ETSI TECHNICAL REPORT

ETR 095

September 1993

Source: ETSI TC-HF

Reference: DTR/HF-3001

ICS: 33.020, 33.040.40

Key words: Human factors

Human Factors (HF);
Guide for usability evaluations
# of telecommunications systems and services

ETSI

European Telecommunications Standards Institute

ETSI Secretariat

Postal address: F-06921 Sophia Antipolis CEDEX - FRANCE
Office address: 650 Route des Lucioles - Sophia Antipolis - Valbonne - FRANCE
X.400: c=fr, a=atlas, p=etsi, s=secretariat - **Internet:** secretariat@etsi.fr

Tel.: +33 92 94 42 00 - Fax: +33 93 65 47 16

New presentation - see History box

Contents

Blank page

Foreword

This ETSI Technical Report (ETR) has been produced by the Human Factors (HF) Technical Committee of the European Telecommunications Standards Institute (ETSI).

ETRs are informative documents resulting from ETSI studies which are not appropriate for European Telecommunication Standard (ETS) or Interim European Telecommunication Standard (I-ETS) status. An ETR may be used to publish material which is either of an informative nature, relating to the use or the application of ETSs or I-ETSs, or which is immature and not yet suitable for formal adoption as an ETS or an I-ETS.

The acceptance of new telecommunications services and systems depends on their usability. It is important to apply usability principles in all phases of a development cycle and to do usability evaluations wherever it is feasible during the development of services or systems.

This ETR provides guidance on usability evaluation methods which give a common conceptual framework for describing usability and the established methods for its evaluation. It also gives guidance for reading and understanding test reports which present results based on these evaluation methods.

It is recommended that the definitions and the approach presented in this ETR be applied in the work of ETSI and in research and development in the telecommunications area in general.

Blank page

1 Scope

The purpose of this ETR is to support the usability evaluation of telecommunications systems and services by:

- improving communication by providing an agreed terminology;
- facilitating our understanding and comparison of test results;
- supporting planning and carrying out of usability evaluations.

This ETR realises the above goals by providing a general framework. This includes a definition of usability and a critique of the current approaches to usability evaluation (Clause 4). The adopted approach within this ETR is described in Clause 5. Here, critical concepts like evaluation methods, measures and metrics are discussed in the context of an overall evaluation process. The process itself is described in terms of eight key steps. These steps can be mapped onto the four main development phases for systems and services (i.e. conception, specification, realisation and operation). Clause 5 also includes recommendations for supporting documentation (e.g. a "test report"). The importance of the context for carrying out successful evaluations is stressed in Clause 6.

Later clauses provide a more detailed description of the recommended testing methods (Clause 7) and some detailed discussion about measurement (Clause 8).

This ETR does not provide precise descriptions or prescriptions on usability testing methods. For example, it will not tell you how to design a questionnaire but it will, in most cases, reference appropriate texts. It should be viewed as a general introduction to issues and problems of undertaking usability evaluations of telecommunications systems and services.

The intended readers of this ETR are:

- professionals who undertake usability evaluations, in particular those not working in the human factors area or having no, or limited, experience of usability evaluations;
- readers of usability evaluation test reports such as designers, engineers, managers and consumers.

2 References

For the purposes of this ETR, the following references apply:

[1] COST 212 (glossary), CEC (1992): "Human Factors in Information Services" (EUR 14277 EN).

[2] Leeuw, A.J.C de (1990): "Eee boekje over bedrijfskundige methodologie: management van onderzoek (Dutch)".

[3] Bury K.F (1984): "The iterative development of usable computer interfaces".

[4] Fowler, C.J.H, Rushton, P and Benton, R, BT Technol. Journal II, Pages 145 to 152 (1993): "Human Factors in the RACE programme - meeting the user's needs".

[5] Whiteside J., Bennet, J and Holtzbett, A, Handbook of Human-Computer Interaction (1988): "Our experience and evolution".

[6] Brooke J.B Usability Engineering in Office Product development (1986): "People and Computers: Designing and evolution".

[7] Sommerville, I, Addison-Wesley (1989): "Software Engineering".

[8] Card S.K, Moran, T. and Newell, A, Hillsdale: New Jersey, LEA (1983): "The Psychology of Human Computer Interaction".

[9] Zoltman, E and Chapanis, A, Behaviour & Technology, 1, pages 55 - 68 (1982): "What do professional persons think about computers?".

[10] Ravden, S.J. and Johnson G.I (1989): "Evaluating usability of human-computer interfaces: a practical method".

[11] Gilb. T. (1977): "Software Metrics".

[12] Whiteside, J., Jones S., Levey, P.S. & Wixon D., Human Factors in computer systems - II- proceedings of CHI'85 conference, San Francisco (1985): "User performance with command, menu and iconic interfaces".

[13] Bjorn-Andersen N., Eason K.D. & Robey, D. (1986): "Managing Computer Impact".

[14] Kidder L.H. & Judd C.M. (1986): "Research Methods in Social Relations".

[15] Fowler C.J.H & Wright, G - Penguin - (1986): "Investigative Design and Statistics".

[16] ETR 051 (1992): "Human Factors (HF); Usability checklist for telephones, Basic requirements".

[17] RACE: Common Functional Specification (CFS) PI30 (1992): "Usability Evaluation Methods".

[18] Oppenheim, A. N. (1966): "Questionnaire Design and Attitude Measurement".

[19] Short, J., Williams, E., Christie, B. (1976): "The Social Psychology of Telecommunications".

[20] Heinrich-Hertz-Institut Berlin (1989 - 1992): "Relevance of Motion Picture Presentation for the Aquisition of Information".

3 Definitions and abbreviations

3.1 Definitions

For the purpose of this ETR, the following definitions and explanations of key terms and concepts used within the main body of the ETR apply. The terms are printed in bold in the text when first mentioned. Where appropriate a reference is provided.

Attitude: complex aspects of the personality involving individual emotions, motivations and expectations. It is mainly dependent on user experiences and on user culture (COST 212 [1]).

Cognitive: mental, having to do with human information processing, involving thinking, learning and remembering.

Communication: exchange of information, according to agreed conventions, among two or more entities and among which the information itself has a meaning (COST 212 [1]).

Effectiveness: measures of the accuracy and completeness of the goals achieved (COST 212 [1]).

Efficiency: measures of the accuracy and completeness of goals achieved relative to the resources (e.g. time, human effort) used to achieve the specific goals (COST 212 [1]).

Expert error-free performance: shortest possible time to complete a specified task with zero errors, when the system is used by the designer.

Field study: investigation in a natural setting (in contrast to a laboratory setting), ("Bedrijfskundige methodologie: management van onderzoek" [2]).

Flexibility: degree of maintaining usability when user category, task, or environment is varied. The smaller the change in usability obtained by varying any of these three variables the higher the flexibility.

Function key: a key (e.g. ENTER or SEND) that causes the system to perform some predefined function for the operator.

Human factors: science studying the following different aspects regarding information services:

- physical aspect: ergonomics, environment;
- syntactic aspect: wording and commands;
- semantic aspect: man-machine dialogue, user-service impact, user procedures;
- pragmatic aspect: service efficiency, user needs, user acceptability, etc.;
- psycho-social aspects: behavioural situations, user expectation, influence on user attitudes etc." (COST 212 [1]).

Iterative design: repeated design process in which a prototype is tested by representative users as early as possible, then changed in a way that is supposed to prevent the occurring problems, and tested again (The iterative development of usable computer interfaces [3]).

Learnability: the differences in usability measured at various degrees of training for the same individual.

Human-machine communication: the communication from a single terminal, between a user on one side and a machine in the telecommunications network on the other side.

Human-machine interface: modality of exchange of information between the user and the system. The physical support of the user interface is the terminal equipment" (COST 212 [1]).

Inter/intra subject: see "reliability".

Menu: list of items from which a selection can be made.

Performance: degree of task achievement by utilizing a telecommunication service. Performance can be described by effectiveness and efficiency.

Pictogram: symbolic, pictorial representation of information.

Rating scale: scale by which numbers, commonly in simple arithmetic progression, are assigned to verbal descriptions of a factor which is assessed subjectively.

Reactivity: the degree of obtrusiveness of an evaluation method.

Reliability (of a subjective test):

a) intra-individual ("within subject") reliability refers to the agreement between a certain subject's repeated ratings of the same test condition;

b) inter-individual ("between subjects") reliability refers to the agreement between different subjects' ratings of the same test condition.

Replication: repeated application of the same test condition within a test in order to increase the reliability of the ratings.

Satisfaction: measures of the comfort and acceptability of a system or service to its users and other people affected by its use" (COST 212 [1]).

Subjective: pertaining to personal opinion or experience.

(Telecommunication) symbol: representation of a piece of information. The association of a symbol to its meaning has usually to be learned.

RACE: research and development in Advanced Communications technologies in Europe.

SUMI: Software Usability Measurement Interface.

System: terminals connected over a telecommunications network. The use of a system for specific tasks or a category of tasks forms the basis for a telecommunications service.

Task: unit of work to be accomplished.

Task behaviour: refers to the interaction of the system with the user within a work context.

Usability: the two components of usability are performance (objective) and user attitudes (subjective).

User: the person who interacts with a system.

Utility: the components of utility are the usability on the one hand, and the balance between the benefit and the financial cost of using a telecommunications system on the other hand.

Validity (of a measure): agreement between the mean value of measurements obtained in a test and the true value which the test purports to measure.

3.2 Abbreviations

For the purposes of this ETR, the following abbreviations apply:

CUSI	Customer User Satisfaction Inventory
ETR	ETSI Technical Report
ETS	European Telecommunication Standard
ETSI	European Telecommunications Standards Institute
GOMS	Goals Operators and Measures Methods
HCI	Human-Computer Interaction
HF	Human Factors
HFRG	Human Factors Research Group
HMS	Home Management System
IBC	Integrated Broadband Communication
I-ETS	Interim European Telecommunication Standard
ISO	Organisation for Internatinal Standardisation
QMS	Quality Management System
QUIS	Questionnaire for User Interface Satisfaction
RACE	Research and development in Advanced Communications technologies in Europe
SUMI	Software Usability Measurement Interface

4 Background

4.1 Introduction

This Clause provides the reader with some critical background information about the general approach adopted in producing this ETR. The two critical pieces of information are the definition of usability adopted and the overall approach or philosophy that underpins the ETR. The clause ends with a summary diagram that shows the interrelationships between the different concepts and issues explored in the later clauses.

4.2 Definitions of usability

4.2.1 The ISO definition

The concept of usability and its components has been defined within the Organisation for International Standardisation (**ISO)** (1992) as follows:

usability: "*the effectiveness, efficiency and satisfaction with which specified users can achieve specified goals in particular environments";*

effectiveness: *"the accuracy and completeness with which specified users can achieve specified goals in particular environments";*

efficiency: "t*he 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 others people affected by its use".*

ISO also notes that effectiveness and efficiency are often referred to as "performance measures".

4.2.2 The RACE definition

The Research and development in Advanced Communications technologies in Europe (RACE) programme is primarily concerned with the definition, design and specification of new Integrated Broadband Communication (IBC) systems and services. RACE recognises that these systems and services need to be usable. Usability is defined from the designers perspective (i.e. what designers need to do to ensure that usable systems and services are developed). From this perspective, some design guidance is offered:

- ensure that the appropriate enabling state exists for each of the users' goal tasks;

- reduce to a minimum the costs to the user in reaching the appropriate enabling states for their goal tasks.

A "goal task" is what the user wants to do, and an "enabling task" is what user shall do to create a state which enables the goal task to be performed.

The RACE approach complements the approach offered in this ETR. The ETR is concerned with evaluating usability rather than designing for usability. Although the two are related, it should be noted that following the RACE design approach cannot guarantee the development of a usable system or service.

4.2.3 The ETSI definition

Within **ETSI** the ISO definition has been generally adopted. ETSI, however, enhances the definition in the following ways:

a) **Usability** is considered as a pure ergonomic concept not depending on costs of providing the system. Usability together with the balance between the benefit for the user and the financial costs form the concept of **Utility.** This means that an ergonomical highly usable system may have low utility for a particular user who considers the cost to be too high in relation to his or her need for using the system. (This ETR deals only with the usability aspect, disregarding financial costs and user needs).

b) **Measures of usability** are assumed to be of two kinds:

1) performance measures, which are "objective" measures or observations of user behaviour and are focused on task performance, i.e. how well the user can achieve a specific task;

2) attitude measures, which are "subjective" measures or observations of the users' opinion of working with the system, i.e. how much they like to use the system.

These two measures of usability are considered to be orthogonal, i.e. they are mutually independent. This means that a system or a service can get a high score on one measure and a low score on the other measure. The two measures however can be dependent through sharing a common set of physical characteristics. These dependencies are often expressed through intermediary concepts (such as compatibility, redundancy or consistency), which themselves can be seen as a set of usability design principles (see BT Technol. Journal, 1993 [4]). The scores on these two usability measures may vary independently for a given system if either the task or the user category is changed.

Performance and attitude measures are also complementary in the sense that both contribute to the complete evaluation of the usability of a human/machine system. An assessment in both dimensions is therefore necessary, unless it can be shown that one attribute remains constant over different implementations of a concept. For example, if the performance of a terminal is not affected by certain changes in the human/machine interface, it would be sufficient to investigate differences in the users' attitudes only in order to evaluate the usability of a variety of solutions.

It is important to remember that the usability definition given here refers to a specific kind of task, user and environment. Usability in this sense cannot be generalized over different kinds of tasks, users and environmental conditions. Variations in these respects are expected to give different values of usability for the same system and require separate evaluations. The magnitude of such differences in usability may be considered as a measure of "flexibility", as is further explained below.

c) **Variations in terms of learnability and flexibility.** The concept of learnability is sometimes considered to be a factor contained within usability. According to the ISO definition**,** usability is restricted to specified users with their given abilities and experiences. The usability measured for an untrained user may therefore be different from that measured for the same individual after various degrees of training. Learnability can then be considered as a temporal dimension added to usability.

