate.
- Consistency: Users should not have to wonder whether different words, situations, or actions mean the same thing.
- Feedback: The system should always keep users informed about what is going on, through appropriate feedback within reasonable time.
- Clearly marked exits: Users often choose system functions by mistake and will need a clearly marked "emergency exit" to leave the unwanted state without having to go through an extended dialogue.
- Shortcuts: Accelerators -unseen by the novice user- may often speed up the interaction for the expert user such that the system can cater to both inexperienced and experienced users.
- Good error messages: They should be expressed in plain language (no codes), precisely indicate the problem, and constructively suggest a solution.
- Prevent errors: Even better than good error messages is a careful design that prevents a problem from occurring in the first place.
- Help and documentation: Even though it is better if the system can be used without documentation, it may be necessary to provide help and documentation. Any such information should be easy to search, be focused on the user's task, list concrete steps to be carried out, and not be too large.

7.5 Focus groups

The focus group can be considered a "discussion-based interview" (Millward, 1995) [22]. Following this definition, it involves multiple respondents who are "focused" (i.e., discuss about an external stimulus), and relatively staged by a moderator.

The focus group began with research into the effectiveness of wartime propaganda, and the social effects of mass communication. Its evolution was made by marketing consultants, for whom focus groups have become central to answering the question of "why" consumers behave as they do.

Although it is considered the quickest and most cost efficient mean of generating consumer relevant information, this is yet to be proved. Currently, as Millward points out, "this method has evolved into a 'quick and dirty' means of fulfilling client needs rather than as a sophisticated research tool".

Nowadays, after adapting the method to different purposes, it has become more popular in other behavioural or social sciences settings, as, e.g., health psychology and marketing research, areas in which this method allow to geta great deal of information, for instance, about the issues involved inmarketing a product. Nielsen (1993) [8] cites this method in his list of methods for usability assessment.

7.5.1 Advantages and disadvantages of focus groups, and applications

The focus groups not only enhance the ability of the usability evaluators to answer research questions, but, and perhaps more important, they can generate questions from new angles and perspectives (Millward, 1995). It has to be noted that it is currently the only method in which people interactuate, and in which the opinion of a group (people interacting) can be studied. One characteristic advantage of using the group as opposed to the individual as the medium of investigation is its "isomorphism to the process of opinion formation and propagation in everyday life", because "opinions about a variety of issues are generally determined not by individual information gathering and deliberation but through communication with others".

This method, however, depends heavily on how rigorously is conducted. There are several issues to guarantee its quality control". In particular, it is important to recall that the aim of focus groups is to get closer to participants' understandings of and perspectives on certain issues. It is not geared to the formal testing of hypotheses in the traditional hypothetico-deductive sense. In a rigorous sense, it can be considered a method of data collection or a supplement to other methods.

This technique is especially applicable when:

- New and different services or systems, to find out first impressions of the service.
- When population is not specialized, and the target is the normal population.
- Marketing issues. To find out the attributes, opinions, attitudes, etc. that can be used to produce a successful marketing campaign.
- Attitude change. When there is the need to know which aspects can be changed in a product.

Special attention can be paid to the use of focus groups together with surveys: used together with surveys , the focus groups can be used for testing results coming from the surveys, or can provide the issues which will be tested with the questionnaires. "They can establish the variety of opinion concerning a topic, establish relevant dimensions of attitudes and identify relevant indicators for the constructs being measured. Focus groups can also ve used to test various questionnaire items for readability, comprehension, wording, order effects and response variation.

Once the survey has been conducted, there are other useful applications: first to assess respondents' reaction to the survey and in particular to trace the cognitive and social processes involved in answering, and second, to aid in the interpretation of survey findings by exploring in greater depth the implications of certain quantitative patterns and relationships.

7.5.2 How to conduct focus groups

The number of participants is, in average, of nine participants per session, with a conventional range from six to twelve. There is evidence that group size is inversely related to the degree of participation. There are more reasons to keep the size small, and they are that large groups are difficult to manage and monitor, and there is the risk of break into subgroups. Sessions last between one and two hours.

The character of "focused" is provided by a particular stimulus object, event or situation. In the context we are talking about, the stimulus might be a scenario for using a service, a particular system, or even a concept (a system or service not yet developed).

The group should be chosen on theoretical grounds as reflecting those segments of the population who will provide the most meaningful information in terms of the project objectives. But, as with any other usability testing method, the concept of "representativeness" of the group of participants is essential.

This particular method largely relies on the moderator style and skills. Moderation is all about process facilitation and what is called "listening and questioning skills". Central to this is the concept of participant empowerment. Basically this means that the moderator is the facilitator of some else's discussion.

Data are collected in the form of transcripts from audio or video tape. With video recording of the sessions, observational data can also be extracted.

There are very few data analysis techniques for this technique, in particular content analysis, or the categorization techniques used in observation.

7.6 Input logging

Input logging consists on the use of a recording device embedded or attached to a system or users to capture and record the users' performance.

The main advantages of input-logging method are:

- the method can be applied on a large sample of users;
- it can be used over a long period of time (longitudinal);
- it does not require the presence of the evaluator. It is accurate;
- it can capture low level and detail data.

So, a large amount of detailed data can easily be obtained in a unobtrusive manner.

The disadvantages are:

- the method may affect system performance (e.g. increased response times);
- it is unfocused, and a large amount of data captured makes analysis difficult;
- if contextual information is not captured, interpretation is difficult;
- the set up and the analysis may be time-consuming if no automatic system for analysis is available.

So the interpretation of the results from logs may be difficult, due to the lack of contextual information.

With input logging the data can be automatically collected for a large number of users without the constant presence of an evaluator. This makes it ideal for longitudinal studies particularly in field settings.

The large amount of data captured makes it critical for the evaluator to have clear and focused goals and if possible to avoid 'data trawling' by capturing too much data. Analysis of such potentially large amounts of data can clearly be improved through the use of relevant statistical packages.

In any case, it is recommended that logging is always used in conjunction with other methods (like questionnaires or interviews). These other methods are necessary to ensure that sufficient contextual information is gathered, which will facilitate the analysis and ensure greater validity of the results.

It should also be noted that experts with specialist skills in computing will be required to write and embed required software.

Logging model

Considering conditions and precautions of use, a logging model is proposed to minimize the disadvantages of input-logging technique.

The logging model design is guided by the following intentions:

a) To reduce the amount of data and to focus the analysis:

- All the data captured will be events at the semantic level; that is to say that the capture will be performed at the application level (not at the window manager level); for example, it will be capture the event "ask for help", not the window manager event "mouse click on the button located at x=10, y=100".
- A set of appropriate capture points will be selected to allow the analysis of main user tasks.

b) To minimize the performance reduction.

- An event model at the semantic level (included in the present document) reduces, classifies and codifies the events to be captured.

c) To facilitate the data analysis.

- An analysis tool would take advantage of the referred event model to facilitate data analysis and interpretation.
- All the logged data won't be associated exclusively with user actions; information introduced by users, the system state, and written comments introduced by users during real operation, can also be logged; this extended input-logging will allow to capture some contextual information or other data required by developers.
- A written daily log of the system use (not embedded in the logging system) should complete the capture of contextual information (like working hours, holidays, other problems, etc.).

Event model

All the events that can be captured can be classified as:

Single event: events that occur in a short time; for evaluation purposes, a single time can be assigned to these events.

Double or continue event: actions that occur during a significant period of time for the evaluation purposes; two single times can be assigned to these events:

Start event: time in which the user begins to perform actions associated with the continue event.

End event: time in which the user performs the last action associated with the continue event.

Every continue event can be successful, when the user managed to reach the goal (for example, pushing an "OK" button), or unsuccessful, when the user did not manage to reach the goal (for example, pushing a "Cancel" button); although, each start event can have many different end events, we consider as continue events only those that finish successfully, and furthermore we consider that only one successful end event exists for each continue event; this decision will focus the analysis on goals achieved, however we will be able to measure unsuccessful tasks by measuring incomplete continue events.

User information event: sometimes the information introduced by users is important to evaluate or characterize some aspect of the system use (like dates, etc.); this event is equivalent to a single event but it also includes the information introduced by the user.

System state event: sometimes the state of a system component is important to evaluate or characterize some aspect of the system use (like document size, etc.); this event is equivalent to a single event but it also includes the system state.

User comment event: a user comment about the system use, obtained in real time, can provide essential information to interpret the data logged or to discover usability defects; this event is equivalent to a single event but it also includes a written user comment.

It is important to remark that only the logging of single events and continue events is mandatory to be considered an input-logging.

On the other hand, a specific application (embedded or called by the application), must be developed if the user comments will be captured during real time operation.

7.7 Surveys

Survey is a research procedure more usually associated with Sociology or Political Science. However, it is also a common technique in Marketing Research, and also in Usability Engineering, although it is not as popular as other methods, since surveys are usually very costly, and not applicable until there is a working system, thus not being useful, in general, in previous phases of the development process.

A survey consists of delivering an instrument of data collection, usually a questionnaire, although it can also be an interview, to a sample of the target population. Many survey designs are longitudinal, i.e., they are made along time to track the change of opinion in the population. Other technique, developed in Marketing, is panel research. A panel is a sample of the population who participates in research during a period of time. Perhaps the most famous application of these methods is the TV audience measurement. Panel research, however, can be very problematic, since there can be many other factors affecting the behaviour, e.g., whether there is an economic benefit in participating in research or not.

The most important stage of any survey design is the sampling. A sample is a representative subset of the target population, and who are chosen using random procedures, trying to guarantee that all components of the population have the same probability of coming into the sample. Thus, these procedures usually require having a census of the population. Sampling procedures are quite complex, and beyond the scope of this tutorial, but the interested reader can consult Tryfos (1996) [23], as a good introduction.

Surveys in usability engineering can be used to obtain user feedback from using the system. Requirements for a successful application are, then, having a working system or service, and that participants actually use the system.

Below you can find a table of advantages and disadvantages (Fife-Schaw,1995) [24]:

Advantages:

- Sampling methods, if appropriately applied, can yield very high-quality results about the attitudes and opinions of a population.

- Easy to repeat, thus providing results along the time.

