INFORMATION SOCIETY TECHNOLOGIES (IST) PROGRAMME

AIDE
IST-1-507674-IP

Review of existing Tools and Methods

Deliverable No. (use the number indicated on technical annex)		**D2.1.1**	
SubProject No.	**SP2**	SubProject Title	**Evaluation and assessment methodology**
Workpackage No.	**WP2.1**	Workpackage Title	**Generic evaluation methodology**
Activity No.	**A2.1.1**	Activity Title	**Review of existing Tools and Methods**
Authors (per company, if more than one company provide it together)		**Claudio Cherri (CRF), Elisabetta Nodari (CRF), Antonella Toffetti (CRF)**	
Status (F: final; D: draft; RD: revised draft):		**F**	
File Name:		**2.1.1 AIDE D2_1_1_v3 Final.doc**	
Project start date and duration		**01 March 2004, 48 Months**	

List of Abbreviations and Glossary

ACC	Adaptive Cruise Control
Acceptability	The characteristic of a product to be perceived positive by an user under the dimensions of Cost, Support, Reliability, Compatibility and Usefulness
ADAS	Acronym for Advanced Driver Assistance System
AICC	(Autonomous Intelligent Cruise Control) relieves the driver from the use of brakes and accelerator in certain conditions
AIDE	Adaptive Integrated Driver-vehicle Interface
Assessment	Process of determining the performance and/or impacts of a candidate application, usually in comparison with a reference case (existing situation or alternative applications), and usually including an experimental process based on real-life or other trials, often involving users
BS	Blind Spot
CA	Collision Avoidance
CAN	Controller Area Network (bus)
CAVE	Cave Automatic Virtual Environment
CWA	Collision Warning and Avoidance
DALI	Driving Activity Load Index
Diagnosticity	The Ability to differentiate the dimensions of the Workload. Only multidimensional scales have this property
DQS	Driving Quality Questionnaire
DWE	Driver Workload Estimator
ECU	Electronic Control Unit
FCW	(Frontal Collision Warning) is a system able to warn the drivers if the host-vehicle is approaching too fast an ahead obstacle. The warning information given to the user can be visual, acoustical or both.
GOMS	Goals, Operators, Methods, and Selection Rules.
GPS	Global Positioning System
GUI	Graphics User Interface
HMI	Human Machine Interface
HUD	Head Up Display
ICC	(Intelligent Cruise Control): it is a system that enables the vehicle to adjust its speed according to leading cars.
Intrusion Degree	In measuring WL, the measure does not influence or interfere with the task performance
IM	Information Manager
IP	Integrated Project
ISA	Intelligent Speed Adaptation
ISO	ISO Standard
ISO/ CD	ISO Committee Draft
ISO/ TR	ISO Technical Report
ISO/ TS	ISO Technical Specification
ISO/ WD	ISO Working Draft
IVIS	In-Vehicle Information & Communication System [15 p8]
LCD	Liquid Crystal Display

LCS	(Lane Change Support) It Informs the driver in case of other incoming vehicles coming along the lateral lanes; moreover, it warns the user if an overtaking manoeuvre is started when other vehicles – not seen by the driver – are approaching
LDWS, LW	(Lane Departure Warning System): It is a system that supports the driver in his / her lateral driving task and in particular when an unintentional lane change manoeuvre occurs. The warnings to the users are provided using different HMI channels, such as visual, tactile, and so on; [14 p7]
LKAS	Lane-Keeping Assist System
LP	Lateral Position
LSS	Lateral Support System
MCH	Modified Cooper-Harper Scale
MP	Mobile Phone
NASA	National Aeronautics and Space Administration
NASA-TLX	NASA Task Load indeX
NS	Navigation System
NV	Night Vision
OW	Overall Workload
PC	Personal Computer
PDT	(Peripheral Detection Task**)** Method whose purpose is to measure driver's mental workload and visual demands by means of a visual stimulus presented at the periphery of the ocular field; the user is asked to press a button in response to the stimulus
Performance	It refers to the quantitative measurement of one or more variables with respect to changes in *speed* (reaction time) and *accuracy* (hit rate).
Reliability	The reproducibility of measurements over time; it refers to the constency of the measure on different occasions or with different sets of equivalent tasks. [13]
RDS	Radio Data System
RSME	Rating Scale Mental Effort
SA	Situation Awareness
SAGAT	Situation Awareness General Assessment Technique
SART	Situation Awareness Reporting Technique
SD	Standard Deviation
Selectivity	The measure has to be sensitive only to the WL dimensions
Sensitivity	The ability to measure small changes [ISO WG8 N266]; The ability of the measure to reflect variations in the task difficulty and so, WL changes (a change in the task difficulty has consequences in the measure value)
SMS	Short Messaging System
SP	Sub-project
Subjective WL	The amount of increasing mental resources to use by an user when interacting with a system according to his own perception of it
Suitability for road trials	The possibility to gather metrics in real conditions
SWAT	Subjective Workload Assessment Technique
SWORD	Subjective WORkload Dominance
TICS	Transport Information Systems

TIS	Traffic Information System
Transferability	The ability of a measure to be used in other applications
TTC	Time To Collision
UMTS	Universal Mobile Telecommunication System
Usability	The extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use. (ISO/IEC 9241-11: Guidance on Usability (1998) [29])
Utility	The functionality of a device to do what it is built for
Validation	Process of verifying that an application performs as expected, often based on assessment results. In this sense, validation is usually considered as an extension to the assessment process, and sometimes the generic term assessment will be used to encapsulate validation.
Validity	The extent to which the variable is diagnostic for concept being investigated. A measure is valid if it measures what it intends to measure. A reliable measure is not necessarily valid.
WL	Workload: the specification of the amount of information processing capacity that is used for Task Performance
VE	Vision Enhancement
VODIS	Voice Operated Driving Information System
WP	Workpackage

Table of Contents

List of Figures

List of Tables

Executive summary

The present deliverable represents the first phase of AIDE Subproject 2. The principal goal of the Subproject is “to develop a cost efficient and industrially applicable methodology for quantifying behavioural effects of IVIS and ADAS functions, and their relation to road safety” [1].
So, in this first phase of work, the state of the art of the methodologies used to evaluate IVIS and ADAS applications has been collected. In particular, this research has focused mainly on techniques used to estimate user mental model, usability and acceptability. Other important factors as Workload and Situational Awareness have been only briefly illustrated here. A deep analysis about techniques used to investigate these dimensions will be given in Deliverable 2.2.1, which purpose is to review all workload methods. For this reason this Deliverable and D2.2.1 are strictly linked each other and they are the starting point to define a common framework of tools and methods in order to conduct behavioural and usability evaluations of IVI / ADA systems.

A general overview of the deliverable organization is described in the first chapter named “Deliverable objective and structure”.
The second chapter gives a general classification of the different methods, using the definitions of usability engineering field. User-Centred Design approach is the main point of reference of all work.
Then, in chapters 3, 4, 5, and 6 detailed descriptions are furnished. Each chapter focuses on a specific factor (Users Mental Model, Usability, Acceptability, etc) giving a theoretical description of the techniques mostly used to evaluate these factors.
In the Appendices some practical applications are described.
At the end, on the basis of the literature review, we have derived to:

- a list of dimension to take into consideration during a behavioural test. The dimensions are grouped on the basis of different modalities of interactions;
- a Step-By-Step procedure for Test Design.

1. Introduction

1.1. Deliverable objectives and structure

The aim of the present deliverable is to collect methodologies for the evaluation of usability and acceptability of IVIS/ADAS in-vehicle applications during different steps of system development.

Fist of all, Chapter 2 will present briefly the design cycle based on the User-Centred Design, the starting point to develop Usable systems, prone to be positively accepted by final users. Then it will present a general overview of the three main categories in which methodologies can be divided:

- Inspection Methods, mainly used as a guiding tool to check system requirements easily and effectively;
- Inquiry Methods, whose result is a clear and often subjective picture of user's expectances and performances with the system, and their main tool is the Questionnaire;
- Testing Methods, where systems are tested in an objective way, to identify performance issues and Safety problems when interacting with IVIS / ADAS

In the following chapters, specific dimensions, important to the Design Cycle, will be explored (User Mental Model, Usability, Acceptance, Subjective Workload, Situation Awareness), describing relative methods and techniques, beyond experimental description and main results.

Chapter 3 is centred on User Modelling and general techniques useful to begin the Development process, to identify user's requirements as well as query users for precise information.

Chapter 4 introduces Usability concepts, and examines several techniques to assess it, differentiating them in Methods involving Experts and Methods involving Users. Usability issues must be identified at the earliest possible stage of development, in order to adjust critical aspects of interaction in time, saving time on wrong development and increasing the ability of users to interact with the System naturally. The second part describes Acceptance, and the techniques used to assess it. Acceptance is a very subjective aspect of a System, and as such it can rely exclusively on Subjective methods. What an user thinks of a System can be understood only by "*asking*" him/her about it.

Chapter 5 deals with Workload, especially Subjective Workload. When Systems become too complex to interact with, Safety might be compromised. For this reason, it is important to understand how much users are loaded when interacting with IVIS / ADAS. Several techniques to assess Workload are described in detail, along with experiments and theoretical considerations

Chapter 6 describes Situation Awareness (SA), its importance in assessing IVIS / ADAS and some techniques to assess it. SA is either subjectively and objectively measurable, and such measures must be interpreted as a whole, in order to capture correct boundaries of Subject' SA.

Chapter 7 includes the Conclusions. The purpose of this Chapter is to summarize the methods described in the deliverable, and to indicate where to use a particular technique, and where to use another one. This task is quite complex, as many techniques might need to be used at once, whereas: the more a technique is used, the deeper the comprehension and understandability of it.

Every methodology listed in Chapters 3, 4, 5, and 6 will be described as follows:

- General description of the technique, where salient points are described. Where available, an algorithmic form is described;
- Identification of Pros and Cons, where interesting aspects of the methodologies are outlined, as well as aspects which require attention when deciding whether to apply such methodologies to project studies;
- A number of Examples, taken from existing European Projects and Literature. Such examples have the aim to show how the methodology has been applied and which results were obtained. Examples, in this case, describe only results. Experiments are often described more in detail in the appendix. The detail of Examples relies much on the source from where it comes from: this led to scarcely described Examples in some cases, and quite complete and technical ones others.

A number of appendices can be found at the end of the present Deliverable. In appendix 9.1, several grids which describe experiments are presented. Each grid relies on a common structure, so that salient and synthetic aspects of the experiment are taken.
In appendix 9.2, the structure of experiments is made concrete as a Step-by-Step Procedure to assess IVIS / ADAS. Such procedure contains all aspects found in the experiments taken into exam, as well as several aspects found in the existing literature, which complete the description of an experiment.
Appendix 9.3 describes a list of important aspects when assessing IVIS / ADAS, making divisions among modalities of interaction (i.e. Visual, Auditive, Haptic and so forth).
In appendix 9.4, two tables have the purpose to identify the different techniques across the various projects, to see how much a technique has been used during the development of IVIS / ADAS to assess them, until now.

2. General Overview of Methodologies

2.1. The User-Centred Design

Due to the increasing diffusion of electronic devices and Information Systems, in all fields of our lives, the need to have Systems which may be used by everyone is more and more felt everywhere. Not just professionals and experts are asked to interact every day with several devices, but each person who just enters a car needs to use an increasing number of facilities, to improve their guiding experience through infotainment devices (radios, mp3…) as well as guiding performance (ADAS). It is therefore important that such devices are intuitive, very easy to use, self-explanatory, and not intrusive with the driving task.
These are the main reasons why the design of systems has shifted towards an User perspective. Users are in the centre of the design process, and they are involved in any stage of the development. The purpose is to understand their needs, expectations, interaction modes, problems, to find effective ways to remove interaction difficulties and give them the information they need.
As quoted in ISO 13407: 1999 [28], “making system more human-centred has substantial economic and social benefits. System can contribute to protect users from risks for their health and safety, meeting users and organizational needs better”.

In this manner systems:

- are easier to understand and use, thus reducing training and support cost,
- improve user satisfaction and reduce discomfort and stress,
- improve the productivity of users and the operational efficiency of organizations, and
- improve product quality, appeal to the users and can provide competitive advantages.

The complete benefits of human-centred design can be determined by taking into account the total life-cycle costs of the system, including conception, design, implementation, support, use and maintenance.
Four main steps must be considered in this type of approach:

1. **Involving actively users in order to clearly understand users and task requirements**.
   In this way it is possible to obtain a valuable source of knowledge about the context of use, the tasks, and how users are likely to work with the future product or system;

2. **Properly Allocate functions between users and technology;**
   That is the specification of which functions should be carried out by users and which by technology. These design decisions determine the extent to which a given job, task, function or responsibility is to be automated or assigned to human performance.

3. **The iteration of design solutions;**
   Users are asked what they think about a System that should do something and how it should do that; a simple prototype (either on paper, or real) is prepared and users test it at an early phase of the development. Interaction issues are identified, and re-design process begins. When problems are solved, the next development phase begins. This type of behaviour, much different from the Cascade Development (where the product was designed and implemented, then tested), allows to proceed by increasing degrees of complexity, only when early stages are certified free from critical errors.

4. **Multi-disciplinary design;**
   A variety of skills are necessary for a user-centred design process to address all the human aspects involved in the design. It is possible to mention design engineers, production engineers, industrial designers, computer specialists, industrial physicians, health and safety practitioners, specialists in human resources and, obviously, final users.

Following this procedure, outcoming systems present interesting properties: they are usable, because they do not contain canonical errors, removed in early stages of development, and because the mechanism of interaction is decided with users who are likely to use the system. Moreover systems are likely to be accepted by users, because they are developed over user's needs, desires, expectations. A summary of concepts and hints for a good design for usable systems, tied to understanding user's behavior, can be found in [5].

To achieve such result, several methods can be used. The main thing to keep in mind is the user, how he/she thinks and behaves, what he expects the system will offer him/her. In the remainder of this Chapter, different methods of highlighting such information by users will be presented. Each method contains different methodologies, with different purposes and validity.

2.2. Inspection Methods

The aim of *Inspection Methods* is to enable examiners to inspect specific Usability-related aspects of a System. Inspection can be used to assess a generic system, anytime since a specific System Design become available up to pre-release. This is particularly useful in order to identify potential problems at an early stage of the development, and not overload developers with wrong or inconsistent elements. Inspection Methods are sensitive to inspectors, that is to say they may vary their results according to the examiner. In general, the defining characteristic of Usability Inspection is the reliance on judgement as a source of evaluative feedback on specific elements of a system. Inspection is an informal method of Evaluation, because it often does not rely on a formal model, and it is completely administered by experimenters through a manual way. Experience is therefore the main variable to obtain sensible results: the more the experimenters are skilled, the better results can be obtained.
Inspectors can be constrained to follow pre-defined Scenarios in order to test specific aspects of a System, or they may be free to interact with the System at their will. The first approach allows to inspect an aspect that is crucial for the tasks the system is developed for, while the latter enables the examiners to test aspects on their own choice, therefore potentially widening the number of peculiarities a system performs.
Such scenarios, especially when interacting with a particularly complex system, can be provided by observing user's interaction with the system or a similar one, in order to understand which tasks are to be executed and how they are supposed to be executed best.
The result of an Inspection Method is usually a list of Usability Problems, which are the starting point of a re-engineerization phase. Nielsen [54] examines with great detail this kind of methods.

PROs	CONs
• System verification at an early stage; can save time on further development • Implementation not explicitly required • More experimenters can improve results in a sensible way	• Relies on experimenter's expertise and knowledge, not on a formal model

Table 2.1: PROs and CONs of Inspection Methods

Several Methods can be identified:

- Cognitive Walkthroughs
- Feature Inspection
- Pluralistic Walkthrough
- Perspective-Based Inspection
- Heuristic Evaluation
- Checklists
- Guidelines

We will discuss methods listed in the right column more in detail in § 4.1.1.