The difference between usability as measured at two different points of time during a training period may be used as a quantitative expression of learnability. The "learning curve", describing the usability as a function of the number of trials to learn specified functions can also be used as a measure of learnability.

In order to make a new system usable for a "naïve user" it is generally necessary to give some help in the form of an instruction or other guidance, which could be provided by different means, e.g.:

- written instruction, possibly including pictograms;
- function keys, designated by words or symbols;
- self-instructing menu;
- instruction by demonstration.

It should be noted that this guidance forms part of the service under consideration and consequently contributes to the overall usability of that service. Any changes in this part are likely to affect at least the performance for enabling tasks and should therefore be treated as one of the independent variables of the system characteristics.

Based on these definitions of usability and learnability, it is possible to set requirements (criteria for acceptance) on both dimensions with respect to the intended user categories. Requirements may be quite different for trained "professional users" and for naïve or occasional users, and the performance has to be measured with the intended users and in the relevant environment for each application.

Another variation that may be studied is how the usability of a system is affected by changes in the task and in the environment where the system is used, as well as between different users' categories. This variation can be measured by treating task, environment and users' category as independent variables**.** The resulting variation in usability is a measure of **flexibility.** A flexible system is able to accommodate a greater range of task, user and environmental variation. A less flexible system may be usable but may not be extendible to a wider range of users, tasks and environment then originally specified. Again, requirements on flexibility can be formulated and the system, or the service thus defined, can be evaluated in the same way as usability itself.

d) **Usability and telecommunications.** When usability evaluation is specifically applied to the use of telecommunications systems, it shall deal with a range of situations. One situation is end-to-end communication between two or more human users. A contrasting situation is human-machine communication from a single terminal between a user on one side and a machine in the telecommunications network on the other side (such as banking or information services), which is similar to any human-computer dialogue situation.

In the first case, the system consists of two (or more) terminals connected to a telecommunication network. The use of the **system** for a specific task or a category of tasks forms the basis for a telecommunications service. With the task oriented usability definition adopted here, the evaluation of a system used for a given task is the same as evaluating the content of the corresponding service. However, a complete telecommunication **service** generally involves both an enabling phase for establishing the connection, a transmission phase when the connection is used for the actual goal achievement, and finally a termination phase for disconnection, The evaluation of the usability of the service will then comprise the enabling task as well as the goal task (also called "device task" and "interpersonal task", respectively).

Furthermore, different implementations of a terminal that is part of a telecommunication system under consideration may in the evaluation process, be treated as variations in the characteristics of the system. Therefore, the object of usability evaluation methods will be designated by the word "system" throughout the ETR but will cover the evaluation of different terminal equipments as well as of complete services.

4.3 Types and approaches to evaluation

Traditionally, social scientists and educationalists have divided evaluation into formative and summative. Formative evaluation is concerned with assessing the process of developing or implementing a programme or system. Summative evaluation on the other hand is concerned with the evaluation of the system itself. For example, the process of producing this usability ETR could be evaluated quite separately from the product of that process (i.e. the actual ETR). Usability assessments are normally of the summative type.

All summative evaluations are attempts to demonstrate that a system, or representation of a system, (e.g. a design concept) possess some quality or attribute that can be defined as being of value to different users or stakeholders. These value attributes can be derived from a number of sources and this can lead to further distinction of different types of evaluation with the summative form. The four types are:

- criteria related evaluations;
- analytical evaluations;
- expert Evaluations;
- comparative evaluations.

In *criteria related evaluation*, criteria that the system needs to meet are usually specified early in the design stage and the system is then assessed at a later stage to determine whether or not it has met these criteria. As the criteria are specified at an early stage in the system development, the criteria are also important as design goals. Consequently this type of evaluation is most closely associated with the *usability engineering approach* (e.g. Whiteside et al, Handbook of Human Computer-Interaction 1988 [5]).

Analytical evaluations are dependent on some model or analytical process to generate the criteria. The model predicts what behaviour would be manifested if a system or more usually its user interface was built to a particular specification (e.g. Goals Operators and Measures Methods (GOMS). Because this approach can anticipate potential usability problems at an early stage it is often referred to as a *diagnostic approach*.

In some cases the model is not made explicit and a tacit model existing in the mind of an expert evaluator is used. In *expert evaluations* the evaluator "walks through", reviews or inspects a system or service drawing upon their expert knowledge and experience. This knowledge and experience is often encapsulated in *design guidelines,* styles or principles.

The final type, the *comparative evaluation,* is the most traditional and common form of evaluation. It usually occurs in the later development phases and involves the comparison of similar systems or different versions of the same systems. A number of dimensions are normally identified to form a common benchmark for making the comparisons. The evaluations are themselves usually undertaken in an experimentally controlled and standard situation to ensure that valid and reliable comparisons can be made. To achieve this level of control generally requires the adoption of established experimental techniques and, consequently, a *laboratory based approach* is commonly adopted. Many Consumer Associations apply the comparative evaluation approach when making recommendations for "best buy" amongst a set of competing products.

The four types also differ according to the participation of the users. The choice is generally between experts as virtual users and non-experts as subjects representing actual users. In the former case there are no subjects in the traditional sense; the expert inspects a design or specification noting problems and faults as they are uncovered. The latter case tends to rely on test subjects who are selected as being typical users of the products or services but have no expertise in evaluation and, therefore, need to be instructed and supervised during the evaluation process. Clearly, in both cases experts are involved in the design and execution of the evaluation study itself.

Further, <u>when</u> the evaluations take place can also differentiate the evaluation types. The distinction is normally made between evaluations that can be undertaken "early" or "late" in the development phases. Early evaluation techniques have the advantage of identifying potential usability problems at a stage where it is relatively cheap and simple to fix. In contrast, late evaluation could result in discovering usability problems after the system has been designed and implemented, which may result in expensive and prolonged redesign activities. On the other hand, late evaluations generally yield more valid results, because the context of evaluation, the interaction with the system and the system itself will be more realistic.

Table 1 below positions the types of the expert/non-expert and early/late evaluations in time dichotomies.

Table 1: Expertise by evaluation timing

	EARLY	LATE
EXPERT	Analytical and Expert Evaluations	Expert Evaluations
NON-EXPERT	Criteria related evaluations	Comparative Evaluations Criteria related evaluations

It is important to develop an evaluation process which maps onto the various key phases of the system development process. Consequently the evaluator should rarely choose a single approach, but should select a number of them to cover both early and late evaluation.

4.4 Summary

The structure of the ETR in relation to the evaluation process being advocated is illustrated in figure 1.

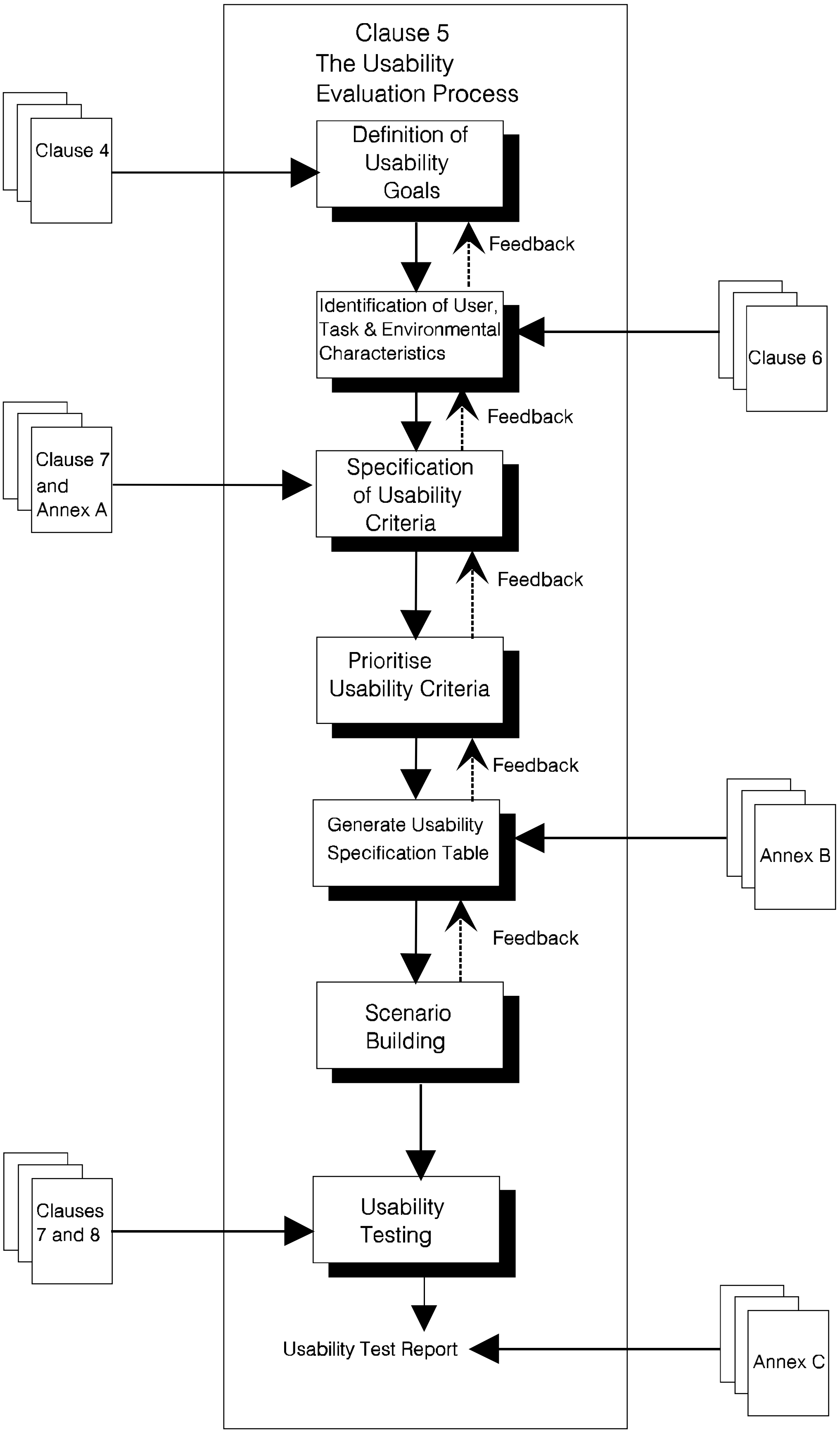


Figure 1: An overview of the structure and content of the ETR

5 The usability evaluation process

5.1 Introduction

The purpose of this Clause is to describe a generic process for undertaking criteria-related and comparative evaluations. Analytical evaluation processes are defined by the model or technique to be used, and expert evaluations are very dependent on the particular level and kind of expertise. The process is generic and is provided for guidance purposes only.

5.2 The usability evaluation process

Traditionally, usability testing, even when part of an iterative design process, has been fairly ad-hoc in nature. The aim of this Clause is to help formalise the usability evaluation process, and thus providing guidance on the procedures or steps required when following such a process. This is an attempt to focus on what exactly one is testing for, and what level of performance is expected to be achieved with the various parts of the system under test (see Usability Engineering in Office product development [6]). Specifically, the whole usability evaluation process involves the following steps:

a) definition of usability goals;
b) identification of the critical user tasks;
c) specification of usability criteria from those goals;
d) generation of a "Usability Specification" table;
e) prioritisation of criteria;
f) scenario building;
g) usability testing;
h) recommendations for improvement.

It can be seen from this list that usability testing is only one part of the process and is dependent on the successful completion of a number of other preceding stages.

There are two peripheral processes of confirmation and verification which should also be considered. A confirmation process checks with the users that the contents of any of the stages are an accurate reflection of their needs. For example, the scenario needs to be checked with users to ensure it is a realistic one. The verification process is more concerned with consistency and completeness of each stage with its preceding stage. For example, verification of step three (i.e. the specifying criteria) would ensure that a complete and consistent set of criteria are specified for a particular set of usability goals (step 1).

Figure 2 shows the relationship between the different steps and processes.