Disadvantages:

- Although it depends of how the sample is approached (e.g., it can be a panel, or it can be used a mail questionnaire, or interviews), getting all planed people to participate in the investigation is very difficult, and usually very costly.

- Sampling is usually very complex, and there will always be groups of the population impossible to reach.

- The instrument has to be fixed for all the participants, thus usually making difficult the study of particular cases.

7.8 Questionnaires

7.8.1 What are questionnaires and when can they be used?

A questionnaire is a set of written questions requiring a written response which describes past behaviours, the user expectations, attitudes and opinions towards the system (ETR 095 [1]).

Questionnaires, whatever its form, have the following advantages:

- They are cheap and easy to apply to large samples of users.

- Can quickly provide both quantitative and/or qualitative data.

But usually have these disadvantages:

- Questions are fixed: there is seldom the possibility to include new questions on request from the respondent, and they cannot be explained in more detail in a standardized way. When questions require clarification, the evaluator should be present, and help the respondent, but in such a way that it does not make any influence on the subject's opinion.

- The evaluator cannot always control the situation or the manner in which the questionnaire is answered.

- As with any other research technique, misrepresentativeness of the sample may produce wrong results.

In general, we can recommend using a questionnaire in usability evaluation for the following reasons:

- to assess, in a standard and formal way, subjective judgements, attitudes, opinions or feelings about the usability of all or part of an existing system, or a prototype or release version of a system.

- to check the acceptance of the total system, usually within the user's normal operating environment.

- Questionnaires can also be used to measure subjective responses in an experimental context.

7.8.2 Deciding the question format

The choice for a specific question format depends on the testing aims, and for the precision with which it has to be measured. Very important issues about the level of measurement achieved with the different formats will be considered below and in the point about data analysis.

7.8.2.1 Open ended questions

These allow the respondent to write their answer in their own words and do not constrain them to choosing from a set of fixed alternatives or a straight yes/no reply. They allow the user to express views which have not been considered by the investigator. With this format, the user can express his or her opinions with complete freedom. The usual trade-off between flexibility and precision makes this way of asking questions very difficult to analyze, and statistical analysis is usually very difficult (content analysis), or even not possible.

In addition it is very easy for respondents to misinterpret the question, especially if the questionnaire has been administered remotely. For these reasons open-ended questions are typically less reliable than closed questions and are not recommended beyond a scoping exercise to determine response ranges or in situations when having opinions not already covered in the other questions is critical.

7.8.2.2 Multiple choice items

With these questions, users are offered a closed set of options, and he or she has to mark (usually only one) of them.

This question format offers significant practical advantages over open-ended questions: the level of detail is determined by the investigator; the questionnaire is usually quick to complete and quick to analyse; the limited response range makes quantitative coding by non-experts possible and easy.

The disadvantages largely relate to the limited response choice. Respondents may be forced to choose between alternatives which do not correspond to their preferred answer for the question, forced to choose a single answer when more than one is appropriate or forced to choose an answer when they do not have enough knowledge to make any informed response. The problem is particularly significant with dichotomous forced choice questions (Yes/No; True/False; Agree/Disagree etc.) where any misunderstanding of the question can change an answer from one extreme to the other.

This type of questions should be used only when the number of possible responses is limited or when there is a fixed number of options the researcher is interested in. Its level of measurement is nominal: this means that the experimenter can only make conclusions about the presence or absence of different responses, it is not possible to make a scale of the options included (but for the percentage of responses for each alternative), and the statistical procedures should be those appropriate for this level of measurement (see the data analysis chapter). Whenever a scale of preference (or any subjective construct) is to be made, more adequate techniques are rating scales.

Caution should be taken with the "don't know" response: this can be useful in some situations, but can be an enormous problem when many participants tend to choose this option, thus having no useful responses. Then, the usual recommendation is to include this option only if it is especially interesting for the testing purposes.

7.8.2.3 Rating scales

Measurement of psychological constructs is usually known as *scaling*. Scaling, and specially judgement scaling, can be dealt with in several ways, which will determine the measurement scale and precision of the resulting scales.

Usability testing by means of rating scales may be more vulnerable to error and bias than other types of responses and their results can suggest a greater degree of precision than is true. Most mistakes are made in the design phase, since the analysis phase is fairly standard. Very specific instructions must be made explicit for really being able to extract useful conclusions.

As far as the measurement level is concerned, scales obtained can be qualitative, or nominal, where the numbers are only codes of some characteristics, but no other properties like order or differences are maintained; ordinal, where order is preserved, but not differences between numbers; and interval scales, where both order and differences between numbers are preserved. This important issue distinguishes different types of rating scales we will review in more detail.

7.8.2.4 Nominal or grade scales

These scales are based on the choice of the level of the subjective attribute being measured made by the subject from a limited set of alternatives, marked with words indicating different grades of the subjective construct being measured (e.g., perceived image quality). Two main requirements have to be fulfilled when we want to use rating scales:

- The stability of the attribute to be measured. That is important since the opinion changes accordingly as a phenomenon changes.
- Divisions between the ranges of opinion. In grading scales, each grade represents a *range* of opinion, and these ranges can be narrow or wide in such a way that the division between them is the fact which in principle might be evaluated, rather than the grades themselves.

Although these scales may not be under complete control, especially due to the fact that words have different meaning across subjects, a well-designed opinion rating experiment ensures a high degree of stability over long periods of time.

It is better to minimize verbal descriptions, leaving the observer as much freedom as possible to establish his own standards. The following example of quality scale (ITU-T BT500-7 standard [25]) shows the words within brackets as a suggestion rather than rigid descriptions.

A Excellent

B Good

C Fair

D Poor

E Bad

The reason of using letters instead of numbers is to avoid that observers treat the scale as one of apparent magnitude. However, the investigator can also use numbers if some explanation is given previously about how respondent should treat them. Very often, respondents avoid extreme points, when their opinion is really that one.

The level of measurement is theoretically nominal, so statistical techniques to be used are restricted to those not assuming higher levels of measurement. However, some statistical techniques are known to be robust to deviations of this assumption, and can be used for scaling purposes.

7.8.2.5 Rankings

This technique involves asking respondents to rank a number of alternative items in order of importance or relevance or against a dimension such as 'effectiveness' or 'ease of use'.

Although the concept is readily understood and appears to give clear cut answers with little risk of misinterpretation by the investigator there are significant drawbacks. The most important one is the level of measurement: although items may be ordered there can be no assumption that the rank differences are equivalent. The ordering can only be interpreted in a relative fashion i.e. 10 items may be ranked in order of importance but they may all be thought to be trivial by the respondent. As the number of items increases the difficulty of the task increases and the reliability of the results declines. On balance, ranking is a less satisfactory technique than the use of rating scales. Additionally, statistical analysis of ordinal data is usually quite difficult, since there are very few techniques to perform it (non-parametric techniques).

7.8.2.6 Discrete numerical rating scales

Also known as Likert-type scales, numbers (1-5, 1-7, 1-9, etc.) denote the scale divisions and the intervals are assumed to represent equal intervals of magnitude of some measure. They simply propose a numerical range with anchors in the extremes, but without marking the intermediate points. Most usual ranges are 1 to 5 and 1 to 9, although this is very arbitrary. However, it has been shown that using more numbers increases the complexity for both the researcher and the respondent, and it adds no more precision.

The most important feature of these scales is that, if some conditions are met, can be considered as interval level of measurement. This means that you can make conclusions not only about the ordering, but also about the quantitative differences between conditions, just in the same way as with temperature. The conditions are that respondents should be very clearly explained how to use the scale. Then, including instructions like those in the example is very important.

These scales are very widespread in many fields, are very easy to respond by subjects, can be analyzed very quickly, and, provided that you follow the recommendations to design them, provide very good results.

7.8.2.7 Comparison scales

This is a variation of numerical scales. In them, the subject has to make a comparison between two tasks or conditions according to an attribute (for instance, quality). This procedure is similar to the rating method known as "paired comparisons" in which a number of conditions are compared in all possible combinations, but is far more economical. The following discrete numerical scale is the usual one. Note that it is centred in zero, which is the reference point.

Both the above type of scales are vulnerable to types of systematic bias. In the 'halo effect' an individual's overall response to a system influences their response to its individual aspects. Thus if they decide that they don't like a particular system they will give consistently poor ratings for ease of use, help facilities, dialogue structure, command names etc. without considering each aspect on its own merits. Similarly, if all the rating scales are laid out with the same orientation (i.e. the 'positive' end of the scale is always on the left or right) then the respondent may tend to move down the page making marks in a vertical 'column'. A simple way of this problem is to randomise the direction of the scales so that the respondent is forced to consider each scale separately.

7.8.2.8 Graphical scales (Continuous rating scales)

They consist of an unbroken line without divisions, representing points on a continuum, or with only one representing the medium point. There also exists the possibility of including several points in the line, each point being defined by a trait label, definition or adjective. However, this is not usual, for all the problems with people's different understanding of the verbal concepts.

They are preferred to grading scales because these ones would introduce appreciable error due to the use of numbers for their analysis. Subjects must be trained on the used of the scale prior to the experiment. They usually express the ratings by means of a small mark on the line in the point between the extremes which best reflects their opinion. Several studies in the most diverse fields have found that consistency and reliability of these scales are very good, but its use will require that several conditions must be assured to preserve the advantages: clear instructions, previous practice with non-experimental items, and explanation of objectives of the measurement and scale are very important to achieve these objectives.

7.8.3 Practical issues in questionnaire design

These recommendations are extracted from the ISSUE Usability Assessment Guidelines [5]:

Question wording

The wording of individual items is a critical aspect of a questionnaire's validity and reliability. They should aim for the simplest wording possible while still conveying the intended meaning. They should not be ambiguous or require any inferences to be made by the respondents. Care should be taken to ensure that the questions are 'neutral' i.e. that they do not imply an expected or "correct" answer. If a question depends on the respondent's memory then the time period in question should be clearly defined. The goals should be clarity, simplicity and intelligibility.

Question length

Lengthy questions should be avoided but where necessary it is better to try to break a long complex sentence into shorter sentences. It may be possible to have an introductory sentence to set the context followed by a short sentence which poses the question. In general aim for twenty words or less per sentence.