2.3. Inquiry Methods

Inquiry Methods rely on users to collect data about the System. They can involve a wide range of people, providing sensible data to examine. They are a subjective source of data, that is they might be influenced by personal likes, needs, background. Because of this, people who participate in Inquiry Methods must be carefully chosen in order to reflect an accurate statistical sample. This way, it is possible to give consistency and validity to observed data. Some major drawbacks are the following:

- user's lack of memory, in some cases; when assessing topics that are easy to find in the user's mind, questions about that aspects are to be done during a test, because they cannot wait until the end;
- high reliance on subjective understanding of concepts described in the Inquiry Method used;
- potential misleading of data according to previous experience of the user.

Inquiry Methods, on the other hand, provide an effective way to know and understand user's thoughts, reasoning models, preferences. They are an useful tool to capture expectancies before developing a system, to understand what must be implemented and what can be left out. At last, they can give a precise picture of user' satisfaction in using the system, along with the will to use and purchase the system under inquiry.

Several methods are available:

- Questionnaires : an *ad hoc* or standardised Questionnaire is given to users to be filled in. Questions, which can accept multiple-choice answers, scales or free answers, have the aim to probe specific aspects of a System. These are the most used method when assessing a System, because they allow to assess specific Usability, Acceptance and Subjective Workload topics [45];
- Interviews: The experimenter asks the user to give free considerations, even vocally, about the system. Easily administrable before the test or after the test;
- Simple Observation of user's behaviour: completely non-intrusive, allows the experimenter to capture specific behaviour when interacting with the system, especially frustration that might not result when using questionnaires [83].

A summary of some projects using the aforementioned methods can be found in § 9.4.

The purpose of Inquiry Methods is rather different, according to the phase of the experiment:

- *Before beginning development*: in this case, users are asked the needed information of which aspects should be inserted into the System, what is the best way to execute a task, which information is superfluous for specific Systems and so on. Interviews are the most suited means for this case;
- *During Development*: this can be done to assess a partial state of System, or a simplified vision of it. Users are presented with a Multiple Choice or Free Answer Questionnaire which

covers specific aspects of the System. Scales can be used in order to have a numerical index of dimensions (e.g. goodness, pleasantness, completeness…)

- *After Development*: Questionnaires and Surveys can extract information about the complete System, allowing the Testers to assess a general performance of System according to a chosen group of final users. Results can highlight Usability issues as well as System's Acceptance level.

PROs	CONs
• Provide an effective way to know and understand user's thoughts, reasoning, models, preferences • Capture expectances before system development • Give a precise picture of User Satisfaction, along with his Willingness to purchase the system	• Need an accurate choice of people to obtain correct data • User's lack of memory drives to the need of administering the method as soon as possible • Subjective Understanding of concepts might happen, along with potential misleading of data according to previous experiences of users

Table 2.2: PROs and CONs of Inquiry Methods

In the present Work, we discuss the following Inquiry Methods:

- Brooke Questionnaire
- Usability Scale
- Semantic Differential Technique
- Uni-Dimensional and Multi-dimensional Subjective Workload Standardised questionnaires
- Willingness to Pay
- Importance Ranking
- Focus Groups
- Driver Behaviour Questionnaire
- Weiner Fahrprobe

2.4. Testing Methods

Testing Methods work on a representative set of Users, whose task is the utilisation of a System, or prototype, in order to verify the compatibility between driving and interacting with the system through an HMI. In this context, it is important to understand how much the system impacts on driving performance, that is whether the driver can safely interact with the system while maintaining full control of the car, and providing no harm to him/herself, the car, pedestrians, other vehicles, etc. Users in this type of Methods are more difficult to find, due to the more time-demanding nature of the trial. Problems of organisation and statistical needs contribute to further increase difficulties in finding a good statistical sample of Subjects.

Two main aspects can be assessed:

- The Primary Task Performance, that is the driving performance according to increasing difficulty tasks (e.g. straight road, curves, darkness, rain…)
- The Secondary Task Performance, that is the performance when interacting with the system while driving (in terms, for example, of Number of Errors and Total Task Time)

The former method can be assessed by means of Objective Measurements: data is collected while the Subject is assessing the System (in one of the ways described below) regarding his/her driving behaviour and car status (e.g. speed, variations, steering angle and so on); the latter is assessed by means of Subjective Measurements, e.g. Questionnaires and Interviews.

Due to this fact, Inquiry Methods and Testing Methods results are strictly tied, and will be discussed together.

Which environment?

IVIS / ADAS can be tested in different environmental conditions, either in a simulator and inside traffic. According to specific System aspects and to the availability of a prototype, one environment is more suitable than another. The different environments are the following:

- **Laboratory Testing**: best suited to test critical and dangerous situations; they offer great savings of time for repetitive testing, do not endanger Subjects, and offer reliable data (if validated); Specific scenarios can be forced to test particular aspects of the Project; due to lack of danger, they often are questioned to be really reliable, because Subjects might not apply correctly for they do not feel the danger of driving in real road.
  The laboratory testing can be shared out in three main groups according to different technologies used and level of realism of the driving scenario.

  Entry-level systems
  - low level of fidelity with a PC monitor that reproduces simplified driving scenario and a minimal mock-up that reproduces driver seat and primary and/ or secondary controls;

Fig. 1: Example of entry-level system (low-level fidelity system).

  - driving school simulators, with one or more screens, simplified mockup (some have a small motion base);

Fig. 2: Example of driving school simulator.

Driving simulators

- fixed driving simulator (real vehicle cab positioned in front a big screen which reproduce real road situation);

Fig. 3: Example of fixed driving simulator.

- dynamic driving simulator (real vehicle cab with motion base and big screen that can be fixed or mounted on the motion base);

Fig. 4: Example of driving simulator with motion base

- advanced driving simulator (real vehicle cab with motion base, circular screen on board and longitudinal and lateral rails);

Fig. 5: Example of advanced driving simulator: the National Advanced Driving Simulator (NADS)[1].

[1] Developed by the National Highway Traffic Safety Administration. For more information see http://www-nrd.nhtsa.dot.gov/departments/nrd-12/NationalAdvancedDriverSimulator.html

Immersive Virtual Reality Driving Simulators

- HMD simulator (Head Mounted Display and mock-up of primary controls and driver seat; data glove for interaction with virtual objects of the car interiors. Some have also a motion base);

Fig. 6: Example of HMD simulator with data glove.

- CAVE© simulator (one to six screens, motion base inside the CAVE© with mock-up of primary controls and driver seat);

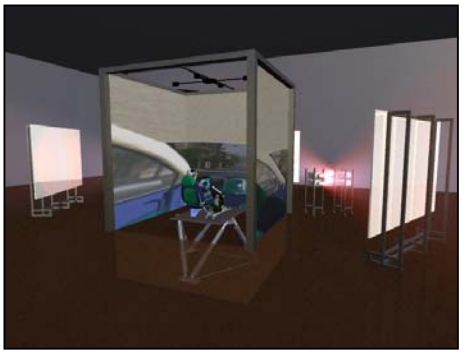

Fig. 7: Example of CAVE (with three screen).

This list is not exhaustive because it is always possible to have different combination of different solutions.

- **Test Track Testing**: high control of variables, it lacks anyway the risks of real traffic. The Subject drives a car into a track which can have different characteristics, so particular aspects of driving behaviour can be assessed (e.g. interacting with IVIS / ADAS and at the same time maintaining control of car);
- **Real Traffic Testing**: the most realistic type of testing. It allows perfect assessment of IVIS / ADAS, also relative to acceptability aspects. Impossible to test dangerous situations to their real extent.

Laboratory and Test Track Testing present a high degree of experimental control, giving the possibility to control conditions and variables (especially when using a simulator).
Real Traffic, on the other hand, gives a great degree of realism, where it is possible to obtain accurate data about critical interaction with ADAS.
In the following table, we summarize PROs and CONs of different Testing Environments [55]:

Environment	PROs	CONs
Laboratory	• Critical and dangerous situations can be tested • Repetitive testing is less time-consuming • No danger for Subjects • Reliable data is collected • Specific Scenarios can be forced to happen • The very *same* initial conditions can be repeated	• Due to lack of danger, Subject might perform in a different way than a "real" in-traffic performance • Realism might be missing, due to low resolution of devices (Screens, audio…)
Test Track	• High control of variables • Subjects are in a car, so IVIS / ADAS can be tested along with driving task	• Lack of danger of real traffic might result in low level of Task Demand
Real Traffic	• Highest level of realism • Accurate data can be gathered • Best way to test visual aspects of ADAS	• Dangerous situations cannot be tested (e.g. CW, ADAS malfunction and failure)

Table 2.3: PROs and CONs of different Testing Environments

According to people's knowledge of being monitored, different kind of information can be taken. Taking data from people who do not know that are being observed can be done, for example, following a car in the traffic and monitoring, from outside, the driver's behaviour [55, p5].

Data collecting according to people's knowledge of being monitored	
People Know	*People do not know*
The vehicle is equipped with suitable devices for logging data and devices such as ADAS Tests are conducted on real traffic situations, dedicated tracks or areas. +: extremely faithful results, knowledge of vehicle and instruments aboard -: few monitorings are available in few time	The vehicle is simply followed by another vehicle, that watches at the behaviour of the driver. Factors such as speed, lateral position, time gap and manoeuvres are monitored. Tests are conducted in real traffic scenarios. +: many observations possible in few time -: Vehicle information is unknown, save the details observable from outside

Table 2.4: Data Collecting according to people's knowledge of being monitored

2.5. Experimental and non-Experimental Research

All the three methods described in the previous paragraphs can be used alone or together in order to develop different types of research design. In particular, designing a research means to structure a study.
Research can have different levels of control (and associated validity): from highly controlled experiment (experimental design) until a observational, without control, research (non-experimental design).
Pedon [64] underlines important aspects of experimental design: when organising an experiment, it is important to identify *Variables*, that can represent several aspects of the experiment, of the subjects, auxiliary conditions and so on. For example, gender and age are variables of the Subjects; weather conditions are Variables of the experiment. Variables are divided into *Independent* and *Dependent*, according to their nature in the experiment:

- *Independent variables* are stimuli, events, facts that are likely to cause changes on other events and behaviours (e.g. age, number of buttons on steering wheel, number of items in a menu...);
- *Dependent Variables* are the events and behaviours that change according to Independent Variables (e.g. response time, errors…).

The weather condition is an Independent variable, because it can not be taken under control, and does not depend on other factors of the experiment; the number of mistakes during an experiment can be, for example, a dependent variable of the gender and age, when studying the influence of these two factors in user's Errors.
When an experiment is planned, a set of independent variables is chosen, and these shall be the fixed point in the experiment. The dependent variables will vary according to variations of independent ones, and they are the object of the testing.
When the number of independent variables used in the experiment is 1, the Design is simple; else, we talk about *Factorial Design*.
When all the variables can be manipulate by the researchers (that is the researcher, randomly, can assign his/her sample to different experimental conditions) and so all the aspects of the procedure can be controlled, we are developing an *experimental research*. On the other hand when it is used some independent variable which can't be controlled (i.e. gender or age) we are developing a *quasi-experimental research*.
In theory only with an experimental research it is possible to establish a cause and effect relationship among variables, practically, also the *quasi-experimental* researches are used to draw these type of conclusions.
Experiments, in general are driven by experts involving *Subjects*, where with this term we indicate a set of people with similar characteristics, carefully chosen to represent a statistical sample of population. The purpose to let Subjects use the system is to understand their behaviour during the interaction with it, their problems when interacting and their Acceptance of the system (how willing they will be to use it in the future).

There are, in literature, three basic types of Experimental Design, each one with certain underlying principles of ensuring equivalence between experimental conditions/groups of subjects and thereby controlling for systematic biases (e.g. "isolating the effect"). They are:

Between-subjects design	Each group uses the system under a single level of independent variable; this method requires more than one group, and each group interacting with a different level of the independent variable.
Within-subjects design	Each group uses the system under all experimental conditions, that is to say each group interacts with all the different levels of the independent variable
Mixed Design	Some variables are managed with *Within-Subjects* Design, and some others with *Between-Subjects* Design. (e.g. age, gender)

Table 2.5: Different approaches for Subjects in the Experimental Design - 1

Many criteria must be taken into account when deciding for an appropriate Experimental Design, like, for instance:

- Required sample size of subjects,
- Necessary Efforts for pre-tests,
- Precision of measures

In the following table, possible biases by carry-over effects are described for the different experimental design [65 p35]:

Possible biases by carry-over-effects	*Between-subjects*	*Within-subjects*	*Mixed*
Sample Size	-	+	+/-
Pretests Required	+	+	-
Precision	-	++	+
Carry-over-effects	+	-	+

+ advantage, - disadvantage.

Table 2.6: Different approaches for Subjects in the Experimental Design - 2

Within-subjects design should be preferred as long as carry-over-effects can be controlled by appropriate measures (e.g. counterbalancing). If these measures are not promising or quite difficult to employ (e.g. because transfer over different orders of conditions is asymmetrical), Between-subjects or Mixed designs have to be applied.

A laboratory testing, obviously, allow a higher control and so the possibility to conduct a real experimental research. However most of the usability researches are semi-experimental ones or, simply, observational.

In particular this second type of research is develop in all the situation in which users or usability expert observe and use a on-board device in order to highlight usability problems.

3. User Mental Model and Requirements

Qualitative analyses are based on User's witnesses and perspectives about a System. What is to be done is the construction of a *Mental Model*, that is a representation of User's way of thinking and behaving, and how people think objects should work, in order to forecast their next action and be therefore able to provide an effective support to them. In order to do this, it is mandatory to deeply understand what a user wants the system to do, and how he/she would like it to do that. A list of requirements are to be carried out, which will be the starting point for the design of the system. Different techniques to gather specific information from users can be used; some of them require different users to sit and talk to each other, others are individual ones. Below there is a brief description of some techniques broadly used in the automotive field.

3.1. Card Sorting

Card Sorting is a technique well suited to identify the way users group items, in order to understand how to group and position items in a way users are most likely to look for them, and to name the grouping done. In this way, system usability is improved because users take less time to find items they are looking for, and therefore they are less distracted from their main task (in an automotive context, driving).
An overview of the method can be found on http://www.usabilitynet.org/tools/cardsorting.htm.

Two methods of Card Sorting exist [48]:

- *Open Card Sorting*: Participants are given cards showing site content with no pre-established groupings. They are asked to sort cards into groups that they feel are appropriate and then describe each group. Open card sorting is useful as input to information structures in new or existing sites and products.
- *Closed Card Sorting*: Participants are given cards showing site content with an established initial set of primary groups. Participants are asked to place cards into these pre-established primary groups. Closed card sorting is useful when adding new content to an existing structure, or for gaining additional feedback after an open card sort.

Test Phases

1. The tester prepares a set of cards, each representing an item; cards may include comments or aids;
2. Users are presented with the cards, and are asked to arrange them in a meaningful way. Some cards will be blank, to let users add items themselves as well;
3. The experiment is usually conducted on one person at a time, to isolate influences from different subjects;
4. When users have finished, aggregation phase begins, and items categorisation is aggregated, in order to identify grouping, misleading meanings, different nomenclatures, potentially dangerous understandings.

Card Sorting requires at least 6 Subjects to give good results. The more the subjects, the better the results, even though when cards are more than 100, analysis time problems might arise.

PROs	CONs
• Easy and cheap to conduct • Helps understanding how people group items, and therefore identify structures for menus, icon positioning and so on • Identifies potentially misleading terminology • Identifies items difficult to categorise and find • Stabilised, as it has been used for a long period of time	• Could be difficult and time-consuming to administer and analyse when items begin to grow high in number • Does not rely on User Tasks; resulting grouping might be unusable • Subjects might ignore the underlying cognitive process tied to tasks and structure items by superficial meanings

Table 3.1: PROs and CONs of Card Sorting

Examples

- *Comprehension Testing of Active Safety Symbols* [10], (More in the appendices).

  Results:

  - Ranking of items representing how people understand them, and therefore can be safely included into car instrumentation;
  - Ranking of items leading to wrong judgement and possibly lower safety, and therefore to avoid in instrumentation

- *Iterative Design Of A New On-Board System For Public Vehicle: From The Idea To The Prototype,* CRF [67] (more in Appendices).