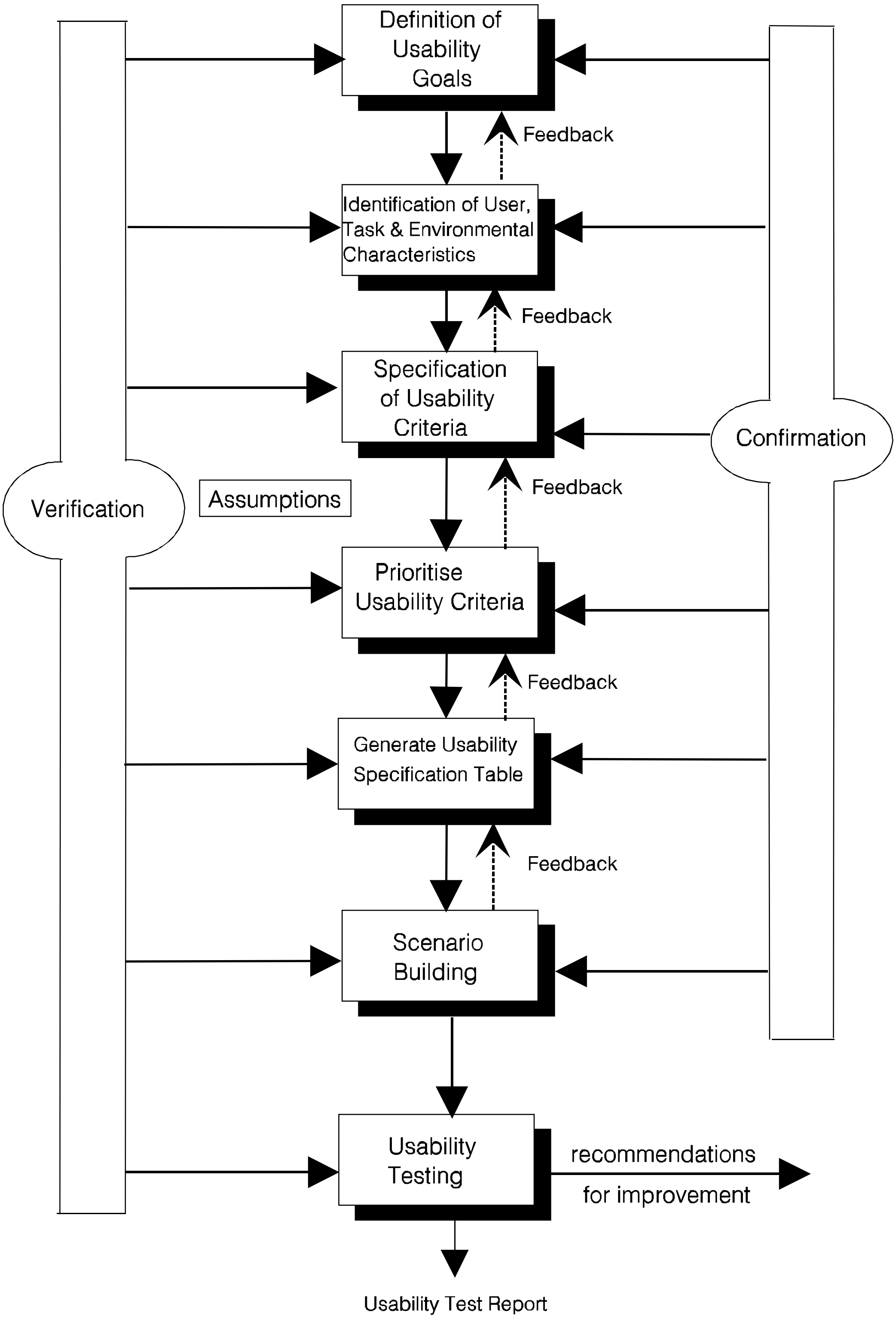

Definition of Usability Goals
Feedback
Identification of User, Task & Environmental Characteristics
Feedback
Specification of Usability Criteria
Feedback
Prioritise Usability Criteria
Feedback
Generate Usability Specification Table
Feedback
Scenario Building
Usability Testing
recommendations for improvement
Usability Test Report
Verification
Assumptions
Confirmation

Figure 2: The usability evaluation process

5.2.1 Step 1: Definition of usability goals

A usability goal is a desired end state which the system should meet in order to be judged as usable. At its most basic level, these goals are derived from the basic dimensions of usability which were described in Clause 4. These dimensions are often expressed as follows:

- effectiveness (of the system);
- efficiency (of the system);
- satisfaction (with the system);
- learnability (of the system);
- flexibility (of the system).

Goals can be expressed in either absolute or relative terms. Absolute goals are simply statements regarding some aspect of usability (e.g. to make the system easy to use). Relative goals are always expressed in a comparative form. These comparisons are often in relation to:

- current/existing methods of achieving the goal;
- the previous version of the system or service;
- a competitor's version of the system or service;
- a customer's or user's requirement.

Hence, the following are examples of relative goals:

- to make the system more satisfying to use than the earlier release;
- to make the system more effective than a competitor's version;
- to make the system easier to learn than the customer's current technology.

When using relative goals, it is important to compare "like with like" in specifying objective measures.

5.2.2 Step 2: Identification of the critical user, task and environmental characteristics

The second step requires the identification of the critical user, tasks and environmental characteristics. These characteristics are normally identified, described and analysed at the specification or requirements capture phase.

This step is important for expressing the key user needs which shall be operationalised into the usability criteria in Step 3. It also provides crucial descriptive information required for building realistic scenarios (see subclause 5.2.6, step 6), and selecting representative samples of users (see subclause 5.2.7, step 7). The step is not only important but quite complex as is Clause 6 which is devoted to giving a fuller description of the procedure.

5.2.3 Step 3: Specification of usability criteria

Usability criteria specify how any one particular usability goal is to be achieved. These criteria are expressed as statements against which a usability goal can be evaluated. Only if the system can satisfy the usability criteria can it be said to have achieved its usability goal. Normally, several criteria have to be met to achieve a goal. The procedure described below is based upon the work of Whiteside et al, Handbook of Human Computer Interaction (1988) [5].

Criteria can also be absolute or relative depending on the goal. An absolute goal demands absolute criteria and relative goals demand relative criteria. For example:

- to complete a specified task within 5 min (absolute);
- to complete a specified task on the first attempt (absolute);
- to complete a specified task 10% quicker than the previous version (relative);
- to complete a specified task in fewer attempts than the previous version (relative).

Deciding whether to use absolute or relative goals and criteria is often influenced by the practicalities of usability testing. For example, to test relative criteria involves the specification of benchmark measures against which improvements of the new system can be compared. This naturally involves considerably more data collection and therefore additional expense.

The danger in using absolute goals and criteria centres around the possible introduction of arbitrariness into the performance specifications. One way to minimise this is to use the notion of "expert error-free" performance as the benchmark for defining the criteria. Expert performance can be defined as "the shortest possible time to complete a specified task with zero errors, when being used by the designer of the system", and, by definition, is independent of usability or the quality of the user interface. Hence, the ratio of the expert user's time to the naive user's time is an indication of the ease of use of the interface. Verification and confirmation processes can also be used to minimise the arbitrariness.

In constructing usability criteria, it is crucially important that they are as complete as possible, and that any assumptions about the context, the performance criteria, or the measurement method are made explicit. Failure to do this may result in retrospective amendments to the criteria and cause difficulties in scenario building.

5.2.4 Step 4: Prioritising the criteria

In large scale evaluations (e.g. if the number of criteria exceeded 100), it may not be possible to test all the criteria and, therefore, some prioritisation of criteria may be necessary. Prioritisation is also important for supporting the decision-making process. The failure to meet a low priority criterion may been seen as less critical. Two types of prioritisation can take place.

The first type of prioritisation should take into account the relative importance of the different criteria. Relative importance of the criteria may be determined by their association with key functionality, by expressing a marketing requirement (e.g. when a system is "sold" on its "ease of use") or, criticality, to other functions or facilities (e.g. where access to a critical number of functions is dependent on achieving a particular usability goal).

NOTE: Criteria of low priority may be tested if they are built into a scenario associated with higher priority ones (see subclause 5.2.6, Step 6).

The second type of prioritisation should take into account the costs of the evaluation. The evaluation team may have specialist skills and expertise and the testing of criteria that require those skills and expertise may take priority. Equally, some methods for usability evaluation are more expensive to use then others. Therefore, there may be a number a practical constraints imposed upon the evaluation and these constraints would in turn influence important requirements placed upon the criteria.

In practice both cost and importance will influence prioritisation. In making the final decision it may be necessary to explore the trade-off between importance and cost constraints.

Finally, it is important to recognise that on some occasions it may not be appropriate for the evaluator to make the decisions about priorities. Their role may be simply to present the information to users and customers in a form that facilitates their decision making process.

5.2.5 Step 5: Usability specification tables

The criteria needs to be "operationalised". In other words details about exactly how they are to be assessed should be provided. The details to be operationalised include:

- the evaluation context (who is doing it, what are they doing, under what circumstances are they doing it);

- the performance criteria (what is the acceptable "worst case", "planned level", "best case", and, for relative criteria, "current level" of performance);

- the measurement method (the "metric" or "measure", "data collection technique", and "when & where" it is measured).

These details can be expressed in the form of a usability specification table. This table provides all the necessary information for scenario building. Annex B provides an example of a partially completed specification table for a Home Management System (HMS) which can be remotely accessed through the Public Telephone Network.

The first column of each table outlines the usability goal being addressed. These correspond to the usability goals given earlier. The next three columns cover the evaluation context, i.e. the person performing the task, a brief description of the task, and under what circumstances the task is being performed. In this example, an "information intensive individual" has been chosen as being a typical user of the system. The first time in the office has been chosen as the typical circumstances under which the tasks are performed. The performance criteria columns specify acceptable levels of performance in either absolute or relative terms, and occasionally both. In terms of "best case", timings have been based on the notion of "expert error-free" performance. "Planned level" and "worst case" timings are based on the "best case" time. The "current level" column is only used for relative goals and criteria which did not apply to this example. The measurement method columns simply specify the metric (e.g. time to open/close curtains), the data collection method to be used (e.g. observation), and the specific technique for data collection (e.g. timing device).

It is also important to note that the specification table can, or should, form part of the requirements definition, and therefore can be viewed as a set of design goals or objectives.

5.2.6 Step 6: Scenario building

The usability specification tables form the basis for the planning of the usability test. As it would be entirely unrealistic to present subjects with these criteria, it is necessary to assemble them into meaningful blocks or scenarios. A scenario represents a sequence or flow of user actions required to achieve a specific task. The scenarios should be based on realistic user tasks. Each scenario normally tests more than one usability criterion. Another important function of scenarios is that they specify exactly the nature of the tasks.

Only after a complete set of scenarios have been assembled, is it possible to perform the usability test.

In constructing a scenario there will be a feedback process with users whereby aspects of the criteria may have to be modified to ensure that a sensible scenario may be constructed.

As well as assembling tasks into scenarios, attention needs to be given to the attitude or satisfaction assessment criteria. These data can be collected through interviews, examining user protocols or, more commonly, by use of questionnaires.

5.2.7 Step 7: Usability testing

At this stage of the process it should be possible to design and conduct the usability testing. This includes the preparation and administration of the usability test. Much of the necessary information required for the test should have been completely and fully specified by this stage. The usability specification tables contain all the relevant information for the evaluator, the scenarios are the means of translating the evaluation context into a form suitable for the subjects, and the usability guide provides advice on selecting methods and analysing the results. It is also important at this stage to check the fidelity or state of the system being evaluated.

5.2.8 Step 8: Recommendations for improvements

The first stage in making recommendations for improvement is to clearly identify areas of concern. Unfortunately, usability evaluation does not readily lend itself to statistical decision making and some form of expert judgement is required. It would be optimistic to expect a system to satisfy every one of the criteria (unless the criteria were artificially conservative), hence the need for a judgmental stage in the evaluation process. This is one of the reasons for why it is important to prioritise the criteria before the test takes place.

This stage is relevant in such circumstances when not all the criteria are met. In some situations the failure to meet a particular criterion may affect whether or not a particular usability goal is satisfied. If it is not met, there could be grave consequences for the overall usability of the system, whereas in another case a usability goal could be judged acceptable despite the failure of a number of low priority usability criteria. In the latter case the system might be judged usable but with recommendations to rectify those minor weaknesses in, for example, a later release. Obviously, sufficient data would have been gathered from the usability testing to be able to identify any usability defects that arise through the failure to meet any particular criteria.

Once the judgements have been made and agreed, it is advised that a Usability Test Report is completed. This report would highlight the testing procedure, problems identified and the evaluators recommendations for improvement. This report is discussed in more detail in subclause 5.4.

5.2.9 Verification and confirmation

Running in parallel with the main process are verification and confirmation procedures. These are similar to verification and validation procedures adopted in software engineering (see "Software Engineering" [7]). Verification is normally undertaken by trained personnel (e.g. Human Factors experts) in order to review the goals and criteria to ensure that they were complete and consistent, and to check the process through which they were generated. In addition, the scenarios would be verified to check that the tasks within them were still compatible with the original criteria. Confirmation would be done by a representative selection of users. Essentially, this would ensure that, from the outset, the whole evaluation would be focused on the critical factors, and the usability of these factors. Also, part of the confirmation process is the prioritisation of the criteria. This is an important stage of confirmation because if, after evaluation, the system fails to satisfy all of the criteria, which is highly likely, it provides a means to judge whether or not the system can be accepted as usable.

The extent and rigour to which the process is subject to verification and confirmation is dependent on the size of the evaluation. For a "small" evaluation, it would be acceptable for verification and validation to be carried out informally by the evaluator; whereas for a large evaluation, this would be less appropriate.

5.3 Measures, metrics and methods

The purpose of this Clause is to describe what measures, metrics and methods can be used in usability evaluation. The scope is limited to a general introduction which highlights issues and concerns. A more detailed description is provided in Clause 8.

5.3.1 Measures

Measurement can be generally defined as the assignment of numerals to objects, persons or events according to rules. A critical issue is to determine what is to be measured. In usability evaluation it is the behaviour of a user operating a system that is usually measured.

A simple but useful distinction can be made between computer behaviour, user behaviour and task behaviour. The first two are self-explanatory. Task behaviour refers to the interaction of the system with the user within a work context. This assumes computers are tools for users to carry out tasks within a work environment via an application. In system terms, the user and system can be thought of subsystems operating within a total system.

With user behaviour and task behaviour the main measurement issues are concerned with performance. Performance is traditionally measured in terms of speed and accuracy. Speed is usually expressed in time to respond to an input or stimulus and accuracy in terms of frequency of errors made. Many of these measures are assessed against a criterion based upon some understanding or normative description of the processing and learning characteristics of the human or the machine. These criteria are norm-related and not task-related and so should be considered as task independent.

Examples of user performance measures are ratio of successes to failures, time spent in errors, percentage of errors, number of trials required to achieve error-free performance and number of good or bad features recalled by user. Although clearly the user has to perform on some task for these measures to be taken, the critical factor is that acceptable performance is not specific to that task - it applies to the system type. Therefore, the criterion for acceptable performance refers to a task independent norm or expectation. These norms are becoming formalised into standards, checklists or guidelines, and usability testing of norm-related criteria is therefore becoming concerned more with conformance.

The assessment of user and task behaviour can occur either early or late in the system development. The keystroke model (Card et al [8]), for example, can be used to specify acceptable response times and thus inform early design decision, and/or to later evaluate a pre-release version of the system.

User behaviour measures should also include subjective measures. These are normally measures of how users perceive the system and their attitudes towards it. The generic nature of these measure allow the adoption of standard questionnaires (e.g. Zoltman & Chapanis, Behaviour & Technology [9] or Ravden & Johnson, [10]). The value of a generic questionnaire is limited because the richness of the context cannot be captured. More specific task related subjective measures may be required. For example, it may be possible to extract subjective comments from verbal protocols or video recordings of task performance (e.g. number of times the user expresses frustration or satisfaction).