Leading and loaded questions

As stated above, great care should be taken to avoid leading or loaded questions. Loading can take the form of indicating a majority response, explicitly referring to one answer rather than another or by citing a prestigious group or individual.

Positive and Negative wording

It is recommended that questions are worded positively. They are preferred by respondents and are understood more easily. Double negatives should never be used, because it has been demonstrated that they introduce great confusion in subjects.

Selecting modifiers for response alternatives in rating scales

e.g. Easy/Difficult, Frequently/Infrequently, Friendly/Unfriendly

- When adjectives or adverbs need to be used to generate response alternatives, it is important that they are understandable and also represent equidistant points on a continuum to the respondent.
- If possible, descriptors in rating scales should reflect the question wording i.e. if the question concerns service usefulness then the descriptors should be "very useful, of little use" etc.
- Descriptors from different continua e.g. Difficult, Frequent should never be mixed.
- The term "average" should be avoided since it is always a relative valuation (the average performance of one group may well be above or below the of another).
- Alternatives should be balanced whenever possible. i.e. if the term 'easy' is used then the contrary term 'difficult' should be used. The simplest way of generating a set of balanced descriptors is to choose a term and its literal opposite (e.g. effective and ineffective), add a modifier to mark the extreme options (i.e. very effective and very ineffective) and then provide a neutral midpoint.
- If respondents are being asked to state a frequency (e.g. Frequency of system use) it is better to provide quantitative examples than purely verbal labels (i.e. every day, once a week, once a month rather than 'frequently', 'sometimes' or 'occasionally') since frequency phrases are interpreted with great variability.

Question sequence

Question sequence is not often a serious issue and is more often a localized problem affecting a group of questions addressing the same subject rather than with the order of different groups of questions each addressing separate topics. The problem arises when an initial question influences the way in which the respondent answers subsequent questions because it establishes a specific, and limiting, point of view. This problem can usually be avoided by way of careful piloting.

Filter questions

This type of question may be may be placed at the beginning of a section to ensure the relevance of subsequent answers. Thus a filter question might ask the respondent to describe their experience with a particular system before asking further questions about problems encountered. If the individual has no relevant experience then they can be directed on to the next set of questions. If this is not done then some respondents will quite happily give answers on the basis of no direct experience at all!

Piloting

This is best carried out by asking a typical respondent to read each question and then explain their understanding of its meaning, the response alternatives if appropriate, and the reasons for their answer. The respondents should be encouraged to comment on anything that appears to be unclear or surprising about the content, format or layout.

After pre-testing, each question should be reviewed. If a high proportion of answers are of the "Don't know" variety then serious revision may be required. Any question which does not add significant information or simply duplicates the results of another should be excluded response rates tend to decline as questionnaires get longer, so only ask those questions which are needed.

7.9 Interviews

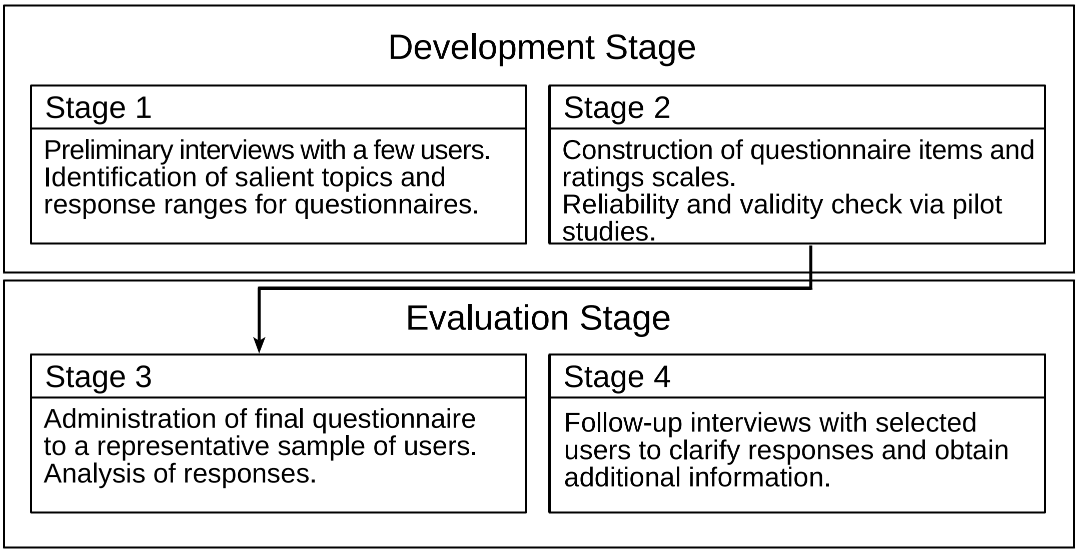
7.9.1 When should interviews be used in usability evaluation?

Interviews are a very useful procedure in the user requirements phase, and in the follow-up stage. In general, they are recommended for situations where a great flexibility is required, but not when you want to generalize to a population of users, since they don't allow the required formality.

Structured interviews (e.g., those following a script) are better considered. The information from these interviews should be used for preparing other data capture techniques, especially for questionnaires.

Although interviews may be capable of discrimination from questionnaires in terms of their degree of formality they should not be considered as less important.

Instead they should be used in a manner that makes best use of their respective strengths. Rubin (1981) suggests the following sequence of stages in a usability evaluation:



Preliminary interviews with a few users. Identification of salient topics and response ranges for questionnaires.

Construction of questionnaire items and ratings scales. Reliability and validity check via pilot studies.

Administration of final questionnaire to a representative sample of users. Analysis of responses.

Follow-up interviews with selected users to clarify responses and obtain additional information.

Figure 5: Stages in a usability evaluation (Rubin, 1988)

In Rubin's model, interviews are used at the beginning and end of an evaluation; initially to gather general information to form the basis for a questionnaire and after the questionnaire to clarify its results and filling gaps. However there are no absolute rules for the use of questionnaires and interviews, as with much human factors research it depends very much on the circumstances and the type of system being evaluated. In many respects, within the context of a usability evaluation, the choice between structured interviews and directly administered questionnaires is likely to be a matter determined by cost and convenience. The evaluator has to make a decision based on knowledge of the domain and of the advantages and limitations of the two techniques as discussed in the following sections.

Interviews designed to elicit information are used in a variety of usability evaluation contexts, e.g.:

- as part of a user requirements analysis in a specification exercise;
- follow-up interviews after an experimental trial;
- to investigate usage problems in a newly implemented system.

The type of interviewing carried out within a usability evaluation context can be compared to the survey interviews undertaken to investigate social attitudes or voting intentions. In these interviews the format will be highly structured, i.e. the questions will be very carefully phrased, presented in a standardized manner and an invariant order. In addition, the respondents' answers are often limited to a restricted number of alternatives since the interviewer is frequently seeking specific answers to well defined research questions. This approach allows the use of sophisticated statistical analysis and the accurate prediction of population trends and intentions without requiring highly skilled staff. The technique's major limitation is its inflexibility; the investigator's conceptual framework is imposed and this may restrict important, but unexpected, information from being offered by the respondent.

In contrast, the style of interviewing employed in usability research is usually much less formal. The sample of interviewees is typically considerably smaller and more tightly defined and the interviewer is likely to require training in interviewing techniques as well as an understanding of the domain. Whilst there is often a standard list of questions, the interviewer may well ask supplementary questions to explore an issue, not usually a possibility with survey interviews. The interview style is intentionally conversational in order to put the interviewee at ease and encourage as much comment as possible (the interviewer is often as interested in the minority comment as in the majority view). The apparent lack of formality does, however, place an extra demand on the interviewer to ensure that the necessary issues are covered in adequate depth and the interaction does not lose its focus. This type of interview can be a highly effective way of establishing basic user requirements at an early point in the design cycle or exploring usability problems with a new design or prototype. The results of these exploratory interviews can then be used to inform more specific evaluation exercises (see figure above).

There are a number of limitations in the use of the interview technique for elicitation of task information. The first problem is that an interview relies on human memory which is frequently inaccurate. In addition memory for task information is likely to become more unreliable if the person is asked to recall it away from the testing place. Finally people often find it difficult to describe tasks which are carried out physically or which they have carried out so often the task has become semi-automatic, e.g. it is very difficult for an experienced driver to explain how to drive a car.

7.9.2 Guidelines for carrying out interviews

The following recommendations are standard interview guidelines (ISSUE Usability Evaluation Guidelines [5]).

Social context

- **Inform the subject**: Try to put the subjects at their ease and begin the interview with a brief summary of who you are (avoiding jargon and unnecessary technical detail), what information you are trying to collect and what it will be used for. The better informed the subjects are the more involved they will feel and the more likely they are to provide the co-operation required.
- **Assure Confidentiality**: assure the interviewees that any information volunteered will not be passed on to other persons. This is particularly true if the interview is to be tape-recorded. In this case, the interviewee's consent should always be requested first. (ethical / legal implications of usability assessment)
- **Be non-evaluative**: assure the interviewee that it is the system's performance that is important and that you are not particularly interested in anyone's personal performance: try and make the interviewee feel like a co-evaluator. It is helpful to assure the subjects that their contribution is important even if their comments or views seem trivial to them.

- **Relative numbers**: It can be helpful to organize group interviews or discussions particularly when usability issues/requirements are not defined. A great deal of information can be elicited and it is sometimes necessary to have an additional note taker in attendance. If only one person is being interviewed it is better to have only one interviewer.

Planning

- **Advance preparation:** despite the apparent informality it is essential to plan interviews carefully.
- **Ordering questions**: Group questions according to the major issues to be investigated. The sequence of groups should seem logical to the interviewee and should reflect, for example, the structure of their own work.
- **Time table**: Determine a maximum duration for the interview (up to 1 hour) and then allocate time accordingly. Write this information on the interviewer's schedule. This will help prevent all the time being taken up on the first few questions.
- **Form of questions**: Questions should be open ended (How?, Why?, When? etc.) rather than closed (Do you?, Is it true that? Have you ever? etc.). Closed questions can be answered with a simple "Yes" or "No" which give very little information. For example, ask "What problems do you encounter when doing ....?" rather than "Has the machine ever failed when you have been using it?". Questions should not indicate a preferred answer (e.g. 'Some experts believe that would you agree with this?')
- **Level of questions**: Make sure that your questions are appropriate for the interviewee concerned. i.e. don't ask operators about scheduling policy or managers about operator job satisfaction.
- **Pilot before use**: An interview should be treated like any other experimental 'tool' and should be tried out before it is implemented for data collection purposes. This should involve collecting and analysing sample data. This precaution should determine whether the questions can be fully understood, whether they are liable to misinterpretation, and whether they will produce comprehensive answers.