  Users can choose the favourite function from a list provided by the experimenters; the list includes functions concerning driving safety (i.e. Collision avoidance, aided vision, etc.), information / entertainment (i.e. radio, CD, etc.), comfort (i.e. Air Conditioning system, etc.) and travelling aids (i.e. navigation system, available park searching system, etc.).
  Due to the qualitative technique used (Focus Group) these results are collected considering the whole users' sample and not segmented by gender and age.

3.2. "Potato Head"

Potato Head (PH) is a technique that allows users to build their own instrumentation. It has been used broadly at UMTRI by Paul Green [23]. The customization is nearly absolute, at least theoretically, that is, users are able to choose the shape, dimension, color, label, position of any interaction device among a large number of available switches. In this way, users place controls where they prefer, thus raising Usability of resulting layout. PH owns its name from the homonymous toy, where children arrange heads by combining different kinds of eyes, mouths, noses, and so on. The principle here is the same.
The main issue for users who are asked to create the best layout in their opinion, is the very layout, because it is difficult to create an easy-to-use layout when not at all accustomed with it. For this reason, an even simple simulator must be used. Studies showed that the first user's choice was usually partially wrong, and during a driving simulation, users desired to change something in the designed layout: either the position, shape, label of switches.
A real car shell is required, in order to collect reliable data about positioning and distance from steering wheel and pedals, along with the ability for users to correctly and precisely position instrumentation where they wanted to.

The set of switches has to be as large as possible, in order to test different possibilities. This is a very expensive aspect, as the identification and preparation of switches requires much time and materials. However, few switches would not be representative of user's choice, thus leading to an approximate design of user's need.
User's freedom in placing switches must not be in contradiction with common Human Factors rules, i.e. place controls too close, or in objectively wrong positions.
Avoiding sharp edges in switches help maintain safety should accidents happen.
No original labels should be left on switches, as they might lead users to select them, despite the possibility to label them anyway they want. A similar argument can be stated for brands.
Good lighting should be provided when adjusting controls, but during simulation, night-time conditions should be reproduced, in order to let users find controls by remembering their position, rather than searching for them.
PH requires at least 50-100 users to give consistent results, and more than one experimenter should be used, in order to reduce test time. This leads to the need to make information gathering homogeneous. Videotaping testing sessions helps improve data collection quality.
Final Configuration should be visually recorded, either by videotape or picture, and labelled; users often are asked why they chose such configuration, but obtaining useful answers (i.e. different from "I like it", "It is simple" and so on) requires experimenter' skills in finding a good way to communicate with users and to ask for the right questions. To solve this problem, a list of possible general reasons was developed by experts and users together. Tested users had to choose from the list, thus eliminating biases in judgement.
Obtaining good data from experiments is difficult, as many combinations of switches can show up. Using a proprietary software helps diminish time of calculation, and to reduce the probability of mistakes during data recording. Such procedure might increase costs of testing.
Care must be taken in choosing participants, as they must cover all ages, social conditions, occupation, in order to capture a large number of drivers. Specific types of car (e.g. luxury, sports) might concentrate some user characteristics to salient aspects.
It is important, aside from preferences, to see user's performance with the designed system; timings and errors done should be taken into account when judging the final design.

PROs	CONs
• Extremely customizable (shape, position, label of each control) • Users can design their own dashboard, thus having good familiarization and potentially good performance	• High cost of preparation of switches • Time-consuming in executing assemble • Time-consuming in analyzing data • Designed layouts must be verified according to performance and error-making • Subjects must be representative of the type of car interior being designed, and might be difficult to find

Table 3.2: PROs and CONs of "Potato Head"

Examples
CRF is using an adapted version of Potato Head technique in the first design phase:

- PH technique integrated in Virtual Driving Simulator, thus making it possible to arrange car interiors in Virtual Reality. The designed layout is then tested in a simulated driving session, in order to see user's performance and interaction with the system.

- *Iterative Design Of A New On-Board System For Public Vehicle: From The Idea To The Prototype,* CRF [67] (more in Appendices).
  Users could choose the favourite function from a list provided by the experimenters. For each chosen function, participants had to indicate their favourite position inside the vehicle (i.e. dashboard, instrument cluster, etc.), sketched on a graphical schema, considering their stereotypes and perceived usability criteria.

Due to the qualitative technique used (Focus Group) these results were collected considering the whole users' sample and not segmented by gender and age.

3.3. Focus Groups

Focus Groups (FG) are a qualitative method of Research for inspecting a System, at any phase of its development, focused on specific topics and in which a group discussion takes place.
A Focus group is essentially a discussion made by perspective users of a system who are asked to talk (almost) freely about the system, its abilities, and its issues [51].
A professional Moderator is often present, with the purpose of guiding the discussion onto a pre-designed schema, and maintains it as transparent as possible. Users' opinions are tracked, so that it is possible to understand the needs of an user who is willing to use the system.
The Focus Group is useful to understand what users would like a specific System to do and how it should do that. To do this, specific themes are outlined, and explored during Focus Groups. Moderators have the task to keep the participants on track, while leaving them free to talk as they want.
Users are sometimes presented a Demo version of the system to assess, and are introduced to its goals and abilities; they can describe their interaction with the system, maybe relating to very time-consuming tasks; this way, it is possible to save time on watching users do a long procedure, simply by listening to how they would do that.
Focus Groups present the problem that users might be imprecise about their statements, that is there might be a difference between what users say and what and how they really do it.
For this reason, it is not recommended to drive a Focus Group as the only resource for assessing Usability of a System. Direct Observation of users should always be conducted.
Another issue is finding homogeneous users for talk, and for this problem, Focus groups should rely on carefully selected users, who share the same level of knowledge and background, which is suited for assessing the system. When interest of users in the System is quite high, then the discussion may be of much help to experimenters, because much information can be collected.

Six to nine users should be the right number to drive a good Focus Group.

PROs	CONs
• Cheap and fast way to assess a System, when carefully planned • Allows clear understanding of users needs and thoughts about the system	• Difficult to find homogeneous users • Needs a good moderator to work efficiently • Subjective information might be unreliable, and could need to test users onto System to have consistent feedback

Table 3.3: PROs and CONs of Focus Groups

Examples

- *Cars Iterative Design Of A New On-Board System For Public Vehicle: From The Idea To The Prototype* (more in the Appendices).

  Five FG have been driven, to gather opinions about CyberCars vehicles, with particular attention to the following aspects:
  - Comparing CyberCars vehicles to actual transport services;
  - Highlighting advantages, disadvantages and perceived usability;
  - Considering perceived safety, reservation modality and payment service;
  - Considering the possibilities to use shared vehicles (10-20 seats buses) vs. not shared vehicles (4 seats).

  Moreover, functions to be arranged in the on-board system have been analysed.

- *COMUNICAR: Establishment of Users needs* [45].
  Three Focus Groups (two in Italy and one in Sweden) were driven to elicit knowledge from experts of innovative on-board system like Comunicar. The aims were:
    - bring up issues from Product and Research experts about the concept of an integrated multimedia HMI for car drivers;
    - get a first sketch on car drivers' information needs from a Product and Research viewpoint;
    - get recommendations for user-oriented HMI design;
    - get hints on the design of the user tests;
    - get hints on the set of questions to be developed in the questionnaire.

3.4. In-depth Interviews

In-Depth Interviews are quite different from Surveys, because surveys follow a listing of pre-defined questions to be answered. They are a qualitative method, and the assessment is done one person at a time.
The purpose of Interviews is to let interviewees to freely talk about important topics, and the interviewer has the purpose to guide the interview, collect data and analyse it. Well-conducted interviews last long, and therefore are expensive. On the other hand, they provide individuals' perceptions, opinions, facts and forecasts, and their reactions to initial findings and potential solutions [24, 25].
Interviews are an underestimated technique of inquiry, because many researches are reluctant to accept the qualitative nature of Interviews.
Interviews follow a pre-defined structure, but interviewers should take this structure as hidden as possible, in order to let interviewees think conversation is going on naturally. At early stage of a System assessment, less structured interviews may serve to identify clearly system's requirements. The nature, number and content of questions is decided in accordance with the client and or the developer, to ensure all relevant topics will be covered.
Interviews should be conducted without distractions, in order to allow interviewees to focus on questions. It is a good practice to ask permission for recording conversations by tape, alongside written notes, when advisable to use a tape; sometimes, taping might inhibit interviews, and therefore it should not be used. It is up to the interviewer to adapt the use of taping.
Often, "probe questions" are provided, in order to avoid interviewees' bypass of questions with vague answers. These questions are structured in a way interviewees can not get around the question, and they can't answer with Yes or No as well.

Interviewers make the difference in the quality of an Interview. They must have a certain confidence with the topic they are interviewing about, and must be able to build an interpersonal situation, to gain people's confidence and to let them open about topics, and be a good listener as well, because interviewees must talk most of the time. There could be the need to train Interviewers before doing the Interviews.
The choice of the interviewees must be done carefully, in order to allow sensible data to be collected. People should be selected from the environment the product under assessment is to be used.
Interviews usually begin with general questions, and should let interviewees talk for most of the time (nearly 90%), in order to collect complete thoughts of people. Different questions inspecting the same topic might stimulate interviewees to add more information to that already provided.
When the Interview is over, collected data must be checked and organized at once; stacking interviews would lead to poor treating of information. From Interviews, important information should emerge, such as major points of agreement, substantial points of disagreement, conclusions or implications for the design of an intervention.

Interviews are expensive, in terms of time and money. A good interview lasts, all inclusive, at least four hours, and the cost for each interviewee might be very high, provided that interviewees are easily found.

PROs	CONs
• Allow to understand user's thoughts, requirements, desires, doubts in a very free way • Motivations and resistance towards certain markets, products, services or marketing measures can be determined • Provide structured information (when the interview is structured) that can be transformed in numbers	• Skilled interviewers are required to prepare effective interviews and to administer them • Expensive in time and money, (each interview requires 3 to 4 hours, data analysis inclusive)

Table 3.4: PROs and CONs of In-Depth Interviews

3.5. Questionnaires

Questionnaires are one of the most used tools to assess a System. Because of their generality of application, they are suitable to assess almost everything of a System, at any phase of its development. Moreover questionnaires can be used also to specify systems requirements in the first phase of the design, when the real object doesn't exist yet ("ideation" phase). In the present paragraphs the questionnaire technique will be describe in details, then some example of its application to define system requirements will be presented. In the chapter 4 and 5, questionnaire to evaluate specifically usability, acceptance and workload will be described.

Questionnaires are a list of questions, whose aim is probing different topics regarding the System. Answers can be

- Free, where the user has to write down his/her own thought;
- Single/Multiple-choice, where different possible answers are provided for an user to answer a question, either choosing one answer among a set of question, or choosing several answers, up to a maximum, among all choices;
- Marking a scale point, where users must rate a concept in a given scale.

Questionnaires are the same for each Subject who tests the System, therefore they can be easily administered to a great amount of Subjects. The same questionnaire can be re-used across different systems, if it is not too specific for a system. To this extent, several *Standardised Questionnaires* exist, whose purpose is to test general aspects, so that it can be applied to many systems.
Due to its high flexibility, the use of Questionnaires is not limited to System Usability, but there are examples of Standardised questionnaires for Acceptance [85] and for Workload.
Questionnaires are a subjective technique, because they rely on Subject's opinion of the system. They are opposed to objective measures about driving behaviour, and are complementary.

Questionnaires have some drawbacks:
1) Subjects may tend to give personal re-elaboration of the event, by introducing biases due to previous experience and knowledge; this may lead the test results to be misleading;
2) Questionnaires are to be done as soon as possible, if they deal with System particular topics, as Subjects tend to forget events;
3) Subjects are often difficult to find, and might have a technical background. This could bring to misleading results, due to the incorrect mirroring of a pure statistical sample.

Creating a good questionnaire requires high System knowledge, clear view of which topics are to be probed, correct expertise of Subjects' language. These points are mandatory, because only who knows the System can know which are the critical parts and how they behave, and therefore can make correct questions to subjects to obtain centered answers. The questions must be clearly asked to Subjects, and without any misunderstanding. If people misunderstand questions, their answers could bring to misleading results.

A Method to create a Questionnaire can be the following:

- Identification and definition of Content areas which will be inserted into the questionnaire;
- Definition of Single Areas contents and formulation or relative items;
- Item Arrangement in a specific grouping and order (general questions before; difficult and complex questions later);
- Analysis through pre-test (a little group of subjects similar to the big sample, to test the goodness of the questionnaire).

Types of Answering Format

Questionnaires can be prepared using different answering formats; beyond the single and multiple-choice, *Ranking Scales* and *Rating Scales* can be identified.
The first type presents the Subject with a list of items, and asks them to order them according to a ranking order (from best to worst, from less / more desirable to more / less desirable, and so forth). This technique allows high involvement by Subjects, but presents problems with long lists.
The Rating Scales ask Subjects to rate a concept on a predefined scale, from a maximum to a minimum. This technique allows more precise results and application of statistical methods to results. On the other hand, subjects could make less precise distinctions among the different proposed alternatives. This type of scale is used, for example, by Van Der Laan, with Semantic Differential Scale (see later in § 4.2.1), for MCH and by Brooke Likert Scale (see next paragraph).

PROs	CONs
• Suitable for testing the system at any phase of development • Provide a common way to assess all users according to pre-defined aspects • Allow to test several aspects of the system at any desired level • The same form of Questionnaire can be used to assess different Systems • Questionnaires used for Usability, Acceptance, Workload, and Situation Awareness due to their flexibility	• Subjective results are obtained • Lack of memory by users might give missing (or even) misleading answers when the questionnaire is administered much after the test • Difficulty in finding a consistent number of Subjects; they might not be "normal" users, (i.e. they might have technical background and knowledge of interaction modalities over the mean)

Table 3.5: PROs and CONs of Questionnaires

To define system requirements different type of approach can be followed. It is possible define a group of functions and ask directly final users if they feel interested or uninterested with the specific services/information that can be provided to the driver. Moreover it is possible to recruiting some participants and ask their level of interest in a specific functions after they have real used them in a vehicle or prototype. Another option is to use multimedia applications. A participant see some situations and different functions displayed in a monitor. Then their preferences about different solutions are asked. The visualisation of situations gives the subjects a vivid description of the relevant traffic scenario and helps them to classify the functions under investigation.

Further information about questionnaires can be found in [61 p81 ,69].

Examples

- *COMUNICAR: Establishment of Users needs* [45].
  A questionnaire with some 80 items was administered to about 60 drivers in Italy and 60 more in Sweden. Over than 50 different information and services were rated with respect to driver's interest.
- *IN-ARTE : User-needs survey* [9].
  A multinational survey by means of a PC-based questionnaire on CD-ROM took place in the countries of the partners in the consortium (France, Italy, Sweden and Germany). The survey have the aim to identify most supporting, relaxing and useful functions and operational modes in order to develop the IN-ARTE system. It is an integration of elementary functions of existing driver support systems (e.g. Adaptive Cruise Control (ACC), Heading Control, combination of longitudinal and transversal control).

3.6. Self-Reported Diaries

Self-Reported Diaries are a technique in which users are requested to keep a diary during a period of time, recording personal information about their behaviour, when interacting with a system. The purpose is to assess the impact of the system in everyday's life. Diaries are totally subjective, and therefore collected information must be verified in some other ways, due to possible biases and personal background.
After data has been collected, it must be aggregated in order to find out agree points and opinions, and topics to refine and adjust. Usually this technique is used in psychology researches (i.e. in the field of human errors).
It is applied also in usability studies to investigate the modality and frequency of use of different functions and devices. This type of information are used to define requirements of a system during the "requirements elicitation phase" [39].