Task behaviour should include the user interface. Measures of clarity, consistency and other task independent concepts can be taken, and check lists of what to measure already exist (see Ravden and Johnson [10]). However, guidance on how to measure these concepts is less available.

5.3.2 Metrics

The measurement of interactive behaviours requires measures that shall reflect the overall context. This has led to the development of new "measures". These have been termed "metrics" by the Human-Computer Interaction (HCI) community (see Gilb.T, "Software Metrics" [11]) for a software engineering definition of a metric). The essence of usability metrics is that they reflect some task or work component. This is achieved through the metric possessing two essential components - a "measure" and a "task action". For example, there are simple metrics like "time (measure) to install (task action)" or more complex metrics containing a number of measure/action pairs expressed in a mathematical relationship (e.g. Whiteside et al's (1985) [12] Work Rate Metric, or Bjorn-Anderson et al's (1986) [13] Task Fit score).

Metrics should be designed to assess task performance. They are user and task dependent. The final choice is therefore be dependent on the tasks the system was designed to support. Unlike system or user behaviour assessment, it is improbable that a standard task evaluation package can be developed. However, a standard procedure should be developed which should include such stages as choosing an appropriate metric.

5.3.3 Methods

Methods are the form of the measurement (Kidder & Judd, 1986 [14]). They determine how the evidence should be collected. For behavioural research, the common methods are observation, interviewing, questionnaires, logging and secondary sources. These methods are usually associated with specific data collection techniques. For example, observation may use a checklists or category systems (e.g. Bale's Interaction Process Analysis); questionnaires can be open or closed; interviews can be structured or unstructured (e.g. clinical interviews).

Usability evaluation can also draw upon physiological methods. For example, measures of event related changes in brain potential and heart rate variability indicate differing levels of attention, or measures of electrodermal activity can reflect different levels of stress. These methods are often used in conjunction with the more traditional behavioural methods, thus providing a multi-method approach. Even if the intercorrelations both within and between physiological and behavioural methods are not always high, the methods are useful to support the interpretation of results from behavioural methods (e.g. reaction times) or to highlight inconsistencies.

Methods of data collection can be distinguished from the research formats required for their collection. The three typical research formats are the case study, the survey and the experiment (see Fowler & Wright, 1986 [15]). These formats can be adopted in a usability evaluation study, but their description lies outside the scope of this ETR.

Although usability evaluation draws upon a variety of different methods and approaches, it primary purpose needs to be made explicit. The purpose of evaluation is to determine whether some system or programme is of value to its developers, users, and/or to society at large. It differs explicitly from a traditional experimental approach, in that there is no theory (or hypothesis) to test. In the experimental approach hypotheses are generated, and operationalised so that they can be empirically tested. The purpose of undertaking an experiment is to advance a discipline through rigorous testing of its concepts and theories. The emphasis is on rigour, both in the generation of the hypothesis (ensuring that they are logically derived from the theory) and the design (ensuring that alternative explanations are reduced or controlled for), and this can result in highly constrained designs or contrived settings.

In evaluation studies the evaluators need to identify those aspects which are valued by the various stakeholders and translate them into criteria which can be demonstrably met (or not met). The purpose of evaluation is not to advance theory but to allow the interested parties to assess the value of decisions made or the value of the decision-making process.

5.2.4 Phases of development

There are many development models or life cycles to choose from. However, in the final analysis, most models fall into one of two classes - cascade or waterfall models, and evolutionary models. The cascade model is the more traditional model which can be idealised as having four main stages:

a) concept;
b) specification;
c) realisation;
d) operation.

In contrast, the evolutionary model is more cyclic and explicitly requires a rapid prototyping approach. The fidelity of the prototype is increased by adding in more and more detail on each cycle until the final system is evolved.

The role of prototyping in system development varies according to the philosophy of the developers. The evolutionary approach involves a high level of user participation. The user and developer evolve the design through a cycle of prototypes which gradually increase in complexity until the final system is agreed. The alternative approach is to view prototyping as one stage in a more traditional software development approach. The prototyping stage is viewed as a formal stage (i.e. realisation phase) where design decisions can be confirmed with users.

The major differences between the two approaches lies in the developer's perception of the importance of formal requirement capture and analysis. The evolutionary role assumes that users are not capable of articulating their requirements without being able to visualise the consequence of satisfying those requirements. The traditional approach views requirements and solution as separate, and prototyping is perceived to be useful for confirming solutions but not necessarily for determining requirements.

In the traditional approach, prototyping still has a role to play but that role is not one of evolving requirements. The prototypes are being developed to confirm requirements, design and implementation decisions. Decisions will be altered on the basis of the users' performance across a number of prototypes. After use, the prototype is "thrown away" and the design can be implemented using traditional coding methods. This is the approach being recommended by RACE.

From the perspective of the evaluation process, the traditional development model has a number of uses. It can be used to indicate what supplementary activities need to take place at the different points in the development lifecycle in order to support and ensure that valid and reliable evaluation takes place. A further important use is that different methods can be used to evaluate systems at different points in the lifecycle. This more precise use is described in more detail in Clause 7.

The development model adopted in this ETR is a four phase, cascade one, and these phases are described as follows:

a) Concept phase:

In this phase, the users, their current tasks and working environment are described and analysed to capture requirements for new services and new applications of existing services. The information may be collected from a variety of sources (e.g. marketing information), using a number of different methods (e.g. focus groups).

b) Specification phase*:*

In this phase, parameters of the new system as well as minimum usability requirements are specified. In order to get the necessary information for specifying requirements, experimental studies may be necessary, related to the physical techniques and logical parameters.

c) Realisation phase:

In this phase, the new services or systems are evaluated with users by, for example, emulation or by experimental use of prototypes, for the prospective tasks and under realistic environmental conditions.

d) Operation phase:

In this phase, a complete assessment of the usability of the new system or service is carried out under operating conditions, normally in field tests, in order to check up the final result on different users.

Table 2 describes the main preparation and testing activities, and the major outputs across a traditional development approach.

Table 2: Testing activities by development phase

	Development phases			
	Concept	**Specification**	**Realisation**	**Operation**
Preparation	Identify and Describe key users and tasks, and the usage context Produce Usability Test Plan	Produce the usability specification tables Review and revise Usability Test Plan	Identify key trial sites Review and revise Usability Test Plan	
Type of testing	User needs validation	Analytical Testing	Laboratory Testing	Field Trials
Outputs	Usability Test Plan	Specification Tables Usability Test Plan (updated) Usability Test Report	Usability Test Plan (updated) Usability Test Report	Field Trial Test report Usability Test Report

The key to all good evaluation lies in preparation, and the success of evaluation process is dependent on good preparation. In particular, there is a requirement for good descriptions of user, task and job characteristics from which the key usability goals can be generated.

The other issue worth highlighting is the need for careful planning. In this respect a Usability Test Plan should be produced which outlines the main testing events, their methods and their associated testing dates. The plan should be continually reviewed and updated.

5.5 Supporting documentation

To help ensure that evaluation testing process is carried out in a quality manner it is advisable to have some accompanying documentation. This enables plans to be recorded, decisions to be traced and actions to be tracked. It can also be useful contractual document. In many cases companies have their own Quality Management Systems (QMS) which may adequately cover the required documentation. However, a standard QMS may not be sufficient to cover the specifics of usability evaluation testing.

This subclause specifies the minimum that should be documented to ensure that quality is built into the usability evaluation process. It is structured around the requirements for three types of documents: a test plan; a test specification; and a test report.

5.5.1 The test plan

The test plan should be written at the early stages of the system's development. The plan should address the following:

- scope of the testing (including a brief description of the system; the key usability issues to tested; as well as some general identification information);

- testing details (including the type of test (e.g. expert walkthrough); the development phase in which the test should take place; some brief details about the tasks subjects are expected to undertake when using the system, the user/subject types (e.g. female managers), any critical user characteristics (e.g. no computer experience) and the required number of users/subjects in each type); a brief description of the usage context (e.g. at home), methodological issues (e.g. design - Latin square; analysis - ANOVA; methods - observation; techniques - logging) and details of the test schedule (when to be undertaken));

- resource details (including who is the project manager, principle evaluator and other members of the evaluation team; as well as financial costs for required human effort and equipment hire or purchase);

- a document history which records the plan's author, the date the document was issued; and status information (e.g. completed; slipped by two weeks, etc.).

Each organisation should consider how best to meet the above requirements. Annex A provides an example form for a test plan. In the ideal situation, the test plan will cover a number of tests required throughout the lifetime of the system's development. If this is not possible then the plan should be updated and reissued as and when the new information becomes available.

5.5.2 The specification tables

The specification tables provide the level of detail required to design and undertake a quality usability evaluation. It can be considered as a table with the following column headings:

- evaluation criteria - specify the usability goal that is being operationalised and tested;

- context - specify for a given goal who (subject) is doing what (task) under what circumstances (usage context);

- performance criteria - these are performance targets (e.g. "worst case"; "planned evel";"best case"; "current level");

- measurement method - specify the metric to be used, the data collection method and technique;

- a document history which record the plan's author, the date the document was issued; and status information (e.g. completed; slipped by two weeks, etc.).

This table should contain all the necessary information for the detailed planning and execution of the required usability trial. Annex B provides an example of a specification table.

5.5.3 The test report

The test report is a brief description of the test results and recommendations. It is a management summary rather than a detailed report. It is not, therefore, designed to provide a description of the data, their analysis and interpretation. It may be used as a front cover for such a larger and more detailed document. The following would be the minimum requirements for the test report:

- system and test identification information - including what is being tested, type of test, when and by whom;

- general findings by usability goals - some indication of seriousness of the problem should be provided;

- recommendations by usability goals - some indication of priority should be provided;

- a document history which record the plan's author, the date the document was issued; and status information (e.g. completed; slipped by two weeks, etc.).

A test report should be generated for each and every usability evaluation regardless of whether the same system was tested in each case. It should provide sufficient details to allow some assessment of "cost-to-fix" although the evaluators themselves may not have the expertise to provide the estimate. Annex C provides an example of a "Test Report".

6 Describing the context

6.1 Introduction

The context of use is implicit in the definition of usability as described in Clause 4 and is required to support Step 2 of the evaluation process described in subclause 5.2.2. In order to specify users, their goals and their environment as well as usability criteria, descriptive data are required. This Clause recommends what characteristics of the user population, the users' tasks and usage environment should be described. Such descriptions can be provided in terms of checklists (see, for example, ETR 051 [16]).

These descriptions are critical for generating the criteria against which a system or service may be evaluated. They are also crucial in providing realistic scenarios which are important for ensuring the construction of a valid test environment.

The approach is not a prescriptive one because what is described and how the descriptions are collected will vary according to the nature of the system and expertise of the evaluator. However, certain general characteristics can be identified and variations from the recommended procedure are noted.

6.2 The general procedure

A general procedure is required to manage the amount of data collected by describing the context. Clearly it is not advisable to describe all users and tasks. Guidance is provided about selecting the critical users/tasks that should be described in more detail and is based upon the work of [RACE/GUIDANCE & URM, ESPRIT/MuSIC & HUFIT, ALVEY/USTM]. The specifics of the procedure is given in the following sections and only an outline is provided below.

The general procedure for tasks and users is to:

- identify (all possible users or tasks);

- select (the critical user or tasks for detailed description);

- describe (the critical users or tasks).

The identification stage involves listing all the users and tasks. Once identified, the key users/tasks can be selected according to their importance to the success of the system or service. Only the critical users or tasks should then be described.

With respect to methods of data collection, a multi-method approach is advocated. This would include the use of secondary sources (e.g. existing system specifications), observations, interviews and questionnaires.

In terms of when the descriptions should be collected, it is strongly recommended that in all cases this occurs at the concept and/or the specification phases (i.e. is part of the requirements capture process). The descriptions provide the required characteristics of the users, tasks or environments and the evaluation assesses whether these characteristics are actually present. This is commensurate with the overall usability engineering approach that underpins this ETR. Early specification not only helps to ensure valid evaluations but assists in the ultimate design of usable systems.

6.3 Identifying and describing users

The identification of users can be usually derived from the market analysis for generic system or the business case for bespoke ones. Generic telecommunications systems normally address a specific segment need (e.g. systems for the Health Sector), or are designed for general public use (e.g. a domestic telephone). In the former case, the users can be specifically identified but this is not advisable in the latter case. The procedure will therefore differ according to the type of system/service which is being provided.

Where users can be identified they should be classed according to a general description of the user's job or role (e.g. secretary) rather then a specific instance of that class (i.e. the named individual). In the case of general public use the extremes of the user population should be specified. This should reflect the range or variability of the user population. It is particularly important to consider people with special needs within the acceptable range. In some cases it may be easier to define the boundary through exceptions.

For example, a boundary could be defined by stating that a particular system is not intended to be used by children under ten.

To recapitulate: for systems designed for specific market segments or usages, list the different types of users;

for general public use, produce a statement defining the acceptable boundary or exceptions.

The selection process can only be applied if a list of user types exists (i.e. in the first case above only). At this stage it is recommended that the list is reviewed and classified according to the importance of the system or service to the particular type of user. Users can be selected on a number of criteria, for example:

- distance from the system or service (e.g. frequent users, occasional users, infrequent users);
- impact of the system on the users' work (e.g. select those who will be required to undergo the greatest change in their work);
- political or economic importance (e.g. buyers, managing directors etc who may not use the system but who have an important say in what is included in the system or functional specification.

The number of criteria chosen is mainly dependent both on the type of system and any development constraints (i.e. available time, money, etc.).

Having selected the critical users or defined the user boundary, it is recommended that the users are described according to a predetermined set of characteristics. Again these characteristics will vary according to the type of system being evaluated. Some examples are given below:

- physical characteristics (e.g. dexterity, sensory levels);
- cognitive characteristics (e.g. memory load, attention span);
- expertise and Skill Characteristics (e.g. knowledge of other systems, typing skills);
- personality characteristics (e.g. levels of stress, interaction styles/preferences).