Implementation

- **Location:** It is better to conduct the interview at the interviewees' normal workplace or desk since their 'home ground' is likely to be less threatening. However, if the content of the interview is potentially sensitive (and privacy is necessary) or there are likely to be frequent interruptions or significant background noise, etc., then a quiet room nearby may be required.
- **Manner**. Try to be businesslike rather than either overly familiar or too formal. Allow the interviewees to express themselves without interruption but ensure that they are reminded of the question if they stray 'off the track'. A moderate degree of positive reinforcement, uniformly distributed, can significantly improve the amount of information volunteered by the respondents. This only requires the interviewer to appear interested, make comments such as "good" and "fine" after the respondent's answers and smile and nod appropriately. Finally, remember to thank the interviewees. You may need to contact them again to confirm a point.
- **Records**: Take notes throughout and do not rely on memory. Although tape recording may seem an attractive support it may inhibit some interviewees, can be unreliable and require considerable transcription effort afterwards. It does however have some use as a back-up to interview notes, particularly if the interviewee speaks rapidly. It is important to gain the informed consent of the interviewees before recording them
- **Remain detached**: Avoid the temptation to argue or offer your own opinion. If there are apparent contradictions in the interviewee's account do not point them out; it is better to say that you have not quite understood and ask them to clarify the issues. Try to ensure that you discriminate fact from opinion. If the subject matter is procedural or technical it is useful to have paper and pencils available so that the interviewees can draw a sketch if they wish.
- **Afterwards**: Try and 'write up' the interview from the notes as quickly as possible. Make a note of any internal inconsistencies and any new issues which should be raised with subsequent interviewees. It may be appropriate to send a copy of your summary to the interviewee so that they can check it for accuracy.

Analysis

The analysis of qualitative data is, in contrast to the analysis of quantitative survey data, much less statistically oriented. This is not only because the size of the respondent sample is often relatively small, but also because the purpose of the exercise is not to predict the behaviour of large groups of people. Instead the information is used to extend and interpret data collected from objective tests i.e. to suggest trends or pick out common areas. The investigator may be able to count the number of respondents expressing particular attitudes and categorise them but often-even measures of variability will be inappropriate. The data analysis is therefore very much content oriented and descriptive.

7.10 Performance measures

It is possible to obtain quantitative estimates of performance, behaviour, events and phenomena through objective measurements. What makes a measurement objective is the absence of interpretation in recording the data. The interpretation of data comes later, but it is not critical part of the measurement process itself.

The different type of performance measures which can be used are:

I) Frequency measures, for example:

- How often a particular input device was used (e.g. mouse versus cursor keys).
- How often particular function keys were used (e.g. "F6" for help or "R" for reload).
- How often certain types of tasks were carried out (e.g. number of local versus remote file transfers) or applications used (e.g. word processor or spreadsheet).
- How often certain types or errors occurred.

2) Time measures, for example:

- The time between two or more actions (e.g. between different keystrokes).
- Total time to complete a task (e.g. to dial a set of digits or add a set of numbers)
- Time spent on supplementary tasks (e.g. using "Help").

3) Error detection measures, for example:

- Wrong keystrokes (which?).
- False or incomplete data entry.
- Use of wrong or inappropriate procedures.

4) Physiological measures, for example:

- Amplitudes and latencies of event related potential components (e.g. P300).
- Heart rate variability.

Advantages

Lack of subjectivity of observation and judgement which leads to questions of validity and reliability.

Disadvantages

They may provide information less related to global user satisfaction than subjective measures.

Certain contexts do not permit objective measurements because the task performance must be overt. Perceptual and cognitive activities must be self reported and dimensions of task and jobs, such as team co-ordination, the objective measurements may be inadequate because they are not designed to measure attributes.

In the following subclauses you will find some of the most usual performance measures in usability testing:

7.10.1 Reaction time

Reaction time (RT) is the time between the occurrence of an event requiring an action on the part of the user and the start of the action demanded by event. The major purpose of measuring RT is to determine how quickly the user can react to an initiating stimulus.

Reaction time is meaningful only when there is a system or task requirement to react as quickly as possible (for example when the user is prompted to notice, or correct, an error).

In general, the shortest time in which an user can respond to a single, discrete stimulus like a light is 200 to 300 milliseconds, but as the stimulus becomes more complex and requires difficult cognitive processes, that minimum period can become longer. When the user response is covert, it may be difficult for the experimenter to know when the user has made his response and this time usually is the addition of several user activities (perception time, process time, decision time, muscular activity time, etc.)

7.10.2 Duration

Duration is the time from initiating stimulus to the time the task or function is accomplished. It therefore includes the RT measure and extends it. Duration is an extremely common measurement and it is important when the system prescribes a maximum duration for a task or group of task.

Ordinarily duration measurement does not have to be extremely precise. Task in a series may flow into each other and there mat be no clearly defined start and stop points to bounds the limits of the measurement. Duration time usually include the user time and the system time.

7.10.3 Accuracy

Accuracy, or its converse, error, is probably the most common and perhaps the most useful measure of personnel performance. There are systems and task in which RT and duration are not important, but accuracy is critical in all.

Some error will always occur because of the inherent variability of the human. In consequence, error data are usually meaningful only in terms of how the system is affected by error, and only in relationship to the number of opportunities to make the errors In the operational environment accuracy/error data are important primarily for diagnosis of problem. An excess of error may be indicative of a design problem.

7.10.4 Frequency

Frequency determine how frequently the user's responses occur or how frequently certain task are performed. frequency is occurrence as a function of same time interval. In systems that make commercial products frequency of outputs is important because determines productivity, but is much less important when the system is performing a mission.

The relative frequency of certain types of error may suggest special difficulties the user has in using the system.

7.11 Thinking aloud

A thinking-aloud test involves having a test subject use the system while continuously thinking out loud. (Lewis 1982) [27]

Verbalizations allow the researcher to understand how users are interpreting the interface. This allows the detection of the main user errors and those parts in the interface that are more problematic. This method was originally applied for psychologist researchers to obtain data about the mechanism and internal structures of the cognitive processes (Ericsson and Simon. 1980), [28].

Advantage:

It allows as the obtaining of a great mount of qualitative data with a few subjects. The information recollected from the user is charged with personal recommendations that can be used in the design process.

Disadvantage:

No quantitative records are possible and only intuitive interpretations can be done.

Practical considerations

- Subjects tend to justify their errors and failures as design problems.
- Experimenter have to attend to what the evaluator are doing during the task and not attend to the rationalizations of subjects.
- A lot of people find that thinking-aloud is a not natural and they have problems in realising the task.
- Verbalizations can interfere with cognitive processing and decrease the subject’s action.
- Verbalizations can motivate to subjects and promote the subject’s action.
- Researcher must put questions to the subjects to facilitate the verbalizations (for example, what are you thinking now?)
- Researcher have to be attentive to the subject’s task and ask himself if the user acts surprised after a system action.
- To avoid the initial surprise with the thinking-aloud task a learning trial is recommended. The user can also see a video of others subjects doing similar tasks.

Different approaches to thinking aloud.

Constructive interaction.

It is identical to thinking aloud, but there are two subjects verbalizing simultaneously. Users must co-operate in the task. The situation is more natural, but the disadvantage is that the co-operative interaction can not be done because each user have different strategies for learning and using computers. It is specially recommended to work with children. (O'Malley et al. 1984) [29].

Retrospective testing.

The subject is working with the interface and his actions are recorded in video format. Later the subject views the recording and makes comments about the task. With this method is possible to get information carefully and without time problems. (Hewett and Scott 1987) [30].

Coaching method.

An expert user teaches to a non-expert user to use the interface. The experimenter writes what parts are having problems and what information it is necessary to give to generate manuals on training. (Mack and Burdett 1992) [31].

7.12 Audio and video records

Audio and video recordings are the best way to register user behaviour with a system, and are usually made in experimental environments, e.g., in the usability laboratory, for experimental or observation purposes. Users interacting with the system are videotaped, and their behaviour analyzed using structured observation techniques. Measurements of errors and performance times are easily taken and can be analyzed using suitable statistical techniques. Use of these procedures provides a way to effectively observe, register and validate the design hypotheses.

The same principles and recommendations as in observation methods can be applied. It is important to emphasize the use of timecodes and task start/finish time notation for later use during analysis and evaluation. This is the kind of simple detail that can save new researchers hours of unnecessary work.

Audio and video recording of participant's behaviour has specific ethical and legal requirements (see clause 9).

8 Data Analysis

Once the usability evaluation has been carried out, the researcher has to make decisions and extract conclusions from the data. Statistical methods are most appropriate for these purposes, and they are introduced in this chapter.

8.1 Measurement Theory and its implications for data analysis

ETR 095 [1] devoted a full chapter to this issue, because data analysis depends on the decisions about the measurement scale of the data gathered, and because there are many misconceptions about this particular issue. For these reasons the present document keeps this part, and reproduces the most up-to-date approach, obtained from the Measurement FAQ (excerpts) [55]:

What is measurement theory?

Measurement theory is a mathematical theory that is directly applicable to many problems of measurement and data analysis. The fundamental idea of measurement theory is that measurements are not the same as the attribute being measured. Hence, if you want to draw conclusions about the attribute ,you must take into account the nature of the relationship between the attribute and the measurements.

Why should I care about measurement theory?

When we measure something, the resulting numbers are usually, to some degree, arbitrary. We *choose* to use a 1 to 5 Likert scale instead of a –2 to 2 scale. We choose to use Fahrenheit instead of Celsius. We choose to use miles per gallon instead of gallons per mile. The conclusions of a statistical analysis should not depend on these arbitrary decisions, because we could have made the decisions differently. We want the statistical analysis to say something about reality, not simply about our whims regarding meters or feet.