PROs	CONs
Users may use the system for a long period of time; this enables them to test it thoroughly, so observations can be more accurate and significative	Observation done are highly subjective and cannot be checked by experts, unless the system provides a logging system and/or camera recordings

Table 3.6: PROs and Cons of Self-Reported Diaries

Examples

- *CRF* used this technique internally in order to identify the most used functions of on-board instrumentation.
- A Large Scale Trial With Intelligent Speed Adaptation In Lund [86]
  A number of test drivers, approximately 25 persons, will be asked to fill in a diary regarding their experience of the ISA. The purpose of the diary is to collect low probability events and other experiences which are difficult to trace in other studies. Examples of variables is social press and "pub-talk", experience of accidents and conflicts, mechanical problems etc. The subjects will be asked to fill in the diary during three periods of one week at each period. The first time shortly after their vehicle has been equipped with the ISA, secondly in the middle of the test-period and finally at the end of the test-period.

3.7. Task Analysis

A *Task* is an objective an user wants to achieve by using a specific system in a specific context, which can be an information about the current status of the system or a modification of the status.
According to this definition, the *Task Analysis* (TA) has the aim to identify the Tasks an user is willing to complete by using the System, and how he/she can do it efficiently. In particular, Task Analysis analyses what a user is required to do in terms of actions and cognitive processes to achieve a task. In fact, it is important to understand the user's behaviour when designing a system, for there is a substantial difference among designers of the system, its implementers, and real users. Things that are simple and immediate for the designer might sound obscure to users, and the programmer might realise an HMI that is suitable for another programmer, but that leads newcomers to panic when using it. The user-centered Design begins with analysing tasks from user's viewpoint. This way, it is relatively safe to implement decisions taken and to be sure a generic user will be able to use the HMI correctly after little (or no) training.
The task analysis must do his or her best to understand the user's task situation well enough to influence the system design given the limited time and resources available.
At a general level, the task analysis involves collecting and representing the answers to the following primary overall questions [40]:

1. What does the system do as a whole?

2. Where does the human operator fit into the system? What role will the operator play?

3. What specific tasks must the operator perform in order to play that role?

There are many techniques to perform a Task Analysis, some involving only experts and others involving final users.

In particular some techniques, called **action oriented approaches**, give a description of the observable aspects of operator behaviour at various levels of detail, together with some indications of the structure of the task. Other techniques focus on the mental processes which underlie observable behaviour, e.g. decision making and problem solving. These will be referred to as **cognitive approaches**.

Task analysis requires information about the user's situation and activities, but simply collecting data about the user's task is not necessarily a task analysis. In a task analysis the goal is to understand the properties of the user task that can be used to specify the design of a system; this requires synthesis and interpretation beyond the data. In the previous paragraphs some techniques to collect data have been already presented (see § 3.3, 3.4, 3.5).

Once the task data is collected, the problem for the analyst is to determine how to represent the task data, which requires a decision about what aspects of the task are important, and how much detail to represent. The key function of a representation is to make the task structure visible or apparent in some way that supports the analyst's understanding of the task. By examining a task representation, an analyst hopes to identify problems in the task flow, such as critical bottlenecks, inconsistencies in procedures, excessive workloads, and activities that could be better supported by the system. Traditionally, a graphical representation, such as a flowchart or diagram, has been preferred, but as the complexity of the system and the operating procedures increase, diagrammatic representations lose their advantage.
One general form of task analysis is often termed task decomposition. This is not a well-defined method at all, but merely reflects a philosophy that tasks usually have a complex structure, and a major problem for the analyst will be to decompose the whole task situation into sub-parts for further analysis, some of which will be critical to the system design, and others possibly less important. For example, one powerful approach is to consider how a task might be decomposed

into a hierarchy of subtasks and the procedures for executing them, leading to a popular form of analysis called (somewhat too broadly) Hierarchical Task Analysis. However, another approach would be to decompose the task situation into considerations of how the controls are labelled, how they are arranged, and how the displays are coded. This is also a task decomposition, and might also have a hierarchical structure, but the emphasis is on describing aspects of the displays in the task situation.
Obviously, depending on the specific system and its interface, some aspects of the user's task situation may be far more important to analyze than others. Developing an initial task decomposition can help identify what is involved overall in the user's task, and thus allow the analyst to choose what aspects of the task merit intensive analysis.

When large systems are being designed, an important component of task analysis is to consider **how the system**, consisting of all the machines, and all the humans, **is supposed to work as a whole** in order to accomplish the overall system goal. This kind of very high-level analysis can be done even with very large systems, such as military systems involving multiple machines and humans. The purpose of the analysis is to determine what role in the whole system the individual human operators will play. Various methods for whole system analysis have been in routine use for some time.
Mission and scenario analysis. Mission and scenario analysis is an approach to starting the system design from a description of what the system has to do (the mission), especially using specific concrete examples, or scenarios.
Function-flow diagrams. Function-flow diagrams are constructed to show the sequential or information-flow relationships of the functions performed in the system.
Function allocation. Function allocation is a set of fairly informal techniques for deciding which system functions should be performed by machines, and which by people. Usually mentioned in this context is the Fitts list that describes what kinds of activities can be best performed by humans versus machines. However, this classic technique is rarely used in real design problems since it is simply not specific enough to drive design decisions. Rather, functions are typically allocated in an ad-hoc manner, often simply maintaining whatever allocation was used in the predecessor system, or following the rule that whatever can be automated should be, even though it is known that automation often produces safety or vigilance problems for human operators.

Once the whole system and the roles of the individual users and operators has been characterized, the main focus of task analytic work is **identify more specific properties of the situation and activities of the human operator or user**. In particular it is important to represent what users needs to know, have to do and might do wrong with the system.

What users have to do

A major form of task analysis is describing the actions or activities carried out by the human operator while tasks are being executed. Such analyses have many uses; the description of how a task is currently conducted, or would be conducted with a proposed design, can be used for prescribing training, assisting in the identification of design problems in the interface, or as a basis for quantitative or simulation modelling to obtain predictions of system performance. Depending on the level of detail chosen for the analysis, the description might be very high-level, or might be fully detailed, describing the individual valve operations or keystrokes needed to carry out a task.
The following are the major methods for representing procedures:
Operational sequence diagrams. Operational sequence diagrams and related techniques show the sequence of the operations (actions) carried out by the user (or the machine) to perform a task, represented graphically as a flowchart using standardized symbols for the types of operations. Such diagrams are often partitioned, showing the user's actions on one side, and machine's on the other, to show the pattern of operation between the user and the machine.
Timeline analysis. Timeline analyses simply display activities, or some characteristic of them, as a function of time during task execution. For example, a workload profile for an airliner cockpit would show a large variety and intensity of activities during landing and takeoff, but not during cruising.

After constructing a timeline display, the analyst looks for workload peaks (such as the operator having to remember too many things), or conflicts, such as the operator having to use two widely separated controls at the same time.

Hierarchical task analysis. Hierarchical task analysis (HTA) involves describing a task as a hierarchy of tasks and subtasks, emphasizing the procedures that operators will carry out, using several specific forms of description. The term "hierarchical" is somewhat misleading, since many forms of task analysis produce hierarchical descriptions; a better term might be "Procedure hierarchy task analysis." The results of an HTA are typically represented either graphically, as a sort of annotated tree diagram of the task structure, or in a more compact tabular form. This is one of most widely-used forms of task analysis.

HTA descriptions involve goals, tasks, operations, and plans. A goal is a desired state of affairs (e.g. a chemical process proceeding at a certain rate). A task is a combination of a goal and a context (e.g. get a chemical process going at a certain rate given the initial conditions in the reactor). Operations are activities for attaining a goal (e.g. procedures for introducing reagents into the reactor, increasing the temperature, and so forth). Plans specify which operations should be applied under what conditions (e.g. which procedure to follow if the reactor is already hot).

Plans usually appear as annotations to the tree-structure diagram that explain which portions of the tree will be executed under what conditions. Each operation in turn might be decomposed into subtasks, leading to a hierarchical structure.

GOMS models. They are closely related to Hierarchical task analysis; GOMS models describe a task in terms of a hierarchy of goals and subgoals, methods which are sequence of operators (actions) that when executed will accomplish the goals, and selection rules that choose which method should be applied to accomplish a particular goal in a specific situation. However, both in theory and in practice, GOMS models are different from HTA. The concept of GOMS models grew out of research on human problem solving and cognitive skill, whereas HTA appears to have originated out of the pragmatic common-sense observation that tasks often involve subtasks, and eventually involve carrying out sequences of actions. Because of its more principled origins, GOMS models are more disciplined than HTA descriptions. The contrast is perhaps most clear in the difficulty HTA descriptions have in expressing the flow of control: the procedural structure of goals and subgoals must be deduced from the plans, which appear only as annotations to the sequence of operations. In contrast, GOMS models represent plans and operations in a uniform format using only methods and selection rules. An HTA plan would be represented as simply a higher-order method that carries out lower-level methods or actions in the appropriate sequence, along with a selection rule for when the higher-order method should be applied.

What users might do wrong

It has been developed a variety of techniques for analyzing situations in which errors have happened, or might happen. The goal is to determine whether human errors will have serious consequences, and to try to identify where they might occur and how likely they are to occur. The design of the system or the interface can then be modified to try to reduce the likelihood of human errors, or mitigate the consequences of them. Some key techniques can be summarized:

Event trees. In an event tree, the possible paths, or sequences of behaviours, through the task are shown as a tree diagram; each behavior outcome is represented either as success/failure, or a multi-way branch, e.g. for the type of diagnosis made by an operation in response to a system alarm display. An event tree can be used to determine the consequences of human errors, such as misunderstanding an alarm. Each path can be given a predicted probability of occurrence based on estimates of the reliability of human operators at performing each step in the sequence (these estimates are controversial; see [70] for discussion).

Failure modes and effects analysis. The analysis of human failure modes and their effects is modeled after a common hardware reliability assessment process. The analyst considers each step in a procedure, and attempts to list all the possible failures an operator might commit, such as to omit the action, perform it too early, too late, too forcefully, and so forth. The consequences of each such failure "mode" can then be worked out, and again a probability of failure could be predicted.

Fault trees. In a fault tree analysis, the analyst starts with a possible system failure, and then documents the logical combination of human and machine failures that could lead to it. The probability of the fault occurring can then be estimated, and possible ways to reduce the probability can be determined.

In conclusion, task analysis includes a lot of techniques (some have been described here) which allow researchers to analyse and interpret in a better way the general organization of a system.
It is critical to understand that these techniques do not in themselves analyze the task or produce an understanding of the task. Rather, they are ways to represent the results of task analysis. They have the important benefit of helping the analyst observe and think carefully about the user's actual activity, both specifying what kinds of task information are likely to be useful to analyze, and providing a heuristic test for whether the task has actually been understood.

Moreover each context have its rules and different important aspects to consider. An automotive context should take into high account the fact that the system will be used when driving. Tasks should be designed in order to represent this major difficulty, to count safety issues.
Once tasks are specified, they can be used to test goodness of the system; if users can complete tasks efficiently, the system is adequate to do what it is designed for. Anyway, many other factors are to be taken into account in automotive field, such as Safety Performance when fulfilling tasks. If a user hits pedestrians when interacting with the system, it does not mean the system is suitable, even if the task is completed correctly and on time.

3.8. Decision Trees

Decision Trees (DT) are a technique used to facilitate the decision making process. The decision problem is represented in the form of a tree with numerous branches and endpoints (sometimes referred to as *twigs*). The DT contains points of decisions where designers or engineers are required to decide the most likely outcome for a given situation. The tree also contains points of action, which represent the recommended course of action, given a series of decisions (e.g. to allocate a given in-vehicle task function to be controlled using speech input). Decision and Action points are connected one by one to form branches and endpoints that represent various phases of the decision-making process, leading to the final result of task-function allocation. Decision criteria such as safety, usability, and driver preference/acceptance provide a logical basis for determining whether a given task-function should be considered for control using an interaction modality instead of another (e.g. vocal vs. manual, as in DINGUS Project). This method can be used in the Design Method and in the Evaluation of a System as well. For this reason, examples of Decision Trees may be pertinence of User Mental Model Creation, Usability and Acceptance Evaluation, and Workload Measurements: the Modified Cooper-Harper Scale (MCH) is an example.

The Method can be described as follows:

1) A Tree is built according to a particular *Task Function*: a function which can be executed by an user when interacting with the System.
2) A series of Choice points are added, each of them questioning about a particular aspect of the system (which kind of user, car, scenario… the function is to be used in/by);
3) A series of Choices for each Choice Point, defined according to the Choice Point, is inserted. Typically, Decision Trees are Binary, in the sense that just two alternatives are inserted (e.g. Yes/No…) for each Choice Point. Each Choice will lead to another Choice Point, or to a Conclusion Point;
4) Conclusion Points end the current flow of the Tree, defining the most suitable tool to support the Task Function.

A number of Decision Trees equal to Task functions to be analysed is created.

PROs	CONs
• Schematic representation of decision process • Helps designers to ensure all possible decisions can be done correctly and adequate interaction systems are available to users	• Might be time-demanding to develop, and might require skilled designers

Table 3.7: PROs and CONs of Decision Trees

Examples

- *The Use Of Speech Recognition Technology In Automotive Applications* [17].
  A series of decision trees were developed to provide designers and engineers of in-vehicle tasks with an analytical methodology for determining whether a given task should be performed using manual controls or using speech input.

4. Usability And Acceptance

4.1. Usability

With the word Usability we mean “the extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use”.

1. *effectiveness*: accuracy and completeness with which users achieve specified goals;
2. *efficiency*: resources expended in relation to the accuracy and completeness with which users achieve goals;
3. *satisfaction*: freedom from discomfort, and positive attitudes towards the use of the product.

(ISO/IEC 9241-11: 1998 Guidance on Usability)

Evaluating an IVIS /ADAS according to its Usability therefore means to identify how much the system is user-friendly and easy-learning to be used by itself and how much it influences the driving experience, and how much it is appreciated and used in a satisfying way by drivers.
In order to verify these aspects, it is important to identify whatever issue the System under Evaluation can present, according to:

1) Aspects related to the interaction between System and driver. This encompasses, for example, issues like visibility, legibility, readability of the visual output, Auditive Messages and Speech Recognition, Multimodal Conflicts, interaction logic, mapping between input and output device…
2) Aspects related to the interaction between driver, system, and driving task. That is to say how much driver’s resources are absorbed by the System and whether it endangers driving safety. This encompasses, for example, System’s interaction modes (by touchscreen, by manual controls, by speech) and how high is the distraction needed to operate a simple task with the system.

The Usability Evaluation can spread among all phases of System Project, from initial Plot up to real implementation. Different methods to assess System Usability are used according the state of the Project, and different figures are involved. The User-Centered Approach involves the final user (that is, the driver, which is the one who will actually use the system once mounted onto a car) even at the beginning of projecting, in order to identify what are the expectancies of Users, and which tools are to be developed in order to fulfil them. Experts are at work in order to examine User’s collected data, and to identify a list of potential issues that could decrease System Usability.

In the following paragraphs are listed some of the most relevant Methods used when assessing System Usability, divided between Methods with Experts, where only Expert carry out the evaluation using different tools, and Methods with Users, where Users are involved directly, in different forms, in the assessing method.

4.1.1. Methods with Experts

In the following paragraph some inspection methods will be briefly described.

4.1.1.1. Heuristic Evaluation

Heuristic Evaluation is a usability engineering method for finding the usability problems in an user interface design so that they can be attended to, as part of an iterative design process. Heuristic evaluation involves having a small set of evaluators examining the interface and judging its

compliance with recognized usability principles (the "heuristics"). It is a fast, cheap method to assess System Usability Issues.
Using Heuristic Evaluation is most effective when several evaluators are used, in order to raise the number of identified issues. It would seem reasonable to recommend the use of about five evaluators, but certainly at least three. The exact number of evaluators to use would depend on a cost-benefit analysis. More evaluators should obviously be used in cases where usability is critical or when large payoffs can be expected due to extensive or mission-critical use of a system. The evaluator experience can help raising the number of detected issues as well, even though not all issues are commonly discovered by a simple use of the System by expert.
During the evaluation session, the evaluator goes through the interface several times and inspects the various dialogue elements and compares them with a list of recognized usability principles (the heuristics). These heuristics are general rules that seem to describe common properties of usable interfaces. In addition to these general heuristics, the evaluator obviously is also allowed to consider any additional usability principles or results that come to mind, that may be relevant for any specific dialogue element. Furthermore, it is possible to develop category-specific heuristics that apply to a specific class of products as a supplement to the general heuristics. One way of building a supplementary list of category-specific heuristics is to perform competitive analysis and user testing of existing products in the given category and try to abstract principles to explain the usability problems that are found. Every evaluator checks the System alone, and when all evaluations are done, the aggregation phase begins. This ensures uninfluenced and reliable results.
Heuristic Evaluation needs not a real System to be performed; evaluators can also work on sketches or incomplete products. This allows the method to be used in every step of development, not just at the end. Furthermore, this enables to identify issues at an early stage of the job, allowing easy recovery and saving time on further wrong developing.