Again it is recommended that ranges are used and, where possible, these are tightly defined. It is these descriptions that can provide a major source for determining the key criteria for the evaluator.

6.4 Identifying and describing tasks

The identification and selection of tasks does not vary according to the type of system being developed (i.e. generic or bespoke).

The recommended procedure would be drawn upon process representations (e.g. flow diagrams) produced at the analysis stage and, if necessary, reproduce them in task hierarchies. The identification of the critical task requires the selection of the correct level within the hierarchy. The desired level should be at the interaction level where the user and system "share" tasks (e.g. "send fax" which requires both a system and user activity). Clearly, if the level chosen is too high (e.g. "prepare letter") or too low (e.g. "hit key") inappropriate tasks will be identified.

Only those tasks which are critical to the success of the system are described. Determining criticality is best left to the expert judgement of the analyst. However, the questions below may be a useful guide. If the response to any of these questions is positive it would suggest the task is critical.

- Is this task essential for the achievement of a given business function?
- Is the task considered to be difficult or stressful?
- Are high levels of knowledge and skills required?
- Is there a lot of variation in how and when the task is performed?

- Does the task have significant dependencies (e.g. inputs and outputs)?

- Does the user have a large say in how and whether the task is undertaken?

- Are there any identified resource issues with the task?

- Is responsibility for the task poorly defined?

- Are there motivation problems associated with doing this task?

Some of these critical questions can also be transformed into useful descriptors. Examples of this and other characteristics are given below:

- task dependencies (i.e. inputs and outputs);

- task frequency (i.e. how often is it undertaken?);

- task fragmentation (i.e. does it have to be completed without interruptions?);

- task importance (i.e. how critical is it to success of the user's job/business?);

- task knowledge & task skills;

- task preparation (i.e. what preconditions need to be satisfied?);

- task autonomy (i.e. how much acceptable variability is allowed in the task execution?).

Yet again these descriptions become an important source of evaluation criteria.

This analysis and description of tasks should be carried out for enabling tasks as well as goal task, as defined in subclause 4.2.2. Consequently, for a complete telecommunication service, both the establishing phase for getting access to the service, the transmission phase for actual communication and the termination phase for disconnection should be dealt with.

6.5 Identifying and describing the environment

The environment in which the system/service is to be used should be identified and described. This information should be contained in the system specification. If it is missing then it is necessary to collect it. In terms of environment it is important that the range of different environments are identified for the more generic systems.

Once the environments have been identified the most extreme and typical examples should be selected for description. Typical characteristics worthy of description are:

- physical characteristics (e.g. dirt, noise, etc.);

- social characteristics (e.g. levels of interactions, frequency and type of interruptions);

- characteristics of the technology (e.g. type of computers/telephones/etc used in the work place).

Such information is crucial for ensuring that the evaluation environment is an accurate reflection of the actual environment in which the system will eventually be used.

Finally, it is also important to note that user, task and environmental factors will interact. For example, a noisy environment (physical characteristic) may cause interruptions to tasks (task fragmentation) which may increase the stress levels of a user (personality characteristics).

7 Evaluation methods

7.1 Introduction

The purpose of this Clause is to provide guidance on selection and application of appropriate methods necessary for successful usability evaluation. Each method is described in terms of:

- constituent components;
- advantages and disadvantages;
- appropriateness for different development phases;
- application and analysis of results;
- conditions and preconditions of use.

The strengths and weaknesses of the methods are discussed with reference to the different development phases, and examples and illustrations are provided.

7.2 Choosing a method

The methods can be positioned along a number of key dimensions. These dimensions are important for selecting an appropriate method and are described below:

a) Reactivity - the degree of obtrusiveness of the method. Some methods (e.g. observation) are more likely to cause reactions in people which may effect the behaviours you are attempting to measure. Others are less obtrusive and, at best, are unnoticed by the users (e.g. logging).

b) Setting - the required naturalness of the setting may be important. The evaluator may require natural settings for data collection (e.g. field studies) or in other situations may prefer to sacrifice naturalness for higher levels of control (e.g. laboratory based studies).

c) Control - the amount of rigour and control required by the evaluator will affect choice of method. For example, an unstructured interview affords far less control than a highly structured questionnaire.

d) Costs - whether a method is cheap or expensive to apply will depend not only on the method (e.g. interviews are more expensive to apply than questionnaires because of the cost of the interviewers time) but also on the amount of time required for preparation, execution and analysis.

e) Sample size - some methods are designed to collect data from a large number of users (e.g. a questionnaire used in a survey), whereas others can be used to collect information from a single user (e.g. an in-depth interview in a case study).

f) Expertise - all methods require expert and skilled application. However, some require more expertise than others (e.g. logging requires considerable computing expertise).

In most cases the evaluator may wish to choose more than one method. The multi-method approach is recommended. It can be used to capture more than one perspective on essentially the same behaviour (i.e. within one of development phases) or to collect data across the different phases. Selecting methods on the basis of their appropriateness to the different development phases is described more fully in subclause 7.4.

7.3 General description of the different methods

The six most commonly used methods are described in table 3 below:

Table 3: General description of the methods

METHOD	DESCRIPTION
Secondary sources	Already existing data which were not collected by the evaluator or not produced by the subject for the purpose of an evaluation study.
Logging	The use of a recording device embedded or attached to the system or users to capture and record the users' performance.
Observation	The selection, recording and encoding of a natural set of behaviours through the use of audio visual techniques.
Questionnaire	A set of written questions requiring a written response which describes past behaviours, the users expectations, attitudes and opinions towards the system.
Interview	A set of spoken questions requiring a spoken response which describes past behaviours, the users expectations, attitudes and opinions towards the system.
Self-descriptive	A method of obtaining verbal protocols which reflect the users' thoughts or opinions about their interactive behaviour with the system.

Experiments are not included as a method in the above table. This ETR considers experiments to be a research format rather than a method. It is, therefore, in the same class as a case study and survey. This does not preclude experimentation being used in an evaluation study, and further details about the experimental approach can be found in appendix 1 of RACE's Common Functional Specification (CFS) P130 [17].

Table 4 summarieses the methods according to their suitability at the different stages of the development phases.

Table 4: Methods by development phases

	Development Phases			
Methods	Concept	Specification	Realisation	Operation
SECONDARY SOURCES				
Statistics and archival analysis	+	=	-	+
Clues analysis	+	=	-	+
LOGGING				
Frequency/Time	-	+	+	=
Error detection	-	+	+	=
Physiol. measures	-	+	+	=
OBSERVATION				
Direct	=	+	+	+
Indirect	=	+	+	=
QUESTIONNAIRE				
Open questions	+	+	+	+
Closed questions	+	+	+	+
Rating scales	+	+	+	+
INTERVIEW				
Structured	=	+	+	+
Unstructured	+	+	+	+
Group discussion	+	=	+	=
SELF DESCRIPTIVE				
On line comments	+	=	+	-
Off line comments	+	=	+	=
Playback comments	+	=	+	+
Legend: - (unsuitable)		+ (recommended)		= (possible)

7.4 The detailed description of the methods

7.4.1 Secondary sources

7.4.1.1 Definition and components

Definition

already existing data which were not collected by the evaluators or not produced by the subject for the purpose of an evaluation study.

Components

in general, for usability evaluations there are two classes of secondary source data. One class is based upon who collected the data and this can be referred to as statistical or archival data. In this case data were not collected by the evaluator. The other class is where data are collected by the evaluator but were not produced as part of an evaluation study. These data are often referred to as “clues” left by users which suggest or indicate certain types of behaviour. A more detailed description of the types of data that can be collected is given below:

1) Analysis of statistics or archives

- system Faults with existing systems (e.g. statistics on ‘down’ time; fault rates; maintenance costs);

- usability difficulties (e.g. statistics on productivity; training time);

- acceptability issues (e.g. statistics on absenteeism; sickness).

2) Analysis of clues

Made by the users, such as:

- wear on keys;
- dog-ears in the user manual;
- assignment and labelling of keys done by the users;
- annotations to manuals.

Or documentation produced by users, such as:

- memoranda;
- problem reports;
- correspondence;
- "crib" sheets.

7.4.1.2 Advantages and disadvantages

Advantages

- It is a quick and economical method for collecting and analysing statistical information.
- It is not time or location dependent.
- It captures behaviours from natural setting.

Disadvantages

- There are no experimental conditions, so the variables cannot be controlled.
- The number of variables available for describing "service usability" is limited and especially psychological factors cannot be measured.
- Archives are often incomplete and gaps or clues are often selective and survive according to their robustness which, in both cases, can make the data non-representative of the domain of interest.

7.4.1.3 Appropriateness for the development phases

It is recommended that this method be applied only in the concept phase (analysing the existing systems) and operation phase (comparing the new system with the previous systems). It is important, particularly in the concept phase of the development, to collect already available information for complying with the needs of the user, based on the technical and ergonomic characteristics of existing systems.

Table 5: Methods by development phases - secondary sources

Method/Phases	Concept	Specification	Realisation	Operation
SECONDARY SOURCES				
Statistics and archives analysis	**+**	**=**	**-**	**+**
clues analysis	**+**	**=**	**-**	**+**
Legend: - (unsuitable) + (recommended) = (possible)				

7.4.1.4 Application and analysis of results

This method is non-reactive and is generally used in exploratory studies to identify problems, and/or to identify and describe certain characteristics of users, their tasks and working environment. However, because the evaluator was not present at the time of collection and because clues are selective and therefore may be misleading, the data collected can be difficult to analyse and interpret.

7.4.1.5 Conditions and precautions of use

A few pitfalls need to be avoided:

a) the frequency of use should not be confused with ease of use. If the use is frequent, it might only show the intensity of a need. Conversely, a seldom used function could still be executed easily enough;

b) notes left by users on interfaces or documents are often an indication of usability problems with the system, but that does not necessarily mean that the difficulties will disappear after a mere application of other solutions, because they may be more of a corrective than facilitating nature.

7.4.1.6 Examples and illustrations

For the collection and analysis of "official" documents:

- organization charts;
- effective modes, intended procedures;
- instructions, rules, guidelines.

Other categories of document to investigate are the "informal" documents such as:

- modifications of procedures;
- code lists, abbreviation lists, macro-order lists.

Reference should be made to any project documentation outlining the requirements of the system and interface and the user's role. In addition, certain companies may have established codes and procedures to carry out work, and these should also be consulted. Consultation of users' training course material is also useful to gather a broad view of *"how it is meant to be done"* (although this is often different from how it is done in practice) especially:

- listings, screen copies;
- mistakes and failures reports, alert;
- already achieved study reports.

7.4.2 Logging method

7.4.2.1 Definition and components

Definition

The use of a recording device embedded or attached to a system or users to capture and record the users' performance.

Components

The components of the method can be considered from the view of the different type of performance measures which can be used in logging. These are:

a) 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 redial);
- how often certain types of tasks were carried out (e.g. number of national versus international telephone calls) or applications used (e.g. word processor or spreadsheet);
- how often certain types of errors occured.

b) 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”).

c) Error Detection Measures, for example:

- wrong keystrokes (which?);
- false or incomplete data entry;
- use of wrong or inappropriate procedures.

d) Physiological Measures, for example:

- amplitudes and latencies of event related potential components (e.g. P300);
- heart rate variability.

7.4.2.2 Advantages and disadvantages

Advantages

- 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 detailed data.

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

Disadvantages

- 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.

7.4.2.3 Appropriateness for the development phases

The logging methods can chiefly be applied in the specification phase (by emulation or simulation) and in the realisation phase (prototypes).

Table 6: Methods by development phases - logging

Methods/Phases.	Concept	Specification	Realisation	Operation
LOGGING				
Frequency measures	-	+	+	=
Time measures	-	+	+	=
Error detection	-	+	+	=
Physiological measures	-	+	+	=
Legend: - (unsuitable)	+ (recommended)	= (possible)		

7.4.2.4 Application and analysis of results

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. However, it is recommended that logging is always used in conjunction with other methods. 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.

Analysis of such potentially large amounts of data can clearly be improved through the use of relevant statistical packages.

7.4.2.5 Conditions and precautions of use

It is recommended that logging is used in conjunction with other methods. It should also be noted that experts with specialist skills in computing will be required to write and embed the required software.

The large amount of data captured also makes it critical for the evaluator to have clear and focused goals and, if possible, to avoid "data trawling" by capturing too much data.

7.4.2.6 Examples and illustrations

Examples and illustrations can be found in subclause 7.4.2.1 which describes the different component parts of the logging method.

7.4.3 Observation method

7.4.3.1 Definition and components

Definition

The selection, recording and encoding of a natural set of behaviours through direct observation or the use of audio visual techniques.

Components

Observation as a research method is generally considered to fall into two types: participant and non-participant. However with respect to evaluations it may be more useful to make a distinction between "direct" and "indirect" observation.

Direct observation requires no specialist recording equipment and involves the evaluator directly observing the user’s behaviour. Indirect observation requires some medium or device that comes between the observed and the observer. One-way mirrors and mechanical recording devices are typically used in indirect observations.

7.4.3.2 Advantages and disadvantages

Advantages

- Observation allows a wide range of behaviours (e.g. gestures or speech) to be captured within a range of different settings (e.g. laboratories or field). It is a very flexible and powerful data collection method.

- The use of the direct observation method is relatively economical to administer and allows the evaluators more opportunities to observe subtle effects which may be context specific (e.g. mood changes).

- The use of recording devices as an indirect method does not require the attention of the evaluator for the whole of the study; provides opportunities to playback recordings, and a permanent record of the evaluation study can be kept.

- The video recordings can be played back to developers to demonstrate particular usability problems.

Disadvantages

- Observation methods can be obtrusive with the presence of observers or recording devices.

- Recording data with direct methods can be difficult, and usually requires highly skilled observers.

- Indirect methods require specialist devices which can be expensive to acquire. Analysing the data collected can also be very time consuming (10:1 for analysis to recording) and expensive. It is not possible to directly observe attitudes, opinions, expectations, or judgements.