What is measurement?

Measurement of some attribute of a set of things is the process of assigning numbers or other symbols to the things in such a way that properties of the numbers or symbols reflect properties of the attribute being measured. A particular way of assigning numbers or symbols to measure something is called a **scale of measurement**.

What are permissible transformations?

Permissible transformations are transformations that preserve the relevant properties of the measurement process. *Permissible* is a technical term; use of this term does not imply that other transformations are prohibited in data analysis any more than use of the term normal implies that other distributions are pathological.

What are levels of measurement?

There are different levels of measurement that involve different properties(relations and operations) of the numbers or symbols that constitute the measurements. Associated with each level of measurement is a set of permissible transformations. The most commonly discussed levels of measurement are as follows:

Nominal

Two things are assigned the same symbol if they have the same value of the attribute. Permissible transformations are any one-to-one or many-to-one transformation, although a many-to-one transformation loses information.

Ordinal

Things are assigned numbers such that the order of the numbers reflects an order relation defined on the attribute. Two things x and y with attribute values a(x) and a(y) are assigned numbers m(x) and m(y) such that if m(x) > m(y), then a(x) > a(y). Permissible transformations are any monotone increasing transformation, although a transformation that is not strictly increasing loses information.

Interval

Things are assigned numbers such that differences between the numbers reflect differences of the attribute. If m(x) - m(y) > m(u) - m(v), then a(x) - a(y) > a(u) - a(v). Permissible transformations are any affine transformation t(m) = c * m + d, where c and d are constants; another way of saying this is that the origin and unit of measurement are arbitrary.

Log-interval

Things are assigned numbers such that ratios between the numbers reflect ratios of the attribute. If m(x) / m(y) > m(u) / m(v), then a(x) / a(y) > a(u) / a(v). Permissible transformations are any power transformation t(m) = c * m ** d, where c and d are constants.

Ratio

Things are assigned numbers such that differences and ratios between the numbers reflect differences and ratios of the attribute. Permissible transformations are any linear (similarity) transformation t(m) = c * m, where c is a constant; another way of saying this is that the unit of measurement is arbitrary.

Absolute

Things are assigned numbers such that all properties of the numbers reflect analogous properties of the attribute. The only permissible transformation is the identity transformation.

These measurement levels form a partial order based on the sets of permissible transformations:

```
                    - Interval -
                   /            \
        Nominal -- Ordinal  Ratio -- Absolute
                   \            /
                    Log-interval
Weaker <-----------------------------------> Stronger
```

In real life, a scale of measurement may not correspond precisely to any of these levels of measurement. It is common to have scales that lie somewhere between the ordinal and interval levels in that the permissible transformations are considered to be smooth monotone transformations. There are more complicated types of measurement that yield more complicated types of scales; for example, it is common to have mixtures of nominal and ordinal information in a single scale, such as questionnaires that have several non-response categories. Unfortunately, there are also many situations where the measurement process is too ill-defined for measurement theory to apply. In such cases, it may still be fruitful to consider what arbitrary choices were made in the course of measurement, what effect these choices may have had on the measurements, and whether some plausible class of permissible transformations can be determined

Is measurement level a fixed, immutable property of the data?

Measurement level depends on the relationship between the measurements and the attribute. Given a set of data, one cannot say what the measurement level is without knowing what attribute is being measured. It is possible that a certain data set might be treated as measuring different attributes at different times for different purposes.

Once a set of measurements have been made on a particular scale, it is possible to transform the measurements to yield a new set of measurements at a different level. It is always possible to transform from a stronger level to a weaker level. For example, a temperature measurement in degrees Kelvin is at the ratio level. If you convert the measurements to degrees Celsius, the level is interval. If you rank the measurements, the level becomes ordinal. In some cases it is possible to convert from a weaker scale to a stronger scale. For example, correspondence analysis can sometimes convert nominal or ordinal measurements to an interval scale.

What does measurement level have to do with discrete vs. Continuous variables?

Measurement level has nothing to do with discrete vs. continuous variables. The same happens with objective versus subjective data. While measurements are always discrete due to finite precision, attributes can be conceptually either discrete or continuous regardless of measurement level.

Does measurement level determine what statistical methods are appropriate?

Measurement theory cannot determine some single statistical method or model as appropriate for data at a specific level of measurement. But measurement theory does in fact prove that some statistical methods are inappropriate for certain levels of measurement if you want to make inferences about the attribute being measured.

If one wishes to make statistical inferences regarding an attribute based on a scale of measurement, the statistical method must yield invariant (or equivariant) results under the permissible transformations for that scale of measurement. If this invariance does not hold, then the statistical inferences apply only to the measurements, not to the attribute that was measured.

Since there has been so much confusion on this point, it must be emphasized that measurement theory does not restrict transformations in a statistical analysis to the class of permissible transformations. That is not what permissible transformation means. The point is that statistical methods should be used that give invariant results under the class of permissible transformations, because those transformations do not alter the meaning of the measurements.

It is important to understand that the level of measurement of a variable does not mandate how that variable must appear in a statistical model. However, the measurement level does suggest reasonable ways to use a variable by default.

How does measurement level relate to statistical methodology?

Measurement level must be considered to avoid making meaningless statements. A typical example of a meaningless statement is the claim by the weatherman on the local TV station that it was twice as warm today as yesterday because it was 40 degrees Fahrenheit today but only 20 degrees yesterday. Fahrenheit is not a ratio scale, and there is no meaningful sense in which 40 degrees is twice as warm as 20 degrees. It would be just as meaningless to compute the geometric mean or coefficient of variation of a set of temperatures in degrees Fahrenheit, since these statistics are not invariant or equivariant under change of origin. There are many other statistics that can be meaningfully applied only to data at a sufficiently strong level of measurement.

The general principle is that an appropriate statistical analysis must yield invariant or equivariant results for all permissible transformations. Obviously, one cannot actually conduct an infinite number of analyses of a real data set corresponding to an infinite class of transformations. However, it is often straightforward to verify or falsify the invariance mathematically. The application of this idea to summary statistics such as means and coefficients of variation is fairly widely understood. For example, a mean is equivariant under changes of origin and scale, hence it is a suitable location estimator for a variable measured at the interval level. The coefficient of variation is *not* invariant under changes of origin, so it is *not* a suitable scale estimator for an interval-level variable.

What's the bottom line?

Measurement theory shows that strong assumptions are required for certain statistics to be meaningful. Measurement theory encourages people to think about the meaning of their data. It encourages critical assessment of the assumptions behind the analysis. It encourages responsible real-world data analysis.

How does all this apply to usability evaluation?

Usability is a construct with no clear definition, or better, with a very complex definition, involving many variables. Until now, there is no usability scale yet. However, this does not imply that there will never be such scale. The work undertaken in other telecommunication areas has produced commonly agreed scales to rate system properties, as the MOS (Mean Opinion Score) scale in audio quality for telephone lines, and it proves that it is fairly possible.

The most usual scales in usability evaluation are interval scales (number of errors or ratings, for example). These scales allow for comparisons on the differences between overall ratings of two systems: for instance, system A obtains 6, system B 8 and system C 12 points. We can say that the difference between A and C is three times the difference between A and B, but we cannot say that system C is twice as usable as system A. In most, if not all, cases, usability measurements (either objective or subjective) will be relative, allowing for comparisons between systems or services, but not absolute.

8.2 Descriptive or exploratory data analysis

Descriptive statistics should be employed in the first steps of a statistical analysis to present raw data into a small number of representative figures. Data sets are summarized in order to describe two important features of the data distributions: the spread of scores and where within that range most of the data falls, e.g. the average of the scores (ISSUE Usability Evaluation Guidelines) [5]. Measures of central tendency are used to assess the 'middle' or 'average' value in a distribution and measures of dispersion in order to estimate the amount of variability contained within the data and thus the degree to which the average is a typical value.

In nominal scale variables (e.g., forced choice questions), the description is made by means of percentages for each alternative. In ordinal variables, the central tendency statistic is the median. With interval scale variables, the descriptive analysis should include measures of central tendency, as the mean, together with dispersion statistics as the standard deviation or the variance. It is very useful to include the confidence interval for the mean (usually 95% confidence level), since this indicates with more precision the range of values of the variable falling in the centre of the population. Remember that statistics was made for measures with error, so it is not exact to describe a whole population with only one value.

Exploring the data is now considered a very important part of the data analysis process, because it can provide very useful insight about 'what went on' in the experimental or evaluation phase, before really testing your hypotheses. A note of caution is to be given about the presence of outliers: these are extreme values and can happen because of coding errors, mistakes, etc. They are usual, for example, in user's response times, if the user does not know what to do in a particular situation. This situation can produce wrong overall results, especially if very few subjects participated in the experiment. For instance, if you have 10 subjects, with a mean response time of 2 seconds. If only one subject has a response time of 10 seconds, your overall mean could be much higher than what is representative for the group of ten people. These outliers can be identified by means of statistical techniques, and if they affect the results or the computations, eliminated from the analysis.

Another concern in the analysis phase is that of statistical model assumptions. These can be of several types: e.g., linearity in ANOVA or regression models, normal distribution of errors. When your data do not follow these assumptions in an extreme way (statistical models are always tolerant for deviations), you can consider a transformation of data. A recommended text about this issue and, in general, exploratory data analysis, is Tukey (1977) [32].

8.3 Reporting trial results and statistical data

APA Publication norms [33] are good advice about how to present the results of usability test. The usability experts should follow these recommendations:

1) Selecting the Method of analysis.

   Authors are responsible for the statistical method selected and for all supporting data.

   Access to computer analyses of data does not relieve the authors of responsibility for selecting the appropriate statistic.

2) Selecting Effective Presentation by means of tables and figures.

The following recommendations apply to tables:

- Tables are efficient, enabling the researchers to present large amount of data in small space.
- Tables show exact numerical values.
- Tables allow comparisons.
- Authors should choose not to present too many tables: readers may lose track of the message.
- Author should reserve tables for crucial data that are directly related to content (results/most important message).
- Tables presenting numbers are only efficient if data are arranged in such a way that it is possible to see meaning at a glance.