PROs	CONs
• Fast and cheap method to assess System Usability Issues • A real system is not required; evaluators may operate on sketches or drafts of interface	• Several evaluators should be used to obtain good results (at least three, better five) • Not all problems are likely to be identified during an evaluation session

Table 4.1: PROs and CONs of Heuristic Evaluation

4.1.1.2. Checklist

The Checklist works as a synthetic Heuristic Evaluation, because it provides a list of topics to be filled in by an examiner, without users' involvement, to drive the design evaluation of a System. Checklists also allow the testing of many users according to the same characteristics and methods, providing a common assessment technique able to give quantitative results. Administering a Checklist is quite simple: it just takes to fill blanks in according to observed behaviour, often with Yes/No answers. The most important part in Checklists is their creation: because they have their source on Heuristics, the creator must be skilled enough to find correct Heuristics to assess the System, and to find the correct way to formulate items to obtain sound answers. Checklists should be used for simple assessment, while more complex phases of testing should be attended by experts.

PROs	CONs
• Very easy to administer (just check to box) • Simple method to assess many systems in the same way, with a reduced degree of freedom in judgement	• Relies only on specified aspects, not much flexible • Requires experienced experts to identify aspects to probe

Table 4.2: PROs and CONs of Checklists

Examples

- *A Safety Checklist For The Assessment Of In-Vehicle Information Systems: Scoring Proforma* (TRL Checklist.pdf)

  A long Checklist is specified in order to assess various aspects of IVIS Interface. The Assessment is about:

  - Interaction
  - Safety
  - Completeness of Information
  - On-Time Information
  - Navigation
  - Language terms

- EUCLIDE: Enhanced human machine interface for on-vehicle integrated driving support system (more in Appendices)
  Use of 3-points Checklists to assess System Impact on Driving Safety. System requirements according Usability, Acceptance, and Workload were defined. The points were

  1) Green : no deficiencies – System fully meets requirements
  2) Yellow: non-critical deficiencies: the system partly meets the requirements
  3) Red: critical deficiencies: the system does not meet the requirement

- RESPONSE Checklist for Assessment of Driver Assistance Systems [41]
  A Checklist to assess ADAS at an early System concept phase has been developed, after an overview of existing literature about checklists.
  Its main purpose are to assess systems
- which are designed to support the main driving task (or parts thereof), where the main driving task consists of all information acquisition, information processing and actions which are directly necessary to get from location X to location Y, and
- which are built in the car by manufacturer or in co-operation with the manufacturer

  The Checklist has been supported by a meta-questionnaire, whose purpose was to test the goodness of the Checklist. The questionnaire was administered to five experts (designers, engineers and physicists).
  Results of the questionnaire were a good performance of the checklist when compared to existing evaluation tools. However, it was also noted that the checklist is more suited for identifying inadequate product ideas than to differentiate between various well thought-out, mature product ideas. Indeed, the current form of the checklist mainly represents an “exclusion test” and less a tool for a differentiating evaluation of various prototypes. Some suggestions for improvements have been done, which rely on precisioning or changing of individual items on the checklist.

4.1.1.3. Guidelines

Guidelines instruct experimenters with general guidances to design a System. There exist a number of Guidelines a System must fulfil, and its goodness could be measured also according to the respect of Guidelines. A great amount of Standard Usability Guidelines are available, and the identification of the most suitable ones has to be done by skilled experimenters. In this case Experts are required to identify which aspects of the system are relevant to the testing, what guidelines state the correct principles able to assess the aspects previously identified, and at last they have to check whether the identified guidelines are respected by the System. A Guideline specifies how an aspect should be in order not to hinder user tasks, according to color, dimension,

position and so on. Checklists are different, because a specific question must be answered to probe an aspect.

This table can be found complete on the web site [81], where complete guideline papers can be found as well. Other information can be found on Usability Evaluation Methods and Guidelines, http://www.usabilitynet.org [84].

Guideline / Practice / Standard, *Title*	Main contents
AAM Alliance of Automobile Manufacturers *Statement of Principles, Criteria and Verification Procedures on Driver Interactions with Advanced In-Vehicle Information and Communication Systems*	Voluntary industry guidelines and best practices for future telematics devices including cell phones, navigation systems, and internet links.
Battelle *Human Factors Design Guidelines for Advanced Traveller Information Systems (ATIS) and Commercial Vehicle Operations (CVO)*	Voluminous document with references to interface design, heavy on trucks. User interface has been said to have a windows flavour, includes physical ergonomics information (e.g., legibility, control sizes) which are not included in the UMTRI guidelines.
EU Commission of the European Communities (1999). *Statement of Principles on Human Machine Interface (HMI) for In-Vehicle Information and Communication Systems ("EU Principles),*	Mostly motherhood statements, some minor revisions are expected soon.
HARDIE *HARDIE Design Guidelines Handbook: Human Factors Guidelines for Information Presentation by ATT Systems*	Commission of the European Communities, Luxembourg, p480 Early set of European guidelines, less data than UMTRI or Battelle.
JAMA *Guideline for In-Vehicle Display Systems*	JAMA (Japan Automobile Manufacturers Association) First set of detailed design guidelines for interfaces. These guidelines are voluntary in Japan but followed by all OEMs and sometimes by aftermarket suppliers, but some aspects are particular to Japan.
SAE J2364 *SAE Recommended Practice Navigation and Route Guidance Function Accessibility While Driving (SAE 2364)*	"15-Second Total Task Rule," specifies maximum allowable task time for navigation system tasks performed while driving when using visual displays and manual controls.
SAE J2365 *SAE Recommended Practice Calculation of the Time to Complete In-Vehicle Navigation and Route Guidance Tasks (SAE J2365)*	Calculation procedure used to estimate total task time (and compliance with SAE J2364).
UMTRI *Preliminary Human Factors Guidelines for Driver Information Systems*	First set of comprehensive design guidelines, including principles, general guidelines, and specific design criteria, with an emphasis on navigation interfaces.

Table 4.3: list of some Guidelines available for review

Developers should also realize that compliance with guidelines does not assure a system will be safe or easy to use, and because of limited research resources and the time required to achieve consensus for a guideline, everything that one needs to know about interface design does not appear in guidelines. Unfortunately, even with all of the research, guidelines, and feedback from several generations of products on the market, many basic safety and usability problems still seem to re-occur, and the need for an awareness of these problems and product usability testing is greater now than ever before [58].

4.1.1.4. European Commission and ISO documents

In this paragraph a list of documents prepared by The European Commission and the ISO TC22 SC13 WG8 are presented. These documents can be a reference during the designing and evaluation phase of different in-vehicles information systems. For each one is provided the state of deliverable (ISO Standard, Technical Report, Technical Specification…)[2], the complete title and a brief description of the contents.

A general official document about design of In-Vehicle Information systems is the following:

- **European Statement of Principles on Human Machine Interface for In-Vehicle Information and Communication Systems.**

  This document, developed by a task force of European Commission, has the scope to describe some general principles about overall design, installation, information presentation, interaction with displays and controls, system behaviour and information about the system. The "system" refers to the functions and parts, such as displays and controls, that constitutes the interface and interaction between the system and the driver.
  In this context the principles consider that the driver's primary driving task is safely controlling the vehicle through a complex dynamic traffic environment.

Another relevant document is:

- **European Statements Principles on Human Machine Interface for In-Vehicle Information and Communication Systems. Expansions of the principles.**

  Also this documents was developed by an HMI European Task Force, supported by the CONVERGE project Telematics Applications Programme-Transport Sector.

  This document contains an expansion of the Principles contained within the European Statement of Principles on Human Machine Interface for In-Vehicle Information and Communication Systems. The intention is to explain the meaning of each Principle in sufficient detail for work to begin on procedures to test if a specific system conforms to the Principles. To be in conformance a system must comply with each Principle which is applicable to the system. The document includes definitions of important terms and concepts. Where appropriate, some of these are re-stated within the expansion of each Principle.

In the following, a list of ISO documents developed by WG 8, about transport Information and Control Systems are shown. Each document is focused on specific topics.

- **ISO 15007- 1:2002 Road vehicles – Measurement of driver visual behaviour with respect to Transport Information and Control Systems – Part 1: Definitions and parameters.**

[2] for clarifications about the abbreviations see "List of abbreviations and Glossary" in this document . For information about ISO standard development and processes see http://www.iso.org/iso/en/stdsdevelopment/whowhenhow/proc/deliverables/pasetc.html

This standard defines key terms and parameters to estimate TICS visual impact on driver visual behaviour. It gives a brief glossary of used terms to correctly understand the standard and its contents.

- **ISO/TS 15007- 2:2001 Road vehicles – Measurement of driver visual behaviour with respect to Transport Information and Control Systems – Part 2: Equipment and procedures**.

  This is the following part of ISO/TS 15007-1, and it gives a guide about equipments to be used for the measures and procedures to be followed.

- **ISO 15005:2002 Road vehicles – Ergonomic aspects of transport information and control systems – Dialogue management principles and compliance procedures.**

  This standard deals with TICS ergonomics development, giving general ergonomics principles for the dialog, independently from specific adopted solutions.

- **ISO 17287:2003 Road vehicles – Ergonomic aspects of transport information and control system – Procedure for assessing suitability for use while driving.**

  This document describes a process to estimate whether a specific TICS, or a combination of a TICS with other on-board systems, is suitable to use by the driver during drive. This standard does not recommend specifical variables to evaluate TICS suitability, nor does it estabilish its suitability for use during drive.
  Suitability is focused on some aspects of usage that mainly influence driving behaviour.
  The main aspects, in the context of system usage learning, are:

  - Interference with driving actions;
  - controllability;
  - efficiency;
  - usage simplicity.

- **ISO 15008 Road vehicle – Ergonomic aspects of transport information and control systems – Specifications and compliance procedures for in-vehicle visual presentation.**

  This document gives a series of minimal specifications for the quality of the images of displays containing dynamic visual information, for such information is readable while the vehicle is moving.
  The standard CANNOT be applied in the following cases:
  - Head-up displays
  - CCTV
  - Maps and Topographical representations (Navigation system settings)
  - Static information (telltales)

- **ISO/TS 16951:2004 Road vehicles – Ergonomic aspects of transport information and control systems (TICS) – Procedure for determining priority of on board messages presented to drivers.**

  This technical specification describes how to establish a priority among TICS messages (navigation, traffic information, system status, electronic Toll / Fee collection, emergency calls an so forth), and non-TICS (telephone, telltales…) to be delivered to the driver, and presented during drive. It does not apply to mandatory messages.
  Several steps must be accomplished:

- *Define Driving context and situation* for each message the TICS can output, in order to identify the priority of the message according to the driving conditions. Hazardous conditions should be taken into account, and one to four different scenarios should be outlined; at least one scenario should describe the "worst case" the TICS message might arrive. Scenarios should deal with trip context, Road environment, Traffic situations, vehicle condition;
- *Selection of experienced evaluators*, who are people accustomed with the TICS and have good knowledge of road safety issues, along with knowledge about local traffic. Ten or more evaluators are required to obtain reliable results, when using the method described in the next point. When evaluators are less than ten, it is advisable to use the method described two points below;
- *Evaluate Criticality and Urgency of a message*, that is all evaluators should share the same idea about a message and its implications. Each evaluator rates a message according to its Criticality and Urgency in the current scenario, assuming he/she is the driver. When evaluating Urgency, Controllability is one of the important factors to be considered. If the situation is uncontrollable, no action shall be expected from drivers. However, if there is a possibility of controlling the situation, then Urgency shall be determined depending on when the system expects drivers to take an actions to handle it. This procedure leads to a Priority Index for each message, given by the average scores of all evaluators about a message;
- An alternative step of evaluating Criticality and Urgency is through the construction of a *Priority Matrix:* this method consists in experts making pair-wise comparisons of all messages, and through a mathematical process dealing with the Standard Deviation of Priority Indexes of evaluators, gives a table of prioritized messages.

Except from messages regulated by law, the driver should be in control of selecting, deactivating, and cancelling messages, independent of priority. Priority Rankings have the purpose to avoid the simultaneous presentation of messages, and in particular auditory messages.

- **ISO 15006 Road vehicles – Ergonomic aspects of transport information and control systems – Specifications and compliance procedures for in-vehicle auditory presentation.**

  This standard give ergonomic specifications to the design and installation of auditive displays that present, by means of vocal messages or sounds, information during drive. The document lists a set or requirements and recommendations for auditive messages on-vehicle, coming from TICS.

- **ISO/CD 16673 Road vehicles - Ergonomic aspects of transport information and control systems - Occlusion method to assess visual distraction due to the use of in-vehicle information and communication systems.**

  This standard provides a procedure for measuring visual demand, which can be used to assess visual distraction due to the use of visual or visual-manual interfaces accessible to the driver while the vehicle is in motion. It applies to both Original Equipment Manufacturer (OEM) and After-Market in-vehicle systems. This standard applies to both permanently installed and portable systems. This standard applies to any occlusion method and is not dependent upon one specific implementation.

- **ISO/CD TR 16352 Road vehicles - Ergonomic aspects of in-vehicle presentation for transport information and control systems - Warning systems.**

  This document will be transformed in a Technical Report, whose purpose will be the Review of Literature about Warning Systems. The report will include experimental experiences on

efficiency and acceptance of different modalities and combinations of warnings, the design of organizative parameters and codes, and of visual, auditive, and tactile warnings, ending with some recommendations.

- **ISO/WD Simulated lane change test to assess driver distraction[3].**

  This standard provides a measurement method for assessing the effect of driver distraction due to information and communication systems used in vehicles while the vehicle is in motion.
  The standard describes a specific dynamic measurement method, the lane change test (LCT). It describes the characteristics, the equipment and procedures to be used.
  The Lane Change Test is a simple laboratory dynamic dual-task method that quantitatively measures performance degradation on a primary driving-like task while a secondary task is being performed. In terms of a dual-task situation, performance in the LCT provides an estimate of driver distraction resulting from a draw of attentional resources from the primary task (driving) to a secondary task (e.g., operating an in-vehicle system).
  It is applicable to all types of user interfaces; manual, visual, haptic and auditory and combinations thereof. It applies to both Original Equipment Manufacturer (OEM) and aftermarket in-vehicle information and communication systems. This standard applies to both integrated and portable systems. Both commercial vehicles and passenger vehicles are included.

4.1.2. Methods with Users

Experts can assess systems only up to a certain point. Due to their expertise, they do not reflect "average user" behaviour. Due to this fact, it is mandatory to use "normal people", those who are likely to use the System in-car, when driving. The *Methods with Users* describe user's interaction with the System according to their performance using it, from Usability viewpoint. In this context, *Secondary Task Measures, Primary Task Measures* and *Self-Reported Techniques* will be examined.

4.1.2.1. Secondary-Task Measures

In usability field the secondary task is the use of a on-board device or any system or part of cabin object of evaluation. These activity, usually, are been doing during the primary and most important task that is driving. In this context driving is called primary task. In the paragraph which describes workload (see § **5.3**) a deep description of secondary and primary – task theory will be given.
Here a description of such measures rely on User Performance when interacting with the System will be analysed. They are: Number of Errors, Total Task Time and Recorded System Logs.