7.4.3.3 Appropriateness for the development phases

The observation method can be applied in all phases. Practically, video observation and recordings may be difficult if they require a special room such as a laboratory. In such circumstances, therefore, they are used only in the specification and sometimes in the realisation phases.

Table 7: Methods by development phases - observation

Method/ Phases	Concept	Specification	Realisation	Operation
OBSERVATION				
Direct	**=**	**+**	**+**	**+**
Indirect	**=**	**+**	**=**	**=**
Legend: - (unsuitable) + (recommended) = (possible)				

7.4.3.4 Application and analysis of results

Observation methods are particularly useful in the later stages of the product's development when a high fidelity prototype or an early version of the system is available for evaluation. They can be used to collect valuable descriptive data on what the users do and how they do it. These data are invaluable for checking protocols, discovering learning difficulties as well as uncovering general usability problems. They are usually carried out for a limited sample of users and over a limited time period either in laboratory or in field situations.

Data analysis can be complex and time consuming. It inevitably involves extracting the required measures through, for example, frame-by-frame analysis to elicit timings or the coding of observational notes. Video recording analysis can be supported by specialist video analysis suites, but these are not commonly available and are expensive to acquire.

7.4.3.5 Conditions and precautions of use

In order to produce valid results, this method, as with other methods, requires careful planning including:

- obtaining the agreement of the user (after explaining the aims and what is at stake);
- avoiding interfering with the task (by avoiding any interruption of the task performance where you sit near the user);
- offering feedback to the user (by giving the results of the completed analysis);
- preserving the anonymity of people, and explaining the procedures to be followed.

A few pitfalls to be avoided are:

a) observation requires some rigour in order to achieve a description and evaluation of the situation;
b) carefully separate the phases of observation and the phases of interpretation. Elaborating observation grids helps overcome this difficulty;
c) motivation for undertaking observation can, particularly when taking place in the work place, be confused for "time and motion" or similar less acceptable studies. In such situations the work force is unlikely to co-operate with the evaluator.

7.4.3.6 Examples and illustrations

The observation requires the elaboration of observable variables (such as changes of places, taking informations, actions, etc.), and should make it possible to explain their interactions. Some criteria to be considered in a grid:

- effective ordering of procedures;
- gestures, positions, changes of places;
- communications;
- results of the activity;
- facial expressions, emotional indices;
- taking information (direction of looks).

All these categories of facts have to be considered during the observation phase. Yet their importance depends on the nature of the activity, and on the final objective of the observation.

7.4.4 Questionnaire method

7.4.4.1 Definition and components

Definition

A set of written questions requiring a written response which describes past behaviours, the users expectations, attitudes and opinions towards the system.

Components

a) Open questions
    - with free responses.

b) Closed questions
    - with preworded multiple choice responses;
    - with responses on given rating scales.

7.4.4.2 Advantages and disadvantages

Advantages

- The method is cheap and easy to apply to a large sample of users.
- It can quickly provide both quantitative and qualitative data.
- The administration and analysis of questionnaires can be delegated.
- For the user, there is no time restriction for the answer.

Disadvantages

- The usually low rate of returns for postal questionnaires may result in an unrepresentative sample.
- The questions cannot be explained in more detail if necessary.
- The evaluator cannot always control the situation or the manner in which the questionnaire is answered.

7.4.4.3 Appropriateness for the development phases

Questionnaires can be used in all phases, especially in the concept and the operation phases.

Table 8: Methods by development phases - questionnaires

Method/Phases	Concept	Specification	Realisation	Operation
QUESTIONNAIRE				
Open Questions	+	+	+	+
Closed Questions	+	+	+	+
Rating Scales	+	+	+	+
Legend: - (unsuitable) + (recommended) = (possible)				

7.4.4.4 Application and analysis of results

Questionnaires can be used to collect data about reactions to a human-machine interface or specific features of a complete system. The purpose of such a survey is usually to obtain a quantitative description of different aspects under investigation and to compare several groups of subjects . They are generally used before, after, or before and after but rarely during an evaluation because this disrupts task performance. The structure and content of a questionnaire naturally depend on the respective system under evaluation. Below are some of the reasons for using a questionnaire:

- to assess 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.

7.4.4.5 Conditions and precautions of use

Although it seems relatively easy to construct and administer a questionnaire, it does require specialist skills to ensure its validity and reliability as a measuring instrument (see Clause 8). It is, therefore, important that questionnaires are tested in pilot studies before being used in the main evaluation study.

Care is needed to ensure that:

- there is a precise and identified purpose;
- clear and precise questions have been formulated;
- the administration conditions have been well-defined.

With respect to analysing questionnaires, care needs to be taken with open questions that they can be coded if any form of quantitative analysis is required. Equally, care needs to be taken before assuming that interval scales have been used. In most cases it is safer to assume that ordinal data has been collected and analyse it using the appropriate non-parametric statistical tests (see Clause 8).

7.4.4.6 Examples and illustrations

Most examples of standardised tests have been developed to assess the usability of software. It is probable that such tests can be amended to satisfactorily deal with telecommunication systems. Below are descriptions of two measures which adopt the questionnaire method.

1) Software Usability Measurement Inventory (SUMI)

SUMI has been developed as part of the MUSiC project (ESPRIT) by members of the Human Factors Research Group (HFRG), University of Cork. It is essentially a 50 item attitude questionnaire which takes about 10 minutes to complete. It provides three measures:

a) an overall measure which is a single score that provides a global assessment of usability. It is particularly useful for quick comparisons between similar products or different versions of the same product;

b) a "usability profile" which reflects five main dimensions or subscales (i.e. affect, efficiency, helpfulness, control, and learnability);

c) as the final output, a list of those items where the software or system is rated significantly better or worse than a comparative standard or norm. The output is produced by a procedure referred to as "item consensual analysis".

SUMI was developed using traditional psychometric techniques to ensure its validity and reliability. However, it has not been validated for telecommunication systems or services.

2) Questionnaire for User Interface Satisfaction (QUIS)

QUIS was developed as a measure of user satisfaction. Again it was specifically developed with software products in mind. QUIS consists of 90 questions and can be purchased with a software package (for IBM PC) to facilitate analysis. Studies have confirmed that it is both a reliable and valid measure of user satisfaction with the human-computer interface. It was developed at the Department of Psychology, University of Maryland, USA. A similar questionnaire, "Customer User Satisfaction Inventory" (CUSI) has also been developed by HFRG, University of Cork.

7.4.5 Interview method

7.4.5.1 Description and components

Definition

A set of spoken questions requiring a spoken response which describes past behaviours, the users expectations, attitudes and opinions towards the system.

An interview is the method most likely to reveal clues on why there is resistance to the use of a technology or why there are obstacles to its acceptance. Verbal exchanges with the affected users are necessary in order to check the relevance and usefulness of the information collected on the service, or to collect data about the user's own perception and feelings towards the system. The interview method needs to vary according to the required degree of formality.

Components

a) Structured interviews

Well defined and focused interviews by the use of prespecified questions. This is similar to a questionnaire except that the evaluator reads out the questions and the respondents provide a verbal reply.

b) Unstructured interviews

General topics or issues are raised with the interviewee to initiate a wider verbal response.

c) Group interviews

These are generally unstructured collective interviews, with the interviewer as a discussion leader for a group of interviewees.

7.4.5.2 Advantages and disadvantages

Advantages

- In face-to-face situations this method allows the capture of important non-verbal cues.
- It allows the evaluator the opportunity to explain or clarify questions.
- It provides the evaluator with the freedom to explore interesting topics in more depth.
- It can capture qualitative as well as quantitative data.

Disadvantages

- It is not easy for the interviewer to remain neutral while facing the user.
- It is difficult to avoid interviewer effects such as: age, sex, appearance.
- It is difficult to standardise across a number of different interviewers and interviews.
- A structured interviews is more expensive than using an equivalent questionnaire.

7.4.5.3 Appropriateness for the development phases

The interview is a common method and can be used in any situation at any phase, especially regarding the collection of data on attitudes and opinions about the issues under evaluation.

Table 9: Methods by development phases - interview

Method/Phases	Concept	Specification	Realisation	Operation
INTERVIEW				
Structured	=	+	+	+
Unstructured	+	+	+	+
Group	+	=	+	=
Legend: - (unsuitable)	+ (recommended)		= (possible)	

7.4.5.4 Application and analysis of results

The interview is usually administered by a single person, and should be limited to not more than one hour. The interviews may be individual or collective (group interviews). The results obtained being different, individual interviews are often more centred on the achievement of the task, and usually collective interviews are better for the organisation of work and group dynamic interaction. They are particular useful for:

- providing general descriptions of the usage context;
- clarifying and/or validating data collected from the use of other methods;
- assessing subjective judgements, opinions or attitudes.

The analysis of structured interviews is similar to that of a questionnaire. Unstructured interviews are more time consuming to analyse because pre-coding is not normally possible.

7.4.5.5 Conditions and precautions of use

In undertaking an interview it is advisable to take the following into consideration:

- make sure at the start that the interviewees understand the purpose of the interview, why they have been selected, and what the information will be used for;

- assure the interviewees that any information collected will be treated in the strictest confidence. This is especially important if any mechanical recording device is being used (e.g. tape recorder);

- be non-evaluative by stressing that it is the system and not the user that is being evaluated.

Some variables may influence the results, such as:

- the personality of the interviewer (sex, appearance, feeling);
- the way to conduct the interview (suggestive replies, etc.);
- the status of the interview (public, commercial);
- the presentation of the purpose of the interview;
- the technique of taking down notes (recording, pencil and paper);
- the place of the interview; the time allowed, and time exceeded,...

A few pitfalls to be avoided are:

a) asking leading questions;

b) having the interviewer's motives misunderstood - the interviewer must be seen as neutral or impartial;

c) the interviewee misunderstanding the interviewers motives and thus providing inappropriate or no answers to the questions.

7.4.5.6 Examples and illustrations

Contents of an interview grid:

- Can you explain to me what you do, what your activity is?
- How and since when have you learnt to use it?
- Is it always the same, the same procedure?
- What are the most important tasks?
- What are the most frequent tasks?
- Are there more sensitive phases?
- What solution do you perceive ?
- Are there difficult procedures?
- How do you do to...?

7.4.6 Self descriptive method

7.4.6.1 Definition and components

Definition

A method of obtaining verbal protocols which reflect the users' thoughts or opinions about their interactive behaviour with the system.

Verbal data describing task behaviours can be collected using a method which does not require the presence of an interviewer or use of a questionnaire. Verbal comments in this situation are linked to the users' real time performance, and focus on their reasoning, strategies, and procedures more than on the users' attitudes and opinions about them. The users are considered as experts on their tasks and are asked to demonstrate how and why they build their own strategies.

Components

a) On line: user comments whilst performing the task - running commentaries or *"thinking aloud"*;

b) Off line: user comments just after performing the task - *review*;

c) Playback: verbal Comments during a video recording replay - *autoconfrontation*.

Variants: reciprocal co-operation whether by *joint discovery* (naive user with naive user, or common use and search of results), by *transfer of learning* (expert user with naive user or results under guidance), or *common confrontation* (expert user with expert user or talks about different shortcuts and strategies). All variants require pairs of users who are more or less skilled in expressing themselves. The role played by the evaluators consists simply in taking notes of the users' verbalized opinions and feelings.

7.4.6.2 Advantages and disadvantages

Advantages

- Provides insight into cognitive activity, in particular the user's reasoning or thought patterns and mental models.

- The relevant information can be checked and clarified by the users.

Disadvantages

- The provision of concurrent protocols (or commentaries) can interfere with the task performance.

- Users have to be articulate and fluent and may need training in order to provide a usable self-description.

- Retrospective protocol analysis whether with or without the use of video recording equipment are prone to post-hoc rationalisations by the subject.

7.4.6.3 Appropriateness for the development phases

This method is particularly powerful when some representation or prototype of the system can be provided. In such circumstances appropriate task scenarios can be developed which the user's protocol can describe.

Table 10: Methods by development phases - self descriptive

Method/Phases	Concept	Specification	Realisation	Operation
SELF DESCRIPTIVE				
On line comments	+	=	+	-
Off line comments	+	=	+	=
Playback comments	+	=	+	+
Legend: - (unsuitable) + (recommended) = (possible)				

7.4.6.4 Application and analysis of results

A powerful method for identifying and describing current work tasks and usability problems with prototypes or early versions of released products.

Analysis can be time consuming as it often involves transcription and coding of the protocols.

7.4.6.5 Conditions and precautions of use

This method is most powerful when applied to the evaluation of some form of representation of the system. The minimum requirement is usually a high fidelity prototype.

Users are normally required to follow a number of specific task scenarios. The subjects should be sufficiently naive to avoid the elicitation of expert or automated behaviours which are difficult to verbalise.

7.4.6.6 Examples and illustrations

Contents of autoconfrontation verbalisation:

- What are you doing at the moment? --> user answer;
- Why do you do (say) that? --> user answer;
- Why do you do it this way? --> user answer;
- What is happening there? --> user answer;
- What is troubling you? --> user answer;
- How do you know that...? --> user answer;
- Where is the problem...? --> user answer.

8 Measurement issues

8.1 Introduction

The purpose of the Clause is to enable the evaluator to make an informed choice of what measurement scales to use. To achieve this purpose the evaluator not only needs a description of the different scales but needs to also understand notions of validity and reliability, and have a general awareness of some of the difficulties of measuring psychological variables. Finally, some applications of the scales are described. These applications have been restricted to the use of rating scales within observation, interview and questionnaire methods.

8.2 Measurement theory

Measurement can be generally defined as:

The assignment of numerals to objects, persons or events according to rules.

The rules vary according to the level of measurement which are represented by the four main types of scales as described in subclause 8.3 below. However, measurement of human behaviour does present certain problems not normally met in the natural sciences. To achieve precise and literal measurement requires an isomorphic or one-to-one relationship between theories being tested and the mathematical system adopted. In evaluation we have to infer the usability of the system from the behaviour of the user. This situation does not allow direct or literal measurement and certain assumptions have to be made (i.e. the observed behaviour measured is reflecting the usability of the system). Essentially, our concepts of usability are imprecise but our measures are not.