In a usability report, tables should at least display:

- Descriptive statistics: mean, median, frequency, percentage, number of data from which these statistics were computed.
- Where means are reported, always include associate measure of variability.
- For specialized analyses, like ANOVA tables and others, refer to APA norms [33], 3.69.

The following recommendations apply to figures:

Figures convey at a quick glance an overall pattern of results.

A well-prepared figure can also convey structural or pictorial concepts more efficiently than text.

A good figure:

- Arguments rather than duplicates the text.
- Conveys only essential factors.
- Omits visually distracting detail.
- Is consistent with similar figures in the same paper.

3) Reference for statistics.

- Do not give reference for statistic in common use. Do give if:
    - You use a less common statistic.
    - A statistic used in a controversial way.

4) Reporting Statistical Tests (Inferential).

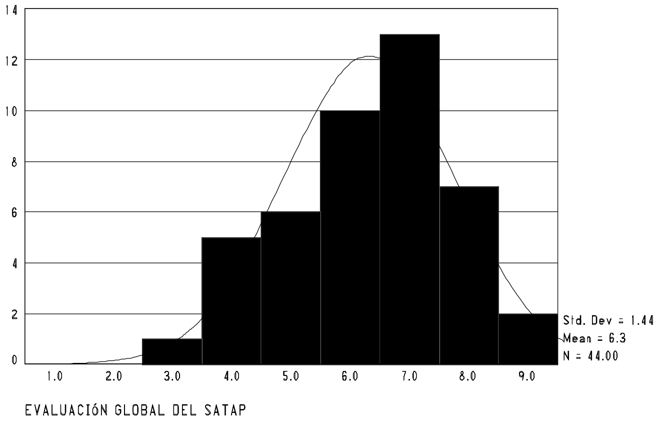
- Include sufficient information to permit the reader to corroborate the analyses, and as a minimum:
    - Obtained magnitude or value of the test.
    - Degrees of freedom.
    - Probability level.
    - Direction of effect.
    - Assume that reader has professional knowledge of statistic.

8.3.1 Plots

Even the simplest usability evaluation experiment can produce large amounts of quantitative data. When these data are presented in their 'raw' form-as lists of errors made, for example- it can be very difficult to appreciate their meaning. A powerful technique for organizing results into a more easily understandable form is to represent the data pictorially in the form of a chart. Most people are familiar with the use of charts as pictorial representations of numeric information. The advantage of the chart is that it allows the investigator to see immediately both the magnitude and distribution of scores within a set of data.

There is a considerable variety of ways to plot different data sets. In the following point you will find general recommendations about which ones are appropriate for different purposes:

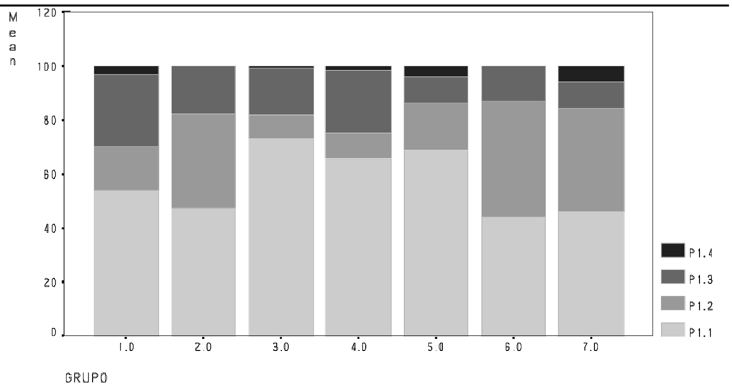
Table 3: Types of plots and their application for presenting results

Graphical Technique	Example
Bar charts and histograms Bar charts plot the frequency or percentage of different objects or events. These are considered to be not quantitative, but discrete events. Histograms are a variation of bar charts used when the x axis is scaled as a function of a continuous variable, which can be time or the range of values for a continuous variable. These are useful for testing distribution assumptions (but there are formal statistical techniques for this) by plotting the normal curve on it. The attached plot shows the distribution of the global evaluation for a prototype	

system.

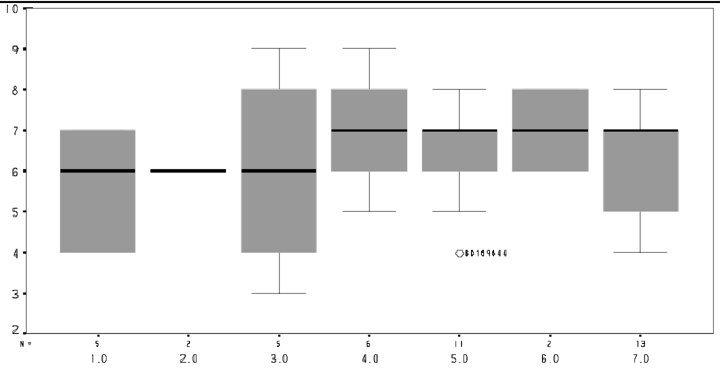
Stacked Bar Charts.

These are a variation of bar charts, and are useful to show the distribution (frequency) of different alternatives or variables as a function of another variable (e.g., group). Care should be taken not to present too many variables, groups, or, in general, making a too complicated plot, since then it is less useful.



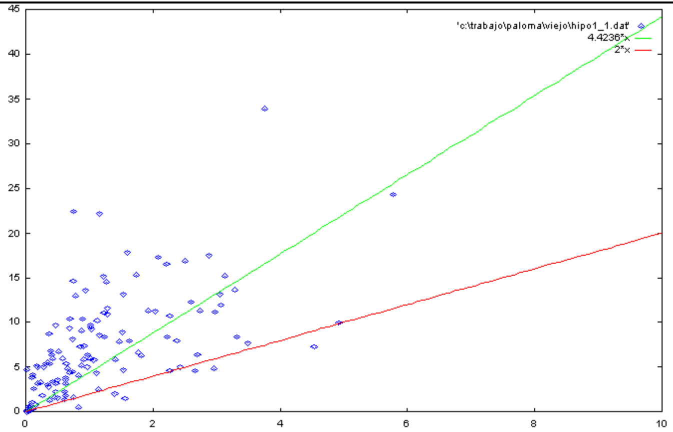
Box and whisker plots

These show the mean, the confidence interval of the mean, and maximum and minimum values of one variable in several groups. They are very useful for detecting the presence of outliers.



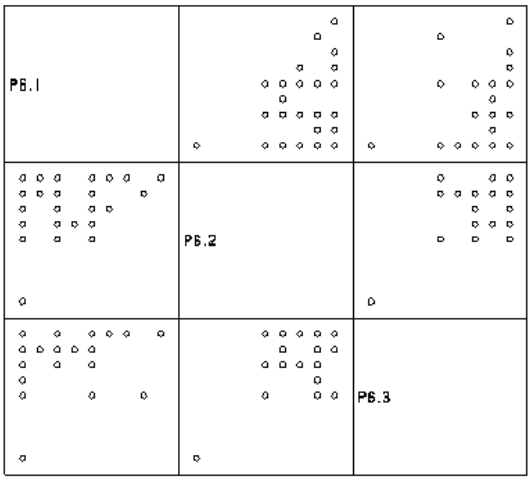
Scatterplots

These show the joint distribution of two variables. They are useful for finding out relationships (linear or non-linear) between two variables (at a minimum of interval level of measurement).



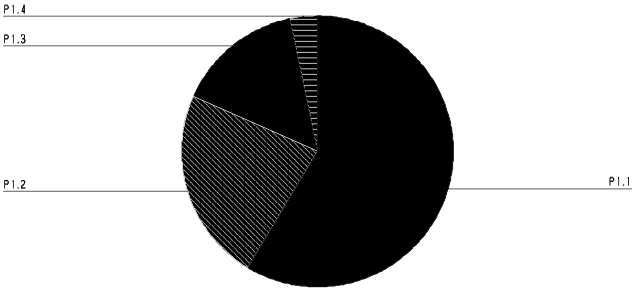
Multiple Scatterplots

They are used to graphically show the relationship between several variables, in a matrix form. They are useful for exploratory analysis in case of trying to find out constructs explaining several variables (those correlating highly between them can be considered to explain the same concept).



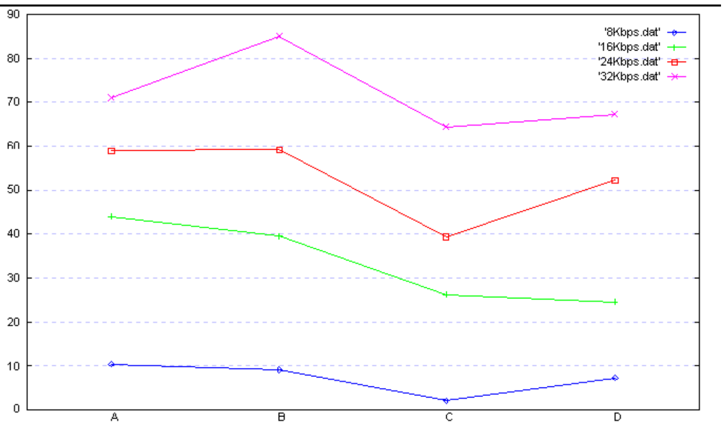
Pie charts

These are very usual in normal life. They depict the percentage for different things out from the total.



Lines plots

These graphics are often used to show the means for different conditions of an experimental factor. Lines connecting the points (means) depict the results for one experimental factor, while x axis scale represents another factor. In this context, they are also called means plots.



8.4 Experimental data analysis: testing comparative hypotheses

The choice for statistical technique to test for comparative hypothesis is completely dependent on the design. However, there is a variety of techniques for similar problems. One of the choices the researcher has to make is about parametric or non-parametric tests. Parametric tests are usually preferred, since they provide more power (capability to detect differences), and are more accepted in the scientific areas. These tests require that several assumptions have to be met but they are quite flexible for small deviations of the assumptions. For each techniques we will explain, we will review up to what extent they are applicable with different violations. However, in case of extreme violations of the assumptions, the researcher should also perform non-parametric tests, which are usually more reliable with 'messy' data, to see if the results are comparable to those obtained with parametric tests.