4.1.2.1.1. Number of Errors

The *Number of Errors* is self-explanatory, and expresses the number of incorrect interactions with the system a user does during a trial. Such errors can be [79]

- *Mistakes*, when the user wrongly interacts with the system due to incorrect mental model and therefore as consequence of wrong decision; such behaviour is conscious, and dictated therefore by poor system interface design, which leads to incorrect – possibly dangerous – conclusions about system's work;

[3] As all the ISO documents are described in this paragraph, this ISO documents is described here, although it refers specifically to workload and not to usability aspects.

- *Slips*, happening when user's attention is diverted by a process similar to the one he is performing. The diverting process is often more frequently performed, and are therefore unconscious.

Mistakes and slips can be identified by means of observation and verbal comments by users, and simply recorded by taking notes. A user can say explicitly he/she wanted to choose another function, but a slip took place and he/she selected another, or he/she might say he/she wanted to activate a specific function, but he/she mistakenly chose the wrong procedure.
Usability is highly related to the number of Errors an user does. Usable systems, by their own definition, should reduce to its minimum the number of errors. Observing a high number of Mistakes might lead to conclude that the System's design, not only is not affordable, but it inspires wrong – and perhaps dangerous – interaction. When observing many slips, it might be stated that Workload is too high in that situation to interact with system and contemporarily maintain safety level acceptable.
Errors can be also self-reported when using PSA-TLX Workload assessment method.

PROs	CONs
• High correlation between System Usability and Number of Errors: the more the errors, the less usable (and safe) the System.	• An expert of the system being evaluated is required to monitor User Interaction with the System • Users might modify their behaviour due to the presence of the expert

Table 4.4: PROs and CONs of Number of Errors Monitoring

Examples

- *Cyber Cars Iterative Design Of A New On-Board System For Public Vehicle: From The Idea To The Prototype* [67] (more in Appendices).
  Users were monitored when interacting with mock-ups and with the final prototype. Error number was in general very low, and the task with more errors (find a radio station to listen to specific content), had so many errors because users mistakenly interacted with interface to achieve a goal, but imagined process was different from the actual one.

- *Subjective Evaluation of The Mental Workload in the driving Context* [63]
  Observation of the global drivers' behaviour by the experimenter in order to identify driving errors and unusual steering wheel movements has been carried out.

- SENECA Project: Users Evaluation [56] (More in Appendices)
  The number of errors when interacting with a vocal system to command a large number of on-board devices was used to assess Usability of the System. Results were good, and showed better results for elderly subjects rather than younger subjects when comparing performances with manual commands.

4.1.2.1.2. Total Task Time

Total Task Time (TTT) is the measure of required time to complete a task. It is a function of Task Complexity and Environmental Conditions, as well as Workload. Typically, low TTT means that it is possible to interact with the System while maintaining acceptable levels of attention on driving Task. When more time is required to complete a Task, then the System must support Task Interruption and Resuming, in order to allow drivers to safely switch from System interaction and Road checking, and vice-versa. A Baseline is calculated, that is how much time the user takes to complete the task when not driving. TTT shall be longer than Baseline mean value, as driver attention is shared between System and Driving Tasks.

PROs	CONs
• Low TTT means the users interacts in a short time with the system, so more time is devoted to the driving task	• TTT must be monitored and causes of its oscillation must be clearly identified, to see whether high TTT should be detrimental to driving safety

Table 4.5: PROs and CONs of Total Task Time Monitoring

4.1.2.1.3. Recorded System Logs

Recorded System Logs track the user's interaction with a system, by detecting the number of interactions an user does. Crucial aspects of interaction are their type and order, timings and comments, the page. System Logs can be captured either by an automatic system, cabled into the system itself, or by an Observational grid. The former method is more accurate, giving the ability to accurately report interaction timings as well as type and order of pressed buttons, sequence of viewed pages particular interaction strategies (e.g. shortcuts), but requires cabling the capturing device inside the system under evaluation. The latter method is less expensive, but also less precise. More discussion about system logs can be found in § 5.3.

PROs	CONs
Give a complete tracking of user's interaction with a system Collected data easily administrable	Might be expensive when automatized Observational Grids must be constructed and filled in real-time, unless a tape is provided

Table 4.6: PROs and CONs of Recorded System Logs

Examples

- SENECA project [56]: users' interactions were tracked by an experimenter sat in the rear seat of the car. An Observational grid was used, where also timings where collected, as well as behavioural driving aspects, such as stopping the car, driving too fast, waiting before beginning a task, and so forth. Complete interaction was tracked, in order to capture the order and number of times interactions took place. The collected data was useful to cunduct deep User Error's analysis. An electronic featured, cabled into the system, allowed to collect correct timings for interactions as well as detailed history of navigation through vocal interface.

4.1.2.2. Primary- Task measures

In order to collect information about vehicle behaviour System Logs is recorded. A log is a comprehensive list of measurements, recorded during a driving session and usually taken by electronic means. The nature of the contents of the log is usually a set of sensitive car parameter values (or its variations) such as speed, steering angle, lateral position, number of brakes, number of Line Crossing, as well as camera recording of Subject visage and interaction with the system. In this context, such data is evaluated to see how usable a system is, when used during driving. A system with high usability scores when the car is parked, but poorly usable when the car is moving, is badly designed due to unacceptable safety level decrease.

Logs are collected electronically, and rely directly on on-board instrumentation. For this reason, taking complete logs might be time-consuming, because a car must be appositely prepared for such probing, and money-consuming, because such instrumentation can be expensive. Conversely, the benefits of having complete values of a driving session apply for statistical elaboration, and further usability probing, relying on camera data, for example.

Almost all experiments described on Appendices make use of System Logs. Their nature in inspecting Usability and Workload relies on statistical analysis of data, compared with *Baseline* Performance. Baseline Performance is the Subject's "normal" performance when driving, and it is intended with the system under testing is deactivated. Several trials are taken, in order to evaluate a mean Baseline performance. Data Collected with the System active shall be confronted with such Baseline, not absolute, but concerning each Subject. This allows to see how much a System impacts on personal behaviour while driving, and allows to decide whether a system can be safely used during a driving session.

4.1.2.3. Self-reported Measures

Self-reported Measures rely on user's self reporting of information. This kind of techniques is different from other techniques because subjective and personal evaluation of self-behaviour and System behaviour is done.

4.1.2.3.1. Questionnaires

Here will be described some examples of questionnaire used to evaluate usability aspects. A general description of this technique has already been given in § 3.5.

Examples

- ADVISORS Final Report [2]: Use of Questionnaires for assessing different ADAS; questionnaire results used for further development
- COMUNICAR Usability Questionnaire (based on Volvo Study) realised as a 10-points scale, from Strongly Disagree to Strongly Agree, for each question [65];
- EUCLIDE [45]: Questionnaires to assess Usability, Acceptance, and Workload
- SAE [10]: Questionnaires about meaning of icons, even though testing pointed mainly on ranking
- SENECA [56]: Use of Pre/Post/Final Questionnaires to inspect subject's Acceptance, expectations, comparison between vocal and manual input systems

4.1.2.3.1.1. Brooke Questionnaire

This questionnaire [7], also known as System Usability Scale (SUS), was developed by John Brooke in 1986, as part of the introduction of usability engineering to Digital's integrated office systems programme. Its objectives were to provide an easy test for subjects to complete (i.e. minimal number of questions), easy scoring, and to allow cross-product comparisons. It has been used extensively in evaluations of projects in Digital World (office systems, system management, technical tools and hardware systems) and has been estimated simple and reliable.

The SUS scale is generally used after a user has tried a system but before any debriefing or discussion takes place. Users should be asked to record their immediate response to each item, rather than thinking about items for a long time.
All items must be checked. If users feel they cannot respond to a particular item, they should mark the centre point of the scale.

Scoring the SUS scale

The Likert scale presents a set of statements. Subjects are asked to express agreement of disagreement on this scale, typically a five-point one. This scale is a Likert scale and yields a single number representing a composite measure of the overall usability of the system being studied. Note that scores for individual items are not meaningful on their own.

To calculate the SUS score, first sum the score contributions from each item. Each item's score contribution will range from 1 to 5. For items 1, 3, 5, 7 and 9 the score contribution is the scale position minus 1. For items 2, 4, 6, 8 and 10, the contribution is 5 minus the scale position. Multiply the sum of the scores by 2.5 to obtain the overall value of SU.
SU scores range is 0 to 100.
The SU scale may be used as long as proper acknowledgement is made of the origin of the scale.

The SU scale was constructed from an original pool of 50 items. Ratings were made by 20 users on all 50 items for 2 systems, one designed for end-user use and one designed for use by systems programmers. These two systems were chosen to represent extremes of usability.
In the end, 10 items were selected which evoked the most consistent and most polarised responses. The items selected have interrelations between 0.7 and 0.9.

System Usability Scale Questionnaire

Question	1 = Strongly Disagree, 5 = Strongly Agree				
1. I think I would like to use this system frequently.	1	2	3	4	5
2. I found the system unnecessarily complex.	1	2	3	4	5
3. I thought the system was easy to use.	1	2	3	4	5
4. I think that I would need the support of a technical person to be able to use this system.	1	2	3	4	5
5. I found the various functions in this system were well integrated.	1	2	3	4	5
Question	1 = Strongly Disagree, 5 = Strongly Agree				
6. I thought there was too much inconsistency in this system.	1	2	3	4	5
7. I would imagine that most people would learn to use this system very quickly.	1	2	3	4	5
8. I found the system very cumbersome to use.	1	2	3	4	5
9. I felt very confident using the system.	1	2	3	4	5
10. I need to learn a lot of things before I could get going with this system.	1	2	3	4	5

PROs	CONs
Easy test for subjects to complete (minimal number of questions), easy scoring, possiblity of cross-product comparison Simple and reliable	

Table 4.7: PROs and CONs of Brooke Questionnaire

See Table 3.5 (p. 28) for general PROs and CONs of Questionnaires

Examples

- COMUNICAR: Human factor tests on car demonstrator The methodology [14]
- Usability Questionnaire (based on Volvo Study) realised as a 10-points scale, from "Strongly Disagree" to "Strongly Agree" , for each question
- ADVISORS [2]: Brooke Questionnaire to assess Usability and Acceptance of LSS (Lateral Support System).

4.1.2.3.2. Thinking-Aloud

Thinking-Aloud is a self-explanatory technique: when a Subject interacts with the System, he/she is asked to think aloud, so to express every thought he/she has about the system. This is suited to test Subjects' expectancies at any point of interactions, what they would like to do, which elements surprise him/her, which is the frustration level they run into, and can be used to assess System Usability. With this method, it is possible to understand which parts of the system are more difficult to reach, and what is the mental process the Subject runs to find a way out to his task completion. Subjects freely express their impressions, as well as their negative impact on system, that will reflect on its Acceptance. Observers collect the flow of thoughts in order to analyse them carefully at a second time.
This method allows capturing very personal and subjective impressions, hardly collectable in a different way, because they are told as they happen. With a questionnaire, for example, Subjects might suffer lack of memory, which might lead to incomplete results or misleading ones. Terminology is also important: Subjects will use their own terms to describe the System and the interaction with it; such terminology can be examined to decide whether it can be used to describe System components.

Thinking-Aloud is easily done with low costs, no extreme need of videotaping (notebooks work well), and allows a great number of subjects to attend it. A drawback is the constraint to observe one Subject at a time, in order to watch correctly his behaviour, and avoid influence by other people about task completion as well as simple distraction.

It can be argued that Thinking-aloud might be disruptive with the task being accomplished, but often it is a means that gives particular important information, whose value compensates disruption. It is, anyway, important to instruct Subjects on how to execute Thinking-Aloud correctly.

PROs	CONs
• Allows to capture subjective way to think, expectations, problems, frustrations • Not expensive	• One subject at a time can be monitored • Might be disruptive • Highly subjective • Sometimes subjects are not willing to express their thoughts to "nobody"

Table 4.8: PROs and CONs of Thinking-Aloud

4.1.2.3.3. Co-Discovery Learning

Co-Discovery Learning (CDL) is about collaborative task completion. Subjects, in pairs, test the System in order to accomplish a pre-defined task, and help each other to do that, as they are working together to accomplish a common goal. A scenario is described, and they are asked to do a list of tasks pertinent to that specific scenario. During task accomplishment, Subjects are asked to express their thoughts. This method makes thoughts verbalization more natural for Subjects, pairing off to Thinking-Aloud. Conversely, this method might lead to biases and influence due to multiple subjects at a time. For this reason, this method should be used with pairs of Subject having the same knowledge and, possibly, who know each other, to reduce inhibition. A two-face aspect is the interactivity between subjects: if on one hand Subject couple comments are more spontaneous than when Subjects test the system alone, on the other it is quite difficult to test performance with the system, because Subjects do not just interact with it, but interact with themselves, too. The Number of Errors will be reduced in CDL, not (just) because of system goodness, but because Subjects are more prone to "think before act". Qualitative data is likely to be much more accurate with this technique rather than Thinking-Aloud, while the latter is more significant for Error and Performance inquiry. CDL can be used at any stage of development: mock-ups can be provided in place of a real system and interactions can be recorded by means of paper or tapes, according to available time and money resources.

PROs	CONs
• High level of acquired Qualitative information • Subjects are more prone to communicate their thoughts to another person than to "nobody" • Usable at any stage of development	• Influences and biases might be inserted when non-homogeneous Subjects are paired off to test the System • Performance and Error Number count difficult to take

Table 4.9: PROs and CONs of Co-Discovery Learning

Examples

- Iterative Design Of A New On-Board System For Public Vehicle: From The Idea To The Prototype, CRF [67] (More in the Appendices)
  A battery of eight tests was realised, to test various aspects of HMI for the Cyber Cars Information System.

4.1.2.4. Expert-reported Techniques

This category includes all Techniques where experts sit by users' side and track their interaction with the System. When an Expert assesses the System, his/her knowledge about the System itself and his/her expertise on previous test is the discriminant to lead a good test. Observational grids are often used, where the Expert just fills a form with information about current user. Other techniques rely on simple observation and note-taking. It is important to understand what and when to observe, and how to record important clues.
We discuss User Monitoring Technique, for it is the most general, less expensive and less demanding in terms of additional furniture.

4.1.2.4.1. User Monitoring

User Monitoring is carried out by experts who look at users interacting with a System, either directly or by means of recorded media. Notes are taken during observation, to capture user's behaviour, at any stage of System development. At early stages, it can be useful to get user's requirements, and information about user task execution. Furthermore, it is the only applicable means when a System prototype is not present and therefore objective measures such as electronic ones cannot be applied. In fact, it would be too expensive to cable a real product.
Observation by means of recorded media is useful to find more clues which would have gone unseen otherwise, while direct observation allows experts to focus on precise areas, at the price of sacrifying other topics. Traditional methods are less complete, but are fast to analyse. Conversely, recording data is more precise, but more time-consuming in the analysis phase.
Because of the presence of the Observer, user might feel embarrassed and their behaviour might result different from normal. Such bias must be contained by observers, in order to maintain data reliability. It is important to make users feel confident with the observer, and they should be informed of the reason of the observation. This is automatically done when users are Subjects recruited for testing purposes.
A single try observation should be driven, in order to take note of timings, required media, and procedures during real testing. This helps saving time and money.
Observation can be the direct prelude to design, but sometimes it is the starting point for organizing Focus Groups, Interviews, Questionnaires in order to clarify given aspects and create a more complete scheme of user's behaviour with the System under evaluation [83].

Experts sometimes rely on use of Observational Grids, pre-defined checklists filled during user's observation. This provides a standard way to assess user's behaviour, and ensures particular aspects to be taken into account during observation. This, of course, requires the definition of scenarios to be run by the user, and this might limit user's freedom in the interaction.

PROs	CONs
Applicable at any stage of development Allows direct information gathering about user's interaction with the system	User behaviour might be modified due to the presence of the observer Skilled observers must be employed to achieve correct results

Table 4.10: PROs and CONs of User Monitoring

Examples

- Iterative Design Of A New On-Board System For Public Vehicle: From The Idea To The Prototype, [67] CRF (More in the Appendices).
- At early stage of development, users were observed when interacting on mock-ups of user interfaces. Errors, interaction behaviour strategies, comments were recorded by the experimenter through direct observation. Along with a semi-structured interview, results were used to choose the best interface scheme and the correct labels to use in the HMI.