The evaluator needs to be careful in the choice of measurement scale. The choice needs to be appropriate to allow accurate and reasonable interpretation to be made of the evaluation results.

8.3 Measurement scales

Before choosing an appropriate scale, the evaluator needs to be aware of what is available.

8.3.1 Nominal scales

A nominal scale, in its simplest form, may be just a classification scheme (e.g. "gender": male, female). The assignment of a qualitative label, although not strictly a measurement, does allow the evaluator to count occurrences and therefore record frequencies.

The following properties apply to nominal scales:

- mutually exclusive: i.e. a particular observation cannot belong to more than one category;
- exhaustive: i.e. all sub-classes of object or events need to be included. (This is often achieved by the inclusion of an “other” category).

Care needs to be taken when numerical codes are assigned to categories (e.g. 1 for females; 2 for males). This assignment is purely arbitrary and, therefore, the application of arithmetic operations (e.g. adding, subtracting, division, etc.) to these numbers would be meaningless.

8.3.2 Ordinal scales

An ordinal scale is a rank order scale where the distances between points on the scale are not specified. An example is a Preference scale where test items can be rank-ordered; "best, next best, third best,.....", etc...

Ordinal scales therefore obey the following rules:

- "greater than" between classes;
- "equivalence" within a class.

8.3.3 Interval scales

An interval scale has equal distances between its points but has no true or absolute zero point. Distances on this type of scale are assumed to be proportional to perceived distances between stimuli. Scale values may be transformed by linear operations only.

Consequently, the magnitude of a response that has been assigned the number of 4 is not necessarily twice the magnitude of a response corresponding to the number of 2.

Interval scales subsume the rules for ordinal scales and, additionally, allow precise information on the magnitude of the interval between classes being measured. This makes it meaningful to apply arithmetical operations to interval data.

8.3.4 Ratio scales

A ratio scale represents the highest scale level and subsumes the properties of all other scales. It has an absolute zero point, and positions on the scale have the same mathematical relations as real numbers. The existence of a true zero point provides ratio scales with an additional property:

- any two points on the scale may be determined independently of the unit of measurement.

For example, a woman who is twice the height of her daughter remains so regardless of whether their heights are measured in inches or metres.

Ratings on these scales are generally comparative and can be done by:

- rating stimuli against an external reference point;

- comparative ratings between stimuli.

Figure 3 provides a summary of the characteristics of the different scales.

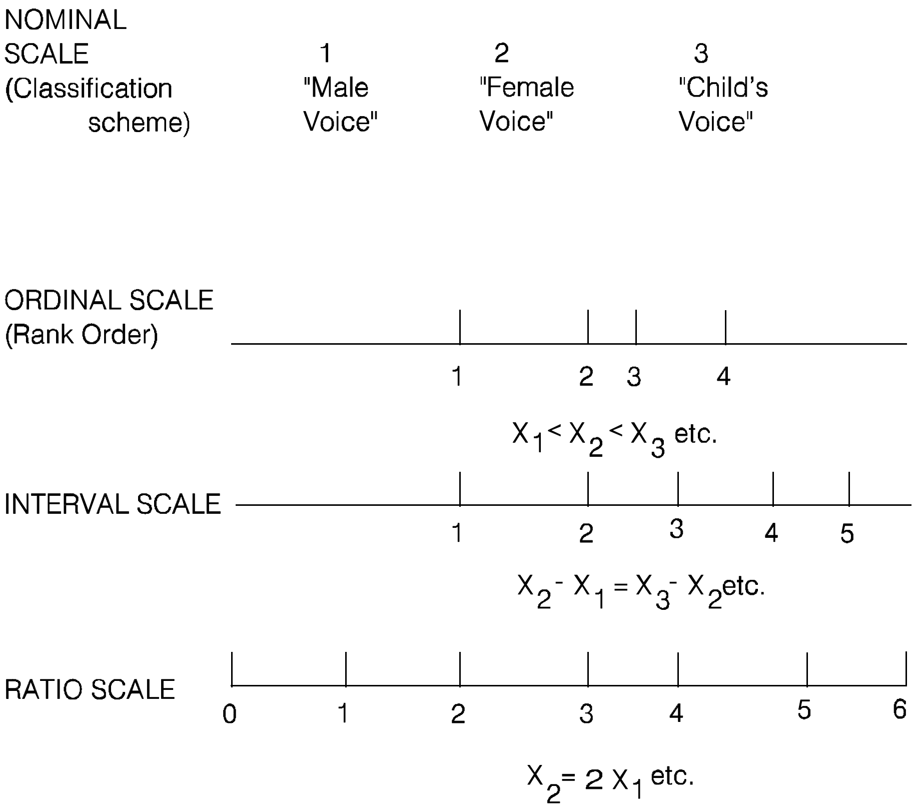


Figure 3: Characteristics of the different scales

8.4 Applying the scales

Scaling methods are used for rating individual users' sensations and opinions in one or more defined aspects (dimensions). Each subject is used as a kind of measuring instrument. The response when exposed to a single physical stimulus, or to all stimuli in a complex working situation, is considered as a quantitative measure of the subject's judgement of his or her own feelings and thoughts. It is, therefore, a method for quantifying sensations, opinions and attitudes.

For usability evaluation, scaling methods are generally used to obtain attitude measures, although they may also be used for rating task performance.

The principal application is for tests under controlled experimental conditions, where the working conditions during the use of a telecommunication service can be emulated and different implementations of terminal equipment be presented to the users (prototyping). For this application, 5- or 7- grade scales with both numbers and verbal definitions are commonly used.

Another application of rating scales is as part of a questionnaire. In this case category scales or pictograms scales are more often used. They should primarily be considered as ordinal scales, and not interval scales.

8.5 Conditions and precautions of use

It is important to carefully define the perceptual dimensions that the subjects are instructed to rate on. Verbal labels should be comparative if the scale is undirectional, or strictly opposite if the scale is bipolar. If words belonging to different dimensions are used as labels on the same scale, the subjects may be confused and the ratings tend to become inconsistent.

It is also important to check the reliability of the subjects' responses. The procedure for doing this is somewhat different for direct rating of stimuli and for ratings in questionnaires.

In the case of assessing reliability through the direct rating of stimuli it is necessary to follow good experimental practice and present the stimuli in random order and have more than one **replication** of the subject's responses. The intra subject variability across a number of replications (e.g. 3 or 4) provides an indication of the internal consistency and thus **reliability** of the scale.

When rating scales are used in questionnaires, the corresponding method for checking consistency is to repeat each question several times in a test with a slightly different wording. In this case, however, it is not possible to calculate reliability in a strict statistical sense. Furthermore, scales in questionnaires are most often category scales without numerical grades or anchors and should be dealt with accordingly.

The **validity** of a scaling method is much more difficult to assess than the reliability, especially if the field of application is new. It is therefore, recommended to use well established methods (e.g. see publication by A. N. Oppenheim [18]), and to limit their application to areas for which the validity has been empirically tested. A number of such methods are standardised internationally and should primarily be used.

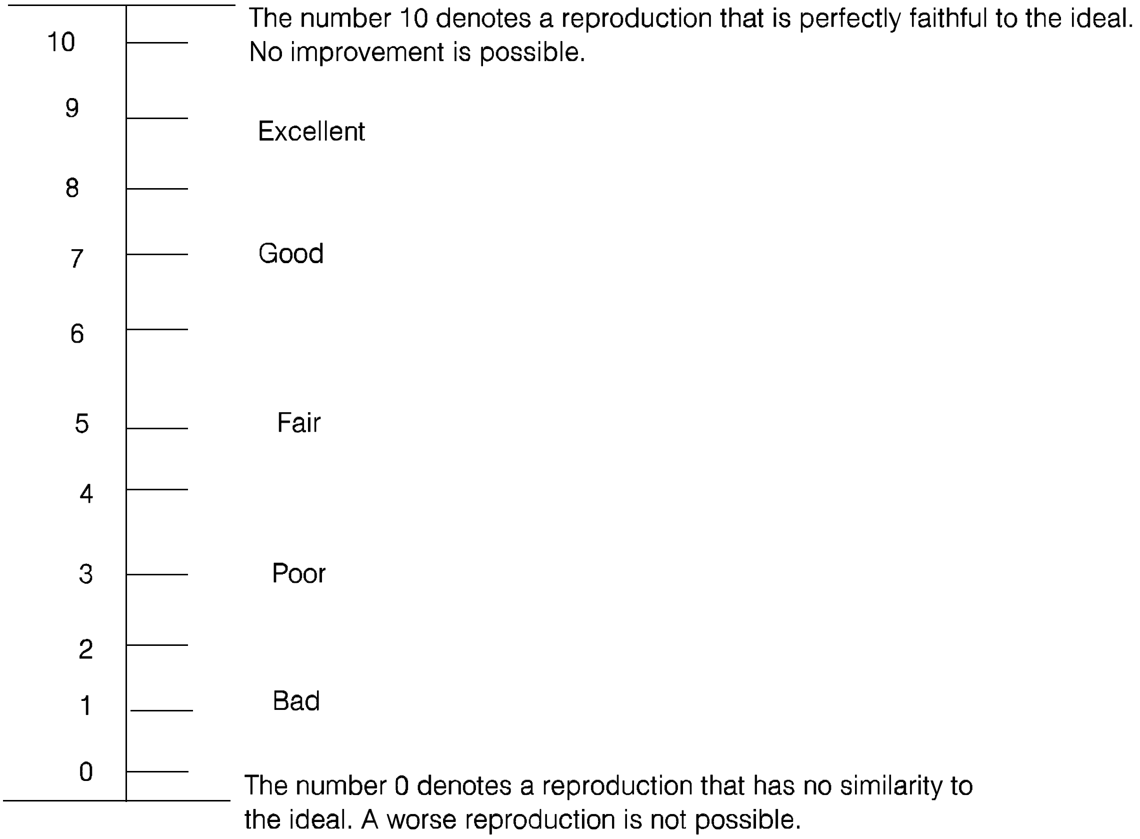
Subclauses 8.8 and 3.1 provide definitions of some key validity and reliability concepts and terms.

8.6 Examples and illustrations

Some examples of scales, both unidirectional and bipolar, with different grades, verbal labels or symbols are given in figures 4a) to 4f).

Excellent	
Good	✓
Fair	
Poor	
Bad	

Figure 4a): An example of a category scale (ordinal)



The number 10 denotes a reproduction that is perfectly faithful to the ideal. No improvement is possible.

The number 0 denotes a reproduction that has no similarity to the ideal. A worse reproduction is not possible.

NOTE: The endpoints of the scale serve as anchors which may verbally defined separately as indicated in the example of a loudspeaker listening tests.

Figure 4b): Numerical interval scale with verbal labels and anchors

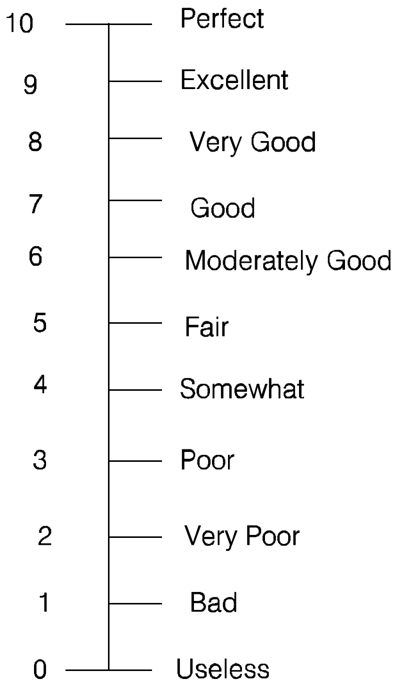


Figure 4c): Eleven point opinion scale (interval) applied for rating speech transmission quality

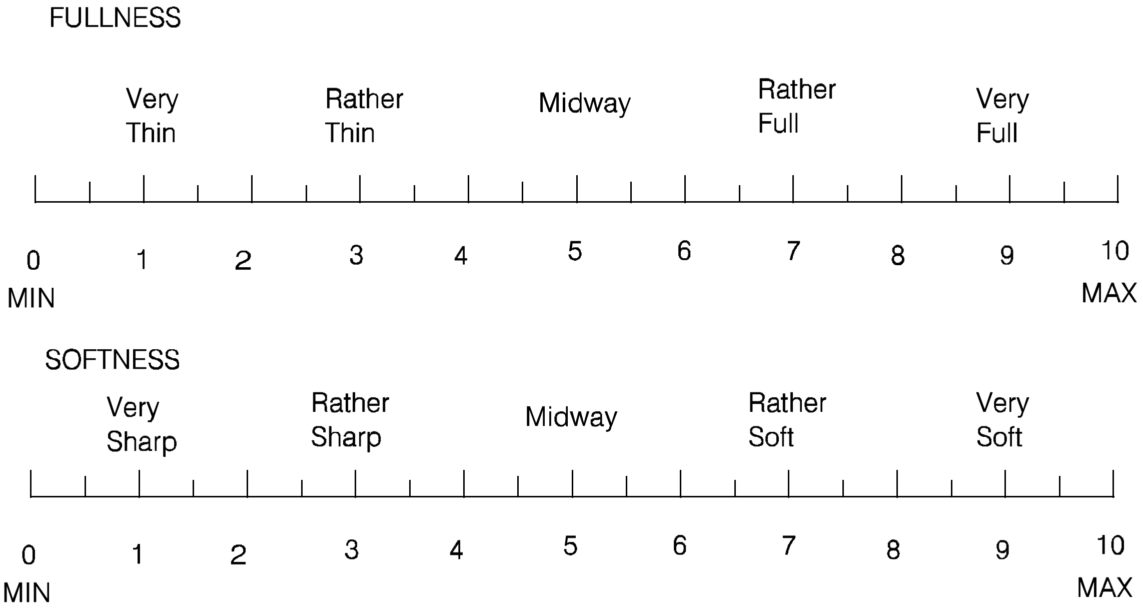


Figure 4d): Bipolar interval scales used for rating specific dimensions of sound quality