8.4.1 Comparing two means: Student's t-tests

This statistical procedure is only appropriate for comparing two means obtained from independent samples of a population (this is the case for matched pairs and independent groups designs with only two levels of the experimental factor). It is inappropriate in all cases for comparing more than 2 means. For instance, in an experiment with 3 levels (a, b, c) of an experimental factor, you have 3 possible comparisons: a vs. b, a vs. c and b vs. c. All statistics books advice that performing simple t-tests to make the three comparisons can lead to mistakes. When you have more than 3 means to compare, you should use analysis of variance procedures (ANOVA), explained below.

Student's t tests assume normal distribution of the scores. However, they are usually very robust to non-extreme violations of this assumption. They also require homogeneity of variance (called homoscedasticity), and they are not so robust for violations of this assumption. The approach in these cases, which are fairly common, is to reduce the degrees of freedom. These procedures are called robust, and reduce the possibility of detecting differences, but make the test more reliable. All statistics packages computing t-tests provide facilities to compute these tests, and should be used whenever the researcher, after exploring data, finds that they don't have homogeneous variance.

Another assumption we will insist on is independence of measures: these procedures are never appropriate for testing repeated measures, this is, when subjects experience the two experimental conditions we are making the test on. For these situations, see repeated measures ANOVA below.

A good reference for these procedures can be found in Hays (1988) [34] and Kirk (1982) [35].

Non-parametric alternative to Student's t tests is Wilcoxon's procedure(again, see Hays, 1988).

8.4.2 Comparing more than two means: ANOVA models

Although not generally known, analysis of variance (ANOVA) models area subclass of the more general linear model which will be exposed in following point, whenever the predictor variables are qualitative (as it is usually the case for experimental factors).

There are complete books dealing with these models, and their application in experimental design and analysis, covering fixed effects, random effects, confounded, fractional models. For these, see Kirk (1982) [35], Maxwell et al.(1990) [36] and Montgomery (1991) [37]. We will only deal with the most simple models for the experimental designs in previous points: independent groups and repeated measures.

The ANOVA model is a linear model formulated in the following way:

$Y_{ij}= \mu+\alpha_j+\varepsilon_{ij}$ i=1..n subjects or observations

j=1..p levels of the experimental factor

Y_{ij} is the observation or measurement

μ is the population mean, a constant in the model

α_j is the effect for each level(j) of the experimental factor

$\varepsilon_{ij} \sim N\ IND\ (0,\sigma^2_e)$ is the measurement error, one for each observation, and should be independent of each other, with 0 mean and constant variance.

The consequences of this formulation is that ANOVA is a very flexible procedure for testing comparative hypothesis dealing with multiple levels of the experimental factor. However, it is not appropriate for non-linear relationships between the factor and the measurement. In the case that the relationship is non-linear, you can perform transformations on your data to have new measurements (see Hoaglin, Mosteller and Tukey, 1991) [38].

The independence of errors assumption excludes from ANOVA models all hypotheses of variation along time, since the error parameter must be independent of each other, which is not the case for repeated measurements in the same condition in different moments.

Again, the assumptions of these tests is normal distribution of scores. However, it has been shown (Kirk, 1982) [35] that ANOVA is very robust against even extreme deviations of normality. However, it is very much affected by violations of other assumption: homoscedasticity. When data show very different variances across groups or levels of the experimental factor, you may find significant differences when they do not really exist, and viceversa. There are specific tests for this assumption, and, in any case, you have robust techniques available. These procedures reduce the likelihood of finding significant differences, making the test less powerful to detect them, and, at the same time, more reliable.

The statistical hypothesis tested in ANOVA can be formulated in the following way:

H_0: $\alpha_j=0\ \forall j$ there is no significant effect for any level of the experimental factor.

H_1: some α_j different from 0; there is at least one significant effect in one level of the experimental factor.

The ANOVA test produces a table which shows the different computations carried out. First, two measures of variability are computed: the sum of squares for the effect of the experimental factor or group (SS between) and a measure of error (SS within). These measures are relative to the number of experimental conditions of groups and of observations for which the test is computed (degrees of freedom). Dividing SS by their degrees of freedom you will have the Mean Square estimates (MS). The F test tells how much of the variability due to the effect of the experimental factor explains the total variability. This computation is compared with the F distribution with its degrees of freedom, and a probability of rejecting the null hypothesis is finally obtained.

If the probability tests of the F contrast is below 0.05, it is generally accepted that there has been a significant effect of the experimental factor. But take into account that this is not a rule of thumb.

Even though differences are considered significant, results of ANOVA do not indicate which particular levels of the factor are significantly different of each other. Many researchers make simple t-tests (pairwise comparisons), which can lead to wrong conclusions. Recommended procedures for post-ANOVA pairwise comparisons are called multiple comparisons procedures. You have a range of techniques, from Bonferroni tests, to Scheffé or Tukey tests (see Kirk, 1982) [35].

8.4.3 Repeated measures ANOVA

This model is appropriate when subjects experience all the conditions in the experimental factor. The model is the same as for independent measures ANOVA, with an added parameter:

$Y_{ij}= \mu+\alpha_j+\tau_j+\varepsilon_{ij}$i=1..n subjects or observations

j=1..p levels of the experimental factor

The new parameter (τ_j) is introduced for all the effects included by using the same experimental unit (subject) for the different experimental conditions.

Assumptions are the same as stated previously, but for a different conception of homoscedasticity (homogeneity of variance). The assumption is now called sphericity, and involves testing the variance-covariance matrix of scores. This is usually very difficult, so always perform robust tests, like Greenhouse-Geisser or Huyhn-Feldt degrees of freedom adjustments (see Kirk, 1982 [35] for a reference on these procedures). Most modern statistical software should provide these tests, by default or as an option.

Sum of squares computations are different. In particular, the SS error is computed in a different way. On the other hand, the interpretation of the results is exactly the same as in independent measures ANOVA.

8.5 An introduction to multivariate analysis models

Linear multivariate models are a general class of mathematical models defined in the following way (matrix notation):

$\mathbf{y}= \mathbf{X}\, \beta + \varepsilon$

where **y** is a vector of criterion variable

X is a matrix of predictor variables

β is the vector of coefficients or weights of the model

ε is a vector of errors, which are assumed to be independent, with 0 mean and constant variance

Two things are important in this model:

- There is only one criterion, or variable being predicted (e.g., performance with the system in the working context). There are multiple predictor variables (e.g., training, etc.). The criterion must be at least of interval level of measurement. Predictors can be either interval level, or lower (ordinal -seldom- and nominal or qualitative). In this case, it can be shown that you will have an ANOVA model.

- The most important assumption is that of the form of the relationship: this is linear. Non-linear relationships can be estimated using other methods, known as non-linear regression. We will not deal with them here. As with ANOVA, the independence of errors assumption prevents these models to be used with time-dependent data.

These techniques are widely used in all knowledge areas, especially for prediction of important real life variables, like performance in the work context (as opposed to the laboratory context), from other observation variables (number of hours of training, satisfaction with new technologies, etc.).The scatterplot shown in the exploratory analysis chapter shows the results of fitting two alternative linear models to some data.

8.5.1 A table to choose among alternative multivariate analysis models

The main problem of these models for testing linear hypotheses come from the lack of a priori knowledge about which variables are better predictors. The iterative nature of the computation algorithms can make the results variable depending on the order of introduction of each variable. There are several statistical procedures devised to solve this: stepwise regression tries to introduce the variables in the equation following iterative forward and backward rules. However, this technique is now considered to be merely exploratory, since results can also depend on computational aspects rather than of substantive things about the predictors. Best subsets regression is a technique to find out the best and most economical model for prediction. However, this is very computation and time-consuming, and is usually only possible when you have few predictor variables.

These models are called multivariate because they are devised to find out relationships between several dependent variables. Dependent Variables (DV), also named criteria, are the variables to be explained by the model. For example, items in a questionnaire are considered as dependent variables in a factor analysis. In some models, Independent Variables (IV) are also considered. In linear regression, for instance, IV or predictors are the variables which will form the model to predict the criteria. In MANOVA (Multivariate Analysis Of VAriance), independent variables may also define groups of subjects.

Therefore, some multivariate models probe the effect of IV on DV (Discriminant Analysis, Multivariate Regression and MANOVA). However, in other methods (Principal Components Analysis, Factor Analysis, Cluster Analysis, Canonical Correlation Models and Log-Linear Models) it is not possible to distinguish between different types of variables. In these models, all variables are considered in the same way, and the particular model will try to find out relationships between them.

The table below intends to present a summary of the different available techniques in such a way that it is easy for the usability researcher to make comparisons between them and to choose the best one for the research purposes.

Table 4: Mutivariate statistical models

Principal Components Analysis

Aim	**To explain the data in a more parsimonious way. To obtain a new set of constructs, fewer than original variables, that explains the whole set of original variables.**
When to use it	We have a variables set and we want to explain them with fewer constructs. In addition, we require that the new set of constructs are uncorrelated.
Types of variables	It involves multiple dependent variables at interval level of measurement and no independent variables.
Comments	The PCA makes a new variables set (named factors) from raw data. The factors have two characteristics: they are uncorrelated and ordered as a function of the variance explained (the first factor is the best explaining the total variance).
References	Gorsuch (1983) [39], Lawley y Maxwell (1971) [40]

Factor Analysis

Aim	**To explain the data in a more parsimonious way. To obtain a new set of constructs, fewer than original variables, that explains the whole set of original variables.**
When to use it	We have a variables set and we want to explain them with fewer constructs. There are several possibilities about the relationship between the constructs: they may be uncorrelated, correlated, and rotated in different ways (orthogonal and non-orthogonal).
Types of variables	It involves multiple dependent variables at interval level of measurement and no independent variables.
Comments	Factor analysis has the same objectives that PCA, but it is based on a different model
References	Afifi and Azen (1979) [41], Harman (1967) [42]

LISRELModels

Aim	**LISREL is the acronym for LInear Structural RELationships. The aim is to test alternative factor structures explored or obtained with factor analysis techniques.**
When to use it	First you need to have alternative factorial models and you need to probe them using goodness of fit statistics.
Types of variables	It involves multiple dependent variables at interval level of measurement and no independent variables.
Comments	These procedures can test different factor structures, and the difference with factor analysis is precisely that the researcher gets a goodness of fit statistic of alternative structures. They are usually complex, but are becoming quite common in social research.
References	Jöreskog and Sörbom (1979) [43], Saris (1984) [44]