4.2. Acceptance

Acceptance is the measure according to which a user expresses the will to use a system and the level of appreciation resulted after use, in terms of safety improvement, reduction of vehicle operation costs, saving in travel time, improvement in driving comfort, HMI friendliness… System performance quality influences Acceptance as well.

Acceptance can be assessed:

- Before a system is evaluated, in terms of how useful it seems to be a system such as the one under evaluation,
- During evaluation, in order to see which parts of interaction bring frustration to users, and
- After evaluation, to see whether the use of the system has modified prior Acceptance positively or negatively.

Acceptance is a subjective measurement. It is not possible to assess it in objective terms, for several factors such as age, expertise, knowledge as well as personal needs and likes are strictly tied to evaluation results. Therefore, the measurement can be assessed through subjective means, such as Questionnaires and Interviews. Below, three different methods to assess Acceptance are presented: Semantic Differential Technique, Willingness to Pay, Importance Ranking.

Examples

- ADVISORS*: An Evaluation Study of the Lateral Support System:* Acceptance Questionnaire (based on Semantic Differential Technique) to test Acceptance of LSS. Results were rather positive. [2]
- *Iterative Design Of A New On-Board System For Public Vehicle: From The Idea To The Prototype* [67]
  A Questionnaire to assess user's Acceptance in using working prototype was administered at the end of test phase.
  Results indicated good Acceptance according to innovation, user expectation consistency, learnability.
- SENECA Project [56]: Users Evaluation Semantic Differential scale to assess Willingness to have a vocal system to command several on-board devices when driving
- IN-ARTE [4]: HMI acceptance is verified with a –2 to +2 Likert Scale. Results are an higher Acceptance of system for experienced drivers and evaluators.

4.2.1. Self-Reported Measures

4.2.1.1. Standardised Attitude Scale Calculus Algorithm

This method [85] is based on a Semantic Differential Scale that measures people's reactions to stimulus world, and concepts in terms of ratings on bipolar 7-point scales defined with opposite adjectives at each end.
In particular, this algorithm tries to represent Acceptance of a Telematic System according to the fact the System is:

1. Useful	2. Pleasant	3. Good
4. Nice	5. Effective	6. Likeable
7. Assisting	8. Desirable	9. Raising alertness

The algorithm is described as follows:
1) Describe the system to be evaluated in terms of *what is your judgement about a system that would…* and present the nine items (before-measurements);
2) Present the nine items again after experience with the system under evaluation with the description: *what is your judgement about the system… you just finished driving with*;
3) Individual items should be coded from *+k* to *–k* from left to right, scores on items 3,6, and 8 should be coded ranging from *–k* to *+k* (the items are mirrored), [where $k \in \Re$ is an opportune value, ndr];
4) Perform reliability analyses on the before-measurement. Use item 1,3,5,7, and 9 for the usefulness scale, and item 2,4,6, and 8 for the satisfying scale;
5) If reliability (Cronbach's α) is sufficiently high (above 0.65), compute per subject the end-score for the two scales by averaging the scores on items 1,3,5,7, and 9 for usefulness score, and averaging scores on items 2,4,6,and 8 for the satisfying score;
6) The usefulness scale can now be averaged over subjects to obtain an overall system practical evaluation. The same can be done with the satisfying scores;
7) Compute difference-scores per subject by subtracting the before-measurement score from the after-measurement score per scale. The difference scores show whether and in which direction subjects' option was altered as a result of experience with the system;

Step 1 and 7 can be left out if evaluation of Acceptance is limited to after-measurement. Step 4 is still required because the scale was translated from Dutch. In this case in Step 4 the after-measurement scores should be used.

Examples

- COMUNICAR [14] used such a method to assess Information Manager
- ADVISORS An Evaluation Study of the Lateral Support System (LSS) [2]
- LSS Acceptance was tested with a technique similar to Semantic Differential, using a range from –3 to +3. Results were rather positive, pointing LSS to be likely to be used in-car.
  Standardised Attitude Scale to assess Acceptance.

4.2.1.2. Willingness to Pay / Use / Purchase

Willingness to Pay (WTP) expresses the measure by which a user is prone to pay to have a particular system installed on-car. It is a very important measure for marketing purposes, to see whether or not the new system could be produced and sold with profit.
As described in [94], "to protect consumers from the potential abuse of monopoly power most utilities are subject to both price controls and quality of service controls. However, there is tension between these controls. The threat of too little or too much quality accompanies the regulator's decision on allowable revenue in pricing reviews. Desirably this threat could be reduced by information on the willingness to pay for different levels of service quality and the cost associated with them."

It is important to understand user's perception of different quality levels, in order to get a clear estimation of user's WTP. In order to do so, users should be carefully chosen to reflect consistent market characteristics [94 p15].

This measure can be assessed in different methods; more about points 4-9 can be found in [94 p18-28]

1) Precise questions in questionnaires or Interviews. The user expresses how much he is willing to pay to purchase the system under evaluation;
2) Through a method similar to a Ranking, users are asked to assign different Systems an amount of limited resources. In this way, it is possible to understand the relative importance of

a system paired off with others, and how much of the limited resources one is willing to assign to such system;

3) A “Bidding Game” is used. Through a binary series of choices, users are asked whether they are willing to pay a certain expense for the system. According to the answer, the expense may grow or diminish, until the user changes his willingness (e.g. if the willingness to pay grows up to a certain amount, and the next step is considered too high for the user, then the previous step represent the user’s willingness to pay for the system under evaluation).
4) Contingent Valuation: the value estimates are contingent on a hypothetical scenario that is presented to respondents for valuing. The original form of Contingent Valuation (CV) constitutes an open ended question, in which respondents are asked to state their willingness to pay (or accept compensation) for a specified system. This method is now rarely used because it has been found to be vulnerable to a range of biases. It faces special problems where the system in question is not purchased directly by the public.
5) Referendum CV: This technique involves asking respondents to make a discrete choice between two alternatives: Pay nothing (extra) and maintain the status quo level of quality, or pay a specified ‘bid amount’ in return for an improved level of quality. For example, a typical question might be: “Would you be willing to pay $A to secure a quality improvement of X units? (yes or no)” The preference data generated using this method is encoded in binaryform, as respondents are only given the option of answering ‘yes’ or ‘no’. This method is inappropriate to test several aspects of a system at once.
6) Choice Modelling: in this method, an existing system is compared with others, where significant differences are outlined, along with the difference of price. The user chooses the system which is best in his opinion, even according to the increase of price. This method allows to see whether the user is willing to accept the increase of price to obtain the perspective increases in performance, or whether he decides that the increases in performance are not worth the increase in price. This method requires more cognitive load, and it makes it harldy adiministrable by phone;
7) Conjoint Rating: users are presented with comparisons of different alternatives with the existing product one at a time, and are asked to express how much they prefer the new system by using a numerical scale. This technique is prone to insert biases in data regression, and does not give any information about the choice between different alternatives, but just how much is the alternative felt different from the original system;
8) Conjoint Ranking: users are presented with all alternatives at once, and are asked to rank them along with the original system. Data collected with this method is difficult to understand;
9) Paired Comparison: two alternatives to an original system are proposed, and users are asked to express their preference towards one system or the other, within a numerical scale.

The measurement can be influenced by several personal factors, such as age and expertise and / or localization and climatic means, that is the geographical position and usual weather conditions. It was observed a slightly different pattern in answers when systems are to be bought separately from the car, or when systems are already part of normal stuff of a new car. Several test show a null Willingness to pay for systems as well.
Systems with too much complexity, boring behaviours, trying to control drivers are examples of systems with low Acceptance levels.

PROs	CONs
Gives a precise idea of how much users are willing to pay to have the System installed in their cars	Personal factors such as age, expertise, localization, weather conditions influence the response

Table 4.11: PROs and CONs of Willingness to pay

Examples

- ADVISORS: An Evaluation Study of the Lateral Support System [2]: ADVISORS Questionnaire. Ratings were 100-500 € for LSS.

4.2.1.3. Importance Ranking

Importance Ranking can be used to identify the way users perceive the relative importance of a set of pre-defined items. They are asked to rank a number of items according to a specific criterion. This can be useful when choosing among different possibilities to realize a system (different mock-ups to elaborate, different interfaces to be presented, different sets of icons, terms an so forth). Some techniques described in the previous paragraph can be used to assess relative importance of different systems of alternatives.

PROs	CONs
Helps identify user's perception of relative importance of a set of pre-defined items	Subjective data that must be treated in a statistical way to give consistent results

Table 4.12: PROs and CONs of Importance Ranking

Examples

- *Cybercars Iterative Design Of A New On-Board System For Public Vehicle: From The Idea To The Prototype* [67] (more in Appendices).
  Users were asked to rank the different types of keyboard; that allowed to choose the most performant keyboard, as analytic tests confirmed to be the one ranked first by users.

- *Comprehension Testing of Active Safety Symbols* [10] (more in Appendices).
  Users were asked to rank different definitions and icons to represent a specific context. This way, it was possible to understand which icons matched together, and which icons were the most indicated, according to users, to represent a specific context, allowing them to realize in a shorter time the correct meaning of control. Such a result allows the Workload decrease, as well as critical incidents better avoidance.

5. Workload

This chapter is an overview of Workload and the techniques used to assess it. A detailed exam of Workload can be found in AIDE 2.2.1 Deliverable named "*Review of existing Techniques*".
Workload (WL) is the specification of the amount of information processing capacity that is used for Task Performance. In the concept of Mental Workload, how the goal is reached (e.g. the order of actions) and individual restrictions imposed upon performance (e.g. in terms of accuracy or speed) are included [16] . The difficulty of a Task is related to the processing effort (amount of resources) that is required by the individual for task performance, and is dependent upon context, state, capacity and strategy or policy of allocation of resources. Tattersall [20 p223] gives practical Guidelines for WL assessment.

WL, when dealing with Drive Task, refers to the concept that an user must share mental resources when driving and doing something else, such interacting with an IVIS / ADAS. It is a great concern to understand how resources are shared, because subtracting too much resources to driving task might lead to unsafe situations, therefore compromising safety. When assessing an in-car system, it is therefore important to verify, beside Usability and Acceptance issues, the level of Workload the system induces. Poor Usability of the System might lead to higher levels of Workload, thus lowering safety. The relation between WL and Performance depending on task demands is a non-linear relation, so an increase in task difficulty might lead to a great increase of Workload.
Here briefly will be listed three workload-measurements groups: self-reported (subjective) assessment measures, physiological measures and objective measures.

5.1. Self-reported Measures

The most common way Subjective Assessment is done is to ask the subject to rate different aspects of efforts using a predefined rating scale while guided by a verbal description for each level of the scale, or in the form of questionnaires, or also structured/unstructured interviews [42]. DeWaard [16] prefers the term "Self-Reported" to the term "Subjective", when referring to measures of Workload, because measures from other measurement groups, in particular physiological measures, are also subjective. They represent the best way to express experienced mental load, for they are expressed by the person who experiences it.
Subjective measures have some of the following properties, corresponding to evaluation's criteria

- Sensitivity
- Diagnosticity
- Selectivity
- *Reliability*
- *Intrusion Degree*
- Validity
- Ease of Use
- Needed Equipment
- *Acceptability by the subject*
- Inter-dependence between criteria
- Transferability

Unidimensionality versus Multidimensionality

A great use of standard scales exists when assessing WL in a self-reported way. Such scales may relate with one aspect of WL at a time (Unidimensional Scales) or deal with several aspects of WL at a time (Multidimensional Scales).

- Multidimensional: preferable when Diagnosticity is of great concern
- Unidimensional: Global Rating of Workload, providing a measure more sensitive to manipulations of task demand.

PROs of Subjective Measures	CONs of Subjective Measures
• Easily accepted and used by Subjects • Require no particular equipment • More sensitive than Objective measures when biases are low • Can reveal some internal mental workload level that can't be directly measured • High face validity • Application Ease • Low costs • Low primary-task intrusion secured as long as the scale is administered after completion of the task	• Do not assess peaks, so Objective measures are required • The diagnosis quality is low, due to the unclear understandability of the word Workload by Subjects • Different personal factors can influence perceived workload, so validity can be altered • Due to low retain of Subjects, they must be conducted as soon as possible • Subject rating might not reflect the real mental workload level, or be influenced by other biases such as dislike or unfamiliarity of the task • Subjects might be reluctant to report that things are difficult • Possible confusion of mental and physical load in rating • Subject's inability to distinguish external demands from actual effort or workload experienced • A Possible dissociation between self-report measures and performance might be an aspect that restricts use. • Limitations in the Subject's ability to introspect and rate expenditure correctly, which, e.g., become obtrusive in conflicting findings in that either peak workload or average workload level determine the final rating

Table 5.1: PROs and CONs of Workload Subjective Measures

5.1.1. Subjective Measures

5.1.1.1. One-Dimensional Scales

As anticipated before, One-Dimensional Scales rely on just one dimension to assess Workload. They range from a minimal number to a maximum number, and they are often discrete. Sometimes, they use verbal descriptors to identify and explain different levels of the scale. The usage of One-Dimensional Scales is very simple to administer to Subject, as they only have to mark the level that better describes their Perceived Workload, and for analysts, for they can treat subjects' outcome statistically very easily.

5.1.1.2. Multi-Dimensional Scales

Because of multiple facets Workload can have, and because of multiple factors Workload can be affected by, it is often insufficient to test overall Workload. In order to understand what factors are more meaningful for WL in a given case, WL is broken into components, each relying on different parts of driver behaviour and resource used. Subjects are required to rate Workload for each aspect, and then data is aggregated in order to give an overall result. This way, it is possible to assess single parts of WL, which is useful to understand what parts of the system need revision, and to have an overall value, useful to compare different systems.

5.1.1.3. Summary of Workload Assessing Techniques

Below, a table which summarizes the scales used to assess WL is presented. For a detailed description of each technique, please see AIDE Deliverable 2.2.1.