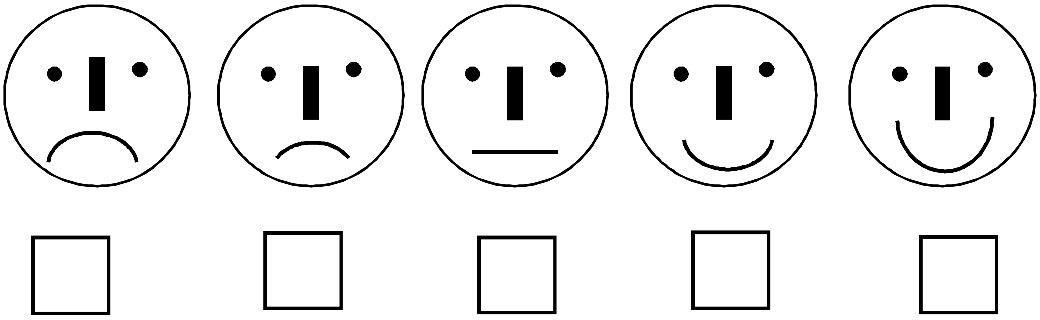


Figure 4e): KUNIN faces for measuring satisfaction on an ordinal scale

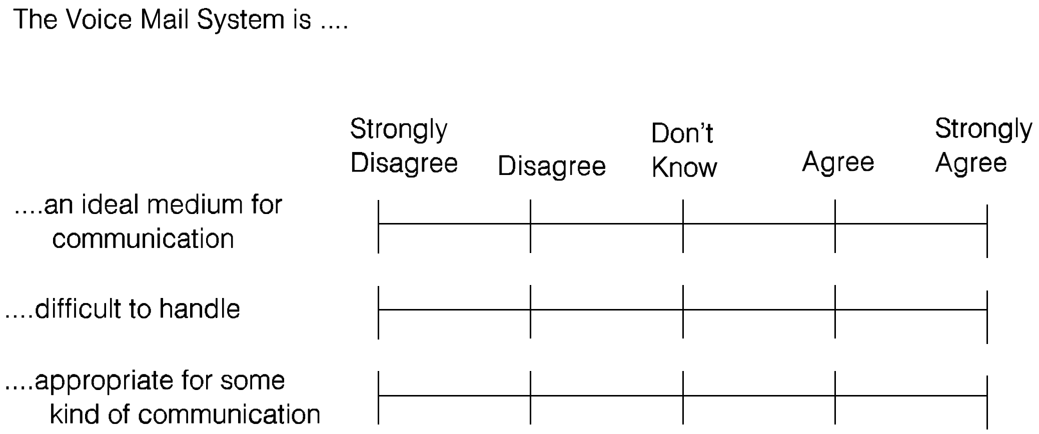

The Voice Mail System is ....
Strongly Disagree
Disagree
Don’t Know
Agree
Strongly Agree
....an ideal medium for communication
....difficult to handle
....appropriate for some kind of communication

Figure 4f): LIKERT scale (ordinal or interval) for attitude measurement e.g. towards a new voice mail system

8.7 Uses of rating scales

As described in Clause 4, usability evaluation involves two classes of measurement of equal importance, i.e. measures of performance and of attitudes.

The attitudes of the user of a telecommunication system depend on a number of factors, which in its turn depend on a larger number of underlying variables related to characteristics of the system, the human-machine interface, the task, the environment, etc. The attitude factors can generally be combined into a few dimensions, which may be extracted from experimental data by a formal factor analysis. These factors may then be used as rating dimensions in attitude measurements. The advantage of such a conceptual framework is that redundancy in the collection of data can be reduced and testing efficiency can be improved.

Examples of generalised attitude factors are given below in subclauses 8.7.1 and 8.7.2, where the underlying dimensions are described in the form of semantic differentials, which define a dimension in both the positive and negative direction. The differentials can often be used directly for defining rating scales. The examples are taken from studies on a number of specific telecommunications applications in two different research institutes.

8.7.1 "Social Presence" and "Aesthetic Appeal"

The "Social Presence" factor has been found to be of major importance for describing user attitudes in person-to-person communication (results from the Communications Studies Group, see Short et al 1976 [19]).

This factor is strongly related to scales marked by:

- unsociable - sociable;
- insensitive - sensitive;
- cold - warm;
- impersonal - Personal.

Another factor which is closely related with "Social Presence" and sometimes confused with it is "Aesthetic Appeal". This factor is particularly related to:

- colourless-colourful;
- small-large;
- constricted-spacious;
- boring - interesting;
- ugly - beautiful.

Both factors are clearly dependent on the communication media and are, therefore, suitable for the rating of telecommunication systems, as well as for rating face-to-face meetings as reference conditions.

Some examples of questions that are relevant to the "Social Presence" factor are quoted below (factor loadings are given in parentheses):

One does not get a good enough idea of how people at the other end are reacting (-0.76);

One gets no real impression of personal contact with the people at the other end of the link (-0.65);

One can easily assess the other people's reactions to what has been said (0.64);

It provides a great sense of realism (0.60);

One gets a good "feel" for people at the other end (0.60);

It isn't at all like holding a face-to-face meeting (-0.59);

It was just as though we were all in the same room (0.54);

People at the other end do not seem "real" (-0.50);

I would be happy to use the system for a meeting in which I intend to persuade other people (0.46);

I couldn't get to know people very well if I only met them over this system (-0.44).

8.7.2 "Satisfaction"

User satisfaction has a very general meaning and can be applied to usability evaluation of most systems. However, experience from tests on some telecommunication services, namely on multimedia information systems, have shown that the following three factors can be used for describing satisfaction in a communication situation:

a) effort versus ease of use;
b) agitation versus relaxation;
c) boredom versus enjoyment.

The meaning of these factors are defined by pairs of opposite adjectives such as:

a) hard - easy;
complicated - simple;
absorbing - restful.

b) stressed - relaxed;
restless - calm.

c) monotonous - entertaining;
dry - animating;
boring - enjoyable.

These dimensions result from factor analytical studies done within a research project of the Heinrich-Hertz-Institute, Berlin: "Relevance of Motion Picture Presentations for the Acquisition of Information" (1989-1992) [20].

Below are examples of statements that are relevant to the three factors cited above. Factor loadings, based on 120 subjects, are given in parentheses. Similar loadings were found in two additional factor analysis studies giving an overall subject base of 300:

a) Effort versus Ease of Use

The use of the service was:

-	troublesome	(0,89);
-	complicated	(0,82);
-	effortless	(- 0,52);
-	straining	(0,74).

b) Agitation versus Relaxation

While using the service I felt:

-	relaxed	(- 0,88);
-	rested	(- 0,74);
-	easy	(- 0,87).

c) Boredom versus Enjoyment

Using the service was:

-	monotonous	(0,82);
-	entertaining	(- 0,80);
-	dry	(0,78);
-	animating	(- 0,73);
-	boring	(0,86);
-	enjoyable	(- 0,74).

8.8 Reliability and validity

Systematic errors of physical measurements are, as far as possible, eliminated by the use of recognised measuring methods, by a careful calibration of instruments, and by instructions about the correct procedure. Often a measuring method can be checked by measuring a known quantity. When systematic errors cannot be eliminated from the measuring procedure a correction can be made instead, provided that the actual size of any such constant error is known. The remaining random error is a measure of the precision or repeatability of the objective method concerned.

For subjective measurements the situation is different. Since the purpose is to measure a sensation, there is no possible way of knowing the "true" value. Still, systematic errors can be compensated for by the use of well-established methods of experimental psychology and by a careful control of irrelevant factors that may affect the results. At best, one then obtains for a certain subject a distribution of observations where the dispersion consists only of casual fluctuations, the intra-individual variation, around the individual's mean value. The variation need not be large for a trained observer and is finally given by his sensory sensitivity.

In psychometrics, it is usual to distinguish between reliability and validity. Reliability is concerned with how consistent a measure is (i.e. do you get the same result when repeatedly measuring the same object). Validity is concerned with whether the measuring instrument is measuring what it is intended to measure (e.g. a barometer is a reliable and valid measure of air pressure but is a less valid measure of height above sea level).

Expressed in error terms this means that reliability is depending on the random error "*e*" only, while validity is primarily depending on the systematic error "*S*", but second also on the size of the random error (the standard deviation of "*e*"). To obtain a high validity, therefore, both the systematic error and the random error need to be small.

When the random error "*e*" is large, validity cannot be considered to be more than fair even if "*S*" is small. However, increasing the number "*N*" of items in a sample improves the reliability and therefore also the validity, in case "*S*" is small. The two concepts are independent, so it is possible for measures to be valid but not reliable or reliable but not valid.

The meaningfulness of subjective tests can be cross-checked by the application of different test methods, including different rating scales and instructions. This "validation" procedure is particularly important when new experimental methods are developed and applied in new areas.

Furthermore, it is necessary to take account of the inter-individual variation, which is caused by the fact that different subjects' sensation of the same stimulus may be consistently different from each other. A laboratory usability evaluation with a small group of users may therefore involve a sampling error, if the group is not representative of the relevant population. Repeated tests with different groups of subjects will then be required in order to increase the general validity.

Annex A (informative): Example test plan

A.1 General information

1. System ID....*Home Management System (HMS)*.......2. Plan Ref....*D1004*..............

3. Brief description of system......*A system that allows remote access via PSTN to basic home systems (e.g. TV, VCR, washing machines, lights etc)*..

4. Critical Usability Issues to be addressed
....*Consistency with other office/home equipment; easy to learn; flexible*..............................
....
....

A.2 Testing details

5. Subject Types, numbers required and critical characteristics (where appropriate)

Type.*Information Intensive* ...Characteristics....*Computer literate/technology aware*..........No.20..
Type.*Children* ....Characteristics....*Over 10, both sexes* ....No.20..
Type.....................................Characteristics...No......
Type.....................................Characteristics...No......
Type.....................................Characteristics...No......
Type.....................................Characteristics...No......
Type.....................................Characteristics...No......
Type.....................................Characteristics...No......

6. Key user tasks (where appropriate)

Task 1.......*Remotely programming VCR*...
Task 2.......*Opening/closing curtains* ....
Task 3............................. ....
Task 4............................. ....
Task 5............................. ....
Task 6............................. ....
Task 7............................. ....
Task 8............................. ....
Task 9............................. ....
Task 10........................... ....

7. Usage Context (where appropriate)

Accessing HMS from work or in transit from home to work. Must be compatible with work place technology, portable and compact...
....
....
....
....
....
....
....

8. Schedule details

Type of Test	Development Phase	Methodological Requirements	Proposed Test Date	Principal Evaluator
Expert Inspection	*Concept Phase*	*Specification documentation*	*1/11/93*	*D Smith*

A.3 Resource details

9. Human Resources

Project Manager.......*Dave Smith*.......................................
Principle Evaluator....*Tom Jones*.......................................
Team members.........*Phil Levy*..
.........*Gutma Singh*...................................
....
....

10. Cost Estimates

Manpower estimate................*50*............................... (days)
Non-manpower (NMP) estimate........*£30*........................(total cost))
NMP Details ....*4 U-matic Video tapes*.... (cost)
....(cost)
....(cost)
....(cost)
....(cost)
....(cost)

A.4 History

Author	**Issue**	**Date**	**Status**
Dave Smith	*Draft A*	*1/4/92*	*Draft for discussion*
Dave Smith	*Draft B*	*1/5/92*	*Draft for Quality Assurance*

Annex B (informative): Example specification table

Evaluation criteria	Context			Performance criteria				Measurement method		
Usability Goal	who	doing what	under what circumstances	worst case	planned level	best case	current level	metric	data collection method	technique
1.Make Remote interface to HMS easy learn	*Information intensive individual*	*opening/ closing curtains*	*for the first time from the office*	*45 secs*	*45 secs*	*20 secs*	*N/A*	*time to open/close curtains*	*observation*	*Check-list & timing device*
1Make Remote interface to HMS easy learn	*Information intensive individual*	*turning on/off any of the lights*	*for the first time from the office*	*45 secs*	*45 secs*	*20 secs*	*N/A*	*time to turn on/off any of the lights*	*observation*	*Check-list & timing device*
2. Make Remote interface to HMS easy to use	*Child over 10*	*Programming VCR to record*	*At home*	*90 secs*	*60 secs*	*30 secs*	*90 secs*	*time to successfully complete programming number of errors*	*observation*	*Video analysis*

Produced by..Issue No...................Date Issued...................Status....................................

Annex C (informative): Example test report

C.1 General information

1. System ID.....*HMS*........................ 2. Type of Evaluation...*Expert*..........

3. Principle Evaluator....*Tom Jones*.......... 4. Date of Test....*1/6/92*................

C.2 Major results

USABILITY GOAL	MAJOR FINDING	PRIORITY
Flexibility - programming sequences	*Children unable to use it due to complex programming sequences esp for TV and VCR operations*	*High*
Flexibility - keying	*People over 60 making keying errors due to inadequate key size and spacing distances*	*High*
Easy of learning - penetration	*Majority of managers only learnt to use 50% the functionality within specified learning period*	*High*
Ease of use - user manual	*Most found the manual incomplete and poorly structured*	*High*

C.3 Recommendations

USABILITY GOAL	RECOMMENDATION	PRIORITY
Flexibility - programming sequences	*Use function keys for TV and VCR operations*	*High*
Flexibility - keying	*Increase spacing in between keys in line with recommendation from ISO......*	*High*
Easy of learning - penetration	*Remove or make optional facilities for remotely controlling washing machine & cooker*	*Medium*
Ease of use - user manual	*Rewrite manual making it more task oriented, and provide a "quick reference guide"*	*High*

C.4 History

Author	Issue	Date	Status
Tom Jones	*1.0*	*1/8/92*	*Complete*

History

Document history	
September 1993	First Edition
February 1996	Converted into Adobe Acrobat Portable Document Format (PDF)