Multi-Dimensional Scaling

Aim	**It is a class of methods for estimating the co-ordinates of a set of objects in a space of specified dimensions, from data measuring the distances between pairs of objects.**
When to use it	You need data in pairs, and you have to generate the proximity matrix (one or more square symmetric or asymmetric matrices of similarities or dissimilarities between objects)
Types of variables	Multiple dependent variables of any type (there are metric and non-metric versions of MDS), with no independent variables
Comments	A variety of models can be used involving different ways of computing distances and various functions relating the distances to the actual data. The MDS procedure fits two- and three- way, metric and non metric multidimensional scaling models.
References	Kruskal and Wish (1978) [45], Arabie, Carroll and DeSarbo (1987) [46]

Discriminant Analysis

Aim	**To find the best linear function to discriminate between two groups.**
When to use it	We have two groups and we want to find the best combination of variables by means of which we can determine if an element belongs to one group or to the other.
Types of variables	It involves a dichotomous (only two values) dependent variable, and multiple independent variables at interval level of measurement.
Comments	There are many techniques and models to make this analysis but all find the best classification function between two groups.
References	Tabachnick and Fidell (1983) [47], Afifi and Clark (1984) [48]

Cluster Analysis

Aim	**To explore relationships in a set of variables or elements and make clusters (homogeneous groups) with them, as a function of the distance between them.**
When to use it	When you have a set of variables or of elements (e.g., subjects) and you have to make homogeneous groups with them. This is a particular type of exploratory analysis, and it is useful to make hypotheses about the relationships between variables or elements.
Types of variables	It involves a multiple dependent or grouping variables at any level of measurement. It has no independent variables.
Comments	The clustering process follows an iterative process, beginning with "one variable=one cluster" and ending with "all variables = one cluster". The researcher has to choose the step that better explains his data
References	Hartigan (1975) [49], Aldenderfer (1984) [50]

Multivariate Regression

Aim	**To find out a linear model to predict the value of multiple criteria (dependent variables) by means of two or more predictors or independent variables.**
When to use it	When you know that a linear function is a good way to explain or predict your criteria, as a function of some predictors. For example, when you want to predict both people's performance and attitudes with a system as a function of previous experience and global ability with computers.
Types of variables	It involves multiple dependent variables at interval level of measurement. It has multiple independent variables at interval level of measurement (predictors) or codes (grouping variables).
Comments	These models are a generalization to multiple criteria of the General Linear Model, as exposed above.
References	Afifi and Clark (1984) [48]

MANOVA (Multivariate ANalysis Of VAriance)

Aim	**To probe if the means of two (or more) DV are different between two (or more) groups.**
When to use it	We have two (or more) grouping variables and two (or more) DV and we want to test if the means of the groups are different.
Types of variables	It involves multiple dependent variables at interval level of measurement. It has multiple grouping variables, and can incorporate also covariates (at interval level of measurement).
Comments	These models are the multivariate generalization of the traditional ANOVA models.
References	Davidson (1980) [51]

Log-Linear Models

Aim	**It uses a logarithmic-transformed model and stepwise methods to find the function that maximizes the expected value of a frequency table.**
When to use it	We have a set of categorical variables and we want find out the most economic model which predicts the expected values of a frequency table.
Types of variables	It involves multiple nominal dependent variables and no independent variables.
Comments	This model is based in a logarithmic linear model, and is equivalent to making anova with nominal level of measurement data (e.g., frequencies or percentages of cross-classifications in a table).
References	Bishop et al (1975) [52]

Logistic regression

Aim	**To obtain a model which best explains a dichotomous criterion (e.g., you like or not a system) as a function of predictors.**
When to use it	You have a dichotomous outcome which you want to predict as a function of a set of variables, which can be of several types.
Types of variables	Only one dichotomous dependent variable and multiple nominal (grouping) or interval level of measurement independent variables (or predictors)
Comments	They are based on maximum-likelihood estimation techniques, and stepwise model-selection procedures. This is a specialized model to deal with regression with a dichotomous criterion (e.g., yes/no response).
References	Hosmer and Lemeshow (1989) [53]

9 Ethical issues in usability testing

Usability evaluation can be considered part of psychological research, since many of the constructs used in the research are psychological concepts, and the basis for any usability evaluation is the final user. Therefore, the psychologists ethical standards clearly apply.

In this section a list of the applicable standards of good practice are provided, obtained from the two best known Psychological Associations, the British Psychological Society (BPS) [54] and the American Psychological Association (APA). The excerpts in the present document are, forcefully, only a small part of the whole codes of conduct, which can be freely consulted at the addresses provided in the bibliography. For the purposes of the present document, both BPS and APA Codes of Conduct apply whenever appropriate.

The main aim of the research procedures outlined in the present document is to obtain valid evidence about how usable a system or service is. Experts and evaluators carrying out these tests shall ensure that research is carried out in keeping with the highest standards of scientific integrity and the ethical norms, as follows:

9.1 Respect for people's rights and dignity

Researchers accord appropriate respect to the fundamental rights, dignity, and worth of all people. They respect the rights of individuals to privacy, confidentiality, self-determination, and autonomy, mindful that legal and other obligations may lead to inconsistency and conflict with the exercise of these rights. Researchers are aware of cultural, individual, and role differences, including those due to age, gender, race, ethnicity, national origin, religion, sexual orientation, disability, language, and socioeconomic status. Researchers try to eliminate the effect on their work of biases based on those factors, and they do not knowingly participate in or condone unfair discriminatory practices.

9.2 Privacy and confidentiality

In general, and subject to the specific requirements of law, usability experts shall take care to prevent the identity of individuals, organizations or participants in research being revealed, deliberately or inadvertently, without their expressed permission.

Specifically they shall:

- endeavour to communicate information obtained through research or practice in ways which do not permit the identification of individuals or organizations;
- take all reasonable steps to ensure that records over which they have control remain personally identifiable only as long as is necessary in the interests of those to whom they refer, and to render anonymous any records under their control that no longer need to be personally identifiable for the above purposes;
- only make audio, video, or photographic recordings of recipients of services or participants in research (with the exception of recordings of public behaviour) with the expressed agreement of those being recorded both to the recording being made and to the subsequent conditions of access to it;
- Confidential Information in Databases.

(a) If confidential information concerning participants in usability tests is to be entered into databases or systems of records available to persons whose access has not been consented to by the recipient, then use coding or other techniques to avoid the inclusion of personal identifiers.

(b) If a research protocol approved by an institutional review board or similar body requires the inclusion of personal identifiers, such identifiers are deleted before the information is made accessible to persons other than those of whom the subject was advised.

(c) If such deletion is not feasible, then before the research responsibles transfer such data to others or review such data collected by others, they take reasonable steps to determine that appropriate consent of personally identifiable individuals has been obtained.

9.3 Informed Consent to Research

Researchers shall normally carry out investigations only with the valid consent of participants, having taken all reasonable steps to ensure that they have adequately understood the nature of the investigation or intervention and its anticipated consequences.

Prior to conducting research (except research involving only anonymous surveys, naturalistic observations, or similar research), the usability investigators enter into an agreement with participants that clarifies the nature of the research and the responsibilities of each party.

- The researcher should use language that is reasonably understandable to research participants in obtaining their appropriate informed consent (except as provided below, "Dispensing with Informed Consent"). Such informed consent is appropriately documented.
- Where it is necessary not to give full information in advance to those participating in an investigation, provide such full information retrospectively about the aims, rationale and outcomes of the procedure as far as it is consistent with a concern for the welfare of the participants;
- Using language that is reasonably understandable to participants, the evaluators inform participants of the nature of the research; they inform participants that they are free to participate or to decline to participate or to withdraw from the research; they explain the foreseeable consequences of declining or withdrawing; they inform participants of significant factors that may be expected to influence their willingness to participate (such as risks, discomfort, adverse effects, or limitations on confidentiality, except as provided below, Deception in Research); and they explain other aspects about which the prospective participants inquire.
- For persons who are legally incapable of giving informed consent, the researcher nevertheless (1) provide an appropriate explanation, (2) obtain the participant's assent, and (3) obtain appropriate permission from a legally authorized person, if such substitute consent is permitted by law.
- Dispensing With Informed Consent.
- Before determining that planned research (such as research involving only anonymous questionnaires, naturalistic observations, or certain kinds of archival research) does not require the informed consent of research participants, researchers consider applicable regulations and institutional review board requirements, and they consult with colleagues as appropriate.
- Informed Consent in Research Filming or Recording.
- Researchers obtain informed consent from research participants prior to filming or recording them in any form, unless the research involves simply naturalistic observations in public places and it is not anticipated that the recording will be used in a manner that could cause personal identification or harm.
- Offering Inducements for Research Participants.
- (a) In offering professional services as an inducement to obtain research participants, researchers make clear the nature of the trials, as well as the risks, obligations, and limitations. (b) Researchers do not offer excessive or inappropriate financial or other inducements to obtain research participants, particularly when it might tend to coerce participation.
- Deception in Research.
- (a) Usability researchers do not conduct a study involving deception unless they have determined that the use of deceptive techniques is justified by the study's prospective scientific, educational, or applied value and that equally effective alternative procedures that do not use deception are not feasible.
- (b) Researchers never deceive research participants about significant aspects that would affect their willingness to participate, such as physical risks, discomfort, or unpleasant emotional experiences.
- (c) Any other deception that is an integral feature of the design and conduct of an experiment must be explained to participants as early as is feasible, preferably at the conclusion of their participation, but no later than at the conclusion of the research.
- Providing Participants With Information About the Study.

- (a) Researchers provide a prompt opportunity for participants to obtain appropriate information about the nature, results, and conclusions of the research, and psychologists attempt to correct any misconceptions that participants may have.

- (b) If scientific or humane values justify delaying or withholding this information, the researchers take reasonable measures to reduce the risk of harm.

History

Document history		
V1.1.1	December 1999	Membership Approval Procedure MV 200005: 1999-12-07 to 1999-02-04