Method	Use		Good Aspects	Non-good Aspects
National Aeronautics and Space Administration Task Load Index (**NASA-TLX**) (Hart & Staveland, 1987) [26]	Multidimensional scale on six factors of WL (0-100 each) 1. Mental Demand 2. Physical Demand 3. Temporal Demand 4. Performance 5. Effort 6. Frustration Level Rating of Relative importance of factors by users: Factors are presented in pairs and Subjects choose the most relevant		❖ WL Measurement over a long period of time ❖ More sensitive than MCH and SWAT ❖ Reliable WL measuring technique due to consistency sensitiveness to increases in WL in different studies ❖ Weighting increases sensitivity and allows to identify causes of WL	❖ Detecting peaks or short-lasting increases in WL ❖ Veltman and Gaillard experimentally proved RSME more sensitive than NASA-TLX The authors argue that this result may be related to confusion caused by the TLX-subscales. [16, p33] ❖ Relationship among NASA-TLX, driver behaviour and critical incidents to be experimentally proven ❖ Duration of Comparison Phase ❖ Comprehension of each factor may be different among Subjects
NASA bipolar rating scale (Original version of NASA-TLX)	Based on 10 bipolar scales: - global WL - Task difficulty - Temporal demand - Performance - Mental/sensory effort - Physical Effort	- Frustration Level - Stress level - Tiredness - Activity Type (skills, rules of knowledge…)	❖ Sensitive to task demand	❖ Factors difficult to understand and differentiate

<table>
<tr><th>Method</th><th colspan="3">Use</th><th>Good Aspects</th><th>Non-Good Aspects</th></tr>
<tr><td rowspan="3">PSA-TLX

(PSA Peugeot-Citroen) (Multidimensional Scale) [66]</td><td colspan="3">Developed by PSA Peugeot-Citroen
Seven-dimensions method, 4 WL factors, 3 axes of evaluation</td><td rowspan="3">• Uses Multidimensional means as NASA-TLX as well as descriptive markings such as RSME
• Better than NASA-TLX in driver’s understanding of scale terminology (leads to better results in questionnaires)</td><td rowspan="3"></td></tr>
<tr><td>Dimensions</td><td>Factors</td><td>Axes</td></tr>
<tr><td>1. Trajectory Control - Vehicle position on the road (lateral)
2. Trajectory Control - Control of speed (longitudinal)
3. Reactivity to dynamic environment
4. Reactivity to static environment
5. Itinerary following
6. Appropriate use of controls and driving equipments
7. Reactivity to safety and status signs</td><td>- Stress (tension, constraint, apprehension, unsafe feeling)
- Fatigue
- general dissatisfaction (discontent, disappointment)
discouragement (loss of motivation, interest missing)</td><td>❖ Disruption
❖ Effort
❖ Driver State

Driver State applies to Factors, while Dimensions are assessed with both other axes</td></tr>
</table>

Method	Use	Good Aspects	Non-Good Aspects
Driving Activity Load Index (**DALI**) (INRETS – France) [Variant of NASA-TLX]	Adapted to driving activity, specifically to evaluate WL generated by in-vehicle system use. Dimensions are: 1. Attention Effort 2. Visual Demand 3. Auditory Demand 4. Temporal Demand 5. Interference (disruption induced by two tasks, system use and drive, carried out at the same time) 6. Situation Stress (constraint or stress level such as tiredness, insecurity, irritation…)	❖ Same as NASA-TLX ❖ Specific for Automotive application	❖ Difficulty to identify the target of assessment: system or driving? ❖ Difficult Comprehension and differentiation of Factors ❖ Factors not relevant to evaluate WL induced in driving activity, only WL induced by use of system while driving ❖ Procedure not yet validated ❖ Advisable to use this method with objective data
Subjective Workload Assessment Technique (**SWAT**) (Reid & Nygren,1988) [43, 72]	Three-dimensional WL assessment scale 1. Time Stress / Temporal Demand 2. Mental Effort 3. Psychological Stress Originally, card-sorting technique: 27 rating scale combinations to be ordered; then, a single WL scale is constructed and a procedure for event scoring created.	❖ Reliable ❖ Sensitive ❖ More Sensitive than MCH Scale	❖ High Time demand for card sorting ❖ More relevant for studies centered in individual differences ❖ Only three WL dimensions used ❖ Less reliable than NASA-TLX ❖ Less sensitive than NASA-TLX ❖ Validity still needs to be proven in research
Modified Cooper Harper Scale (**MCH**) (Wierwille & Casali, 1983) [90] DECISION TREE	❖ Unidimensional Scale ❖ Scale 1-10. 1 is low level. ❖ Applies to cognitive tasks. ❖ Evaluates Task difficulty and effort	❖ Sensitive to variations in task difficulty	❖ Not suited for short-lasting WL variations ❖ Less sensitive than NASA-TLX ❖ Less sensitive than RSME Validity to be determined in research ❖ Less reliable than other methods

Method	Use	Good Aspects	Non-Good Aspects
Overall Workload (**OW**) (Vidulich & Tsang 1987) [13] SIMPLE SCALE	❖ Bipolar scale from 0 (very low) to 100 (very high), graduated by 5 units	❖ Easy to use ❖ More Sensitive than MCH and SWAT ❖ More Acceptable than MCH and SWAT	❖ Less valid than NASA-TLX ❖ Less reliable than NASA-TLX ❖ No Diagnosticity
Test of Estocolmo University SIMPLE SCALE	❖ Perceived difficulty and effort ❖ 1 scale for difficulty: 9 points + verbal description ❖ 1 scale for effort: 0-10 scale + verbal description	❖ Sensitive to the difficulty variations ❖ Used to evaluate subject reasoning and verbal comprehension	
Subjective WORkload Dominance **SWORD** (Vidulich et al, 91) [87]	Adaptation of AHP (Analytical Hierarchy Process) 1. Comparison of different tasks in pairs. For each pair, subjects indicate in 17-levels scale the task inducing the most important WL 2. Creation of a Judgement matrix by the experimenter 3. The means are computed to obtain notes given in each task evaluated	❖ Sensitive ❖ Reliable ❖ Easy to use	❖ Depends on the number of tasks to compare ❖ Scale format problems: Definition of WL may be different for subjects, factors evaluated may not be the same
Rating Scale Mental Effort (**RSME**) (Zijstra & Van Doorn, 1985) [94] SIMPLE SCALE	❖ Unidimensional Scale ❖ Ratings indicated on a vertical line from 0 to 150mm ❖ A number of anchor points labelled with verbal descriptor of effort (no effort, extreme effort) ❖ Higher degree of invested effort considered as an attempt to keep performance at a certain level in response to increased task demands	❖ Easy to administer after and during driving ❖ One of the most sensitive compared to other methods (Verwey & Veltman, 1995) ❖ High reliability: higher workload ratings as a function of task load ❖ Not intrusive ❖ Good Acceptability ❖ RSME more sensitive than NASA-TLX [16, p33]	❖ Still to be experimentally proven by research

Table 5.2: Summary of Subjective Workload Assessment Techniques

<table>
<tr><th colspan="2">Type of Scales</th><th colspan="2">Scales according to Dimension</th></tr>
<tr><td>SIMPLE SCALE</td><td>Synthetic assessment of Workload. Global Evaluation.</td><td rowspan="2">One-Dimensional Scales
Assessing Global WL
Not time-consuming
OW validated in the aeronautic, often used in automotive
RSME used in automotive, not yet validated

Most Reliable: OW, RSME
Main issue: WL Comprehension</td><td rowspan="2">Multi-Dimensional Scales
Assessing Different Dimensions contribute to WL
Allow to find causes
More time-consuming (Administration and Analysis)

Most Reliable: NASA-TLX</td></tr>
<tr><td>DECISION TREE</td><td>Questions are classified: subsequent questions may vary according to previous answers. Answers are of type YES/NO</td></tr>
<tr><td>MULTIDIMENSIONAL SCALES</td><td>Assess different components/factors contributing to WL. Most frequently used because it is more precise. The main advantage is the possibility to identify specific factors contributing to WL, but not the causes</td><td></td><td></td></tr>
</table>

Table 5.3: Explanation and Use of different Types of Scales

5.1.2. Self-reported measures of driving Performance

5.1.2.1. Driver Behaviour Questionnaire

The *Manchester Driver Behaviour Questionnaire* (DBQ) categorises driver behaviour in terms of errors, lapses, and violations and has been found to be a good predictor of crash involvement [62, 71,78]. The DBQ was developed to explore the role of human error in causing crashes. Human error is comprised of three distinct categories: lapses or slips (inadvertent or inappropriate occurrences of highly practised behaviours), errors or mistakes (errors of omission or commission resulting from a lack of knowledge or information) and violations (intentional actions in violation of rules or established practice) [57, 68, 70]. The DBQ contains questions about a driver's propensity to commit lapses, errors, and violations. Comparison of drivers' DBQ answers to their crash histories has shown that the violations score on the DBQ, particularly items classified as aggressive violations, is a good predictor of accident involvement [12p6,62,74].

5.1.2.2. Driving Quality Scale (DQS)

This scale is a self-reporting measure of how well the subject thinks he has driven. It ranges from –100 (I drove extremely bad) to +100 (I drove extremely well). Subjects are asked to answer the question "How well did you drive during the trial, compared to normal?" [14]

5.1.3. Expert-reported measures of driving Performance

5.1.3.1. TRIP

TRIP (Test Ride for Investigating Practical Fitness to Drive) is a tool to identify what aspects of driving are a concern to the driver. Different sections are present, a scoring system determines the Practical Fitness to Drive.
Scoring modes are:

- Insufficient, Doubtful, Sufficient, Good, which have points from 1 to 4;
- Simple scoring ([lower limit] 1,2,3,2,1 [upper limit]), where low score is worse, along with verbal descriptor of ranges limit.

Each section is provided with a blank space, in which to make annotations when the scoring is Insufficient (I).
Expert Judgement is the decision of what measures are to be taken to make the subject able to drive correctly (hours of drive, technical adaptation, restraint in time/space) or whether the Subjects is unfitting to drive, even after a number of followed lessons.

For the detailed prospect, see HASTE *Development of Experimental Protocol* (p 109) [15].

5.1.3.2. Wiener Fahrprobe

The method was originally designed by Risser [73], and developed to study learning drivers, but it can also be used to study driver behaviour in real traffic. The study is carried out with two observers in the car, studying the driver along a specified test route. One of the observers registers standardised variables such as speed adaptation at junctions/obstacles, lane change, interaction with other road users etc. The other observer does free observations, i.e. conflicts, communication, interaction and special events [27].
A representation of Wiener Fahrprobe, as well as a more detailed description, can be found in AIDE Deliverable 2.2.1.

5.2. Physiological Measures

This kind of measures do not require an overt response by the driver, and most cognitive tasks do not require over behaviour. Monitoring can be done continuously thanks to little intrusiveness due to miniaturization. Physiological Measures are influenced by many factors, like physical load, personality and emotions. It is therefore difficult to execute such kind of measurements, even because data analysis is complex.
Below is a short description of some Physiological Measures.

5.2.1. Heart Rate Variability

In general, this metric is more sensitive to increases in WL. This measure may be sensitive to measures of attentional demand in driving experience [16]. Conflicting results during test have been found, which led to conclude that this measure is sensitive to a major factor that is tiredness / physical effort, so data is too difficult to interpret; moreover, speech may influence blood pressure and therefore modify heart rate variability [52, 76].

5.2.2. Respiration Rate Variablity

Respiration rate increases under stressful attention conditions, and in increased memory load conditions or increased temporal demands. But this metric without information about tidal volume is meaningless and leads to inconclusive results; moreover, this measure is also sensitive to tiredness and physical effort. Collected Data is therefore difficult to interpret in real test condition, due to its sensitiveness to many variables. The measurement of such technique is also very intrusive [**13**].

5.2.3. Galvanic Skin Response

It corresponds to the determination of the resistance of the epidermal tissues of the human body to the flow of low-level electrical current, the variation of electrical behaviour of the skin. Such measure depends on emotional load, individual differences are very important, and results are difficult to interpret.

5.2.4. Muscle Tension

Some muscles (e.g. neck muscles and grip pressure) may increase in tension with the increase in WL. This measure is sensitive to tiredness and physical effort, but no many recent reports use this measure, perhaps due to its poor reliability in measurement and interpretation.

Measure	PROs	CONs
Heart Rate Variability	Sensitive to increases in Workload	Conflicting resluts dureing tests have been found; speech and tiredness or physical effort might influence blood pressure
Respiration Rate Variability	Sensible to stressful attention condition and in increased memory load conditions or increased temporal demands	Also sensitive to tiredness and physical effort; Might be meaningless without tidal volume (Wienties, 92,93)
Galvanic Skin Response		Depends on emotional load, and individual differences; results are difficult to interpret
Muscle Tension	Takes into account response of some muscles in precise cases	Poor reliability in measurement and interpretation

Table 5.4: PROs and CONs of Physiological Measures

5.3. Performance Measures

DeWaard [16] refers to three types of performance measures of workload: primary-task measures, secondary-task measures and reference tasks.
In this context we describe the primary-driving task performance and secondary task performance.
We have already mentioned secondary task, speaking of usability (see § 4.1.2.1). Secondary task workload measurement is a widely used approach to Workload measurement, in which the Subject is assigned a primary task and a secondary task simultaneously, while giving the primary task a higher priority. The primary task's workload is measured as the degradation in the secondary task performance under the dual task condition compared to a condition with only the secondary task. Human attentional resource is limited, so that concurrent tasks compete for the attentional resource and information processing capacity. The most widely used secondary tasks are mental mathematics, memory, tracking, reaction time, auditory detection, problem solving, monitoring, time estimation, random sequence generation, and classification. As already mentioned (see § 4.1.2.1), a secondary task is also the use of an infotainment system or another on-board system.

In Meister [50], some guidelines are given for selecting secondary tasks. The tasks should be:

1. Non-interfering with primary task;
2. Easy to learn;
3. Self-pacing: to allow the secondary task to be neglected if it is necessary to ensure the primary task performance;
4. Continuous scoring;
5. Compatible with the primary task;
6. Sensitive;
7. Representative.

Some tasks that meet the aforementioned criteria include reaction time, mental arithmetic, self-adaptive tracking, and monitoring.
Two paradigms for dual-task performance are possible:

- Loading Task Paradigm: *secondary*-task performance is maintained, even if decrements in primary-task performance occur. The addition of the second task results in a total workload shift from region A towards region B (Meister [49] states a diagram where Region A shows highest levels of Performance against lowest levels of Workload, while Region B contain a decreasing level of Performance, against an increasing level of Workload) so that primary-task performance measures can be used as indicators of Workload.
- Subsidiary Task Paradigm: the instruction to maintain *primary*-task performance is given. Consequently secondary-task performance varies with difficulty and indicates 'spare capacity', provided that the secondary task is sufficiently demanding. Spare capacity [8] is a concept that is used frequently in dual task performance, and assumes a total undifferentiated capacity that is available to perform all tasks. In the case of unaffected single-task performance, the unused capacity is called spare capacity, and is in principle available for secondary-task performance.

The largest sensitivity in Secondary-Task Measures is achieved if the overlap in resources that are used is high; in order to perform the secondary task, spare capacity of the *same* resource should be required (Multiple-resource theory [88]). Time sharing is expected to be less efficient if the same resources are used. This large overlap in resources used is at the same time a threat to undisturbed Primary-Task performance because Primary Task intrusion is largest if two tasks that use the same resources have to be time-shared.

Other problems related to secondary task methodology [19]:

- Non-specific intrusion (e.g., peripheral interference),
- Omission of secondary-task performance in the case that primary-task demands are very high,
- Operators' resource allocation policy (the priority given to each task).

This resource allocation policy is important if the primary task has a high ecological validity. Also, the choice for a secondary task is more difficult in tasks approaching everyday performance.
The use of secondary tasks in applied environments is more complex than in laboratory experiments, and for this reason caution is required. Most frequently used as secondary tasks are choice reaction time tasks, time estimation or time-interval production, memory-search tasks and mental arithmetic (see [19, 60, 89], for overviews). Eggemeier & Wilson [19] have compared several multiple-task studies and conclude that results regarding sensitivity of the different measures are mixed.
Primary-task intrusion also differs between studies. They argue that both effects are related to a large diversity in workload levels, tasks and test environments.
This dual-task situation is widely used in automotive workload researches because it reflect the real driving condition of everyday life in which the driver has to maintain control over his vehicle and at the same time absorb and process information coming from peripheral systems (such as phones, navigation systems and entertainment -features) and interact with them [18].

PROs	CONs
• Secondary task workload measurement has a high degree of face validity. • Administering the secondary task to different primary tasks gives comparable workload measurement results (Wickens et al., 2000).	• Requirement of additional instrumentation, • Possible compromises to system safety (primary-task intrusion) • Lack of operator acceptance, partially resolvable with embedded task techniques.

Table 5.5: PROs and CONs of Secondary Measurments in Usability Assessment

Here, a set of metrics will be listed that can be used to evaluate both primary and secondary – task performance.
The metrics listed refer to a number of indicators particularly significant in order to evaluate in-vehicle workload. A brief description is given for each of them. For further details see AIDE deliverable 2.2.1 and the relevant RoadSense Deliverable [13].

Lateral Control

Number of major lane deviations	Any part of the vehicle exceeding either the central line or the roadside lane boundary by more than a half of the vehicle width
Steering wheel position variance	Position of the vehicle centre with respect to the centre of the lane
Standard deviation of steering wheel angle	SD of previous measurement
Standard deviation of lateral position	Variability of vehicle position with a lane
Behavioural entropy of steering wheel angle	Method for computing predictability of driver behaviour and for characterising WL
Steering wheel reversals rate	Number of times per minute (or covered kilometer), that the directionof steering wheel movement is reversed through a small, finite angle, or gap.
Yaw rate	Variation of vehicle angular velocity

Visual management

Time on road perception information inside the vehicle	Time spent looking at mirrors or systems giving visual road information
Time on driving information inside the vehicle	Time spent looking at the dashboard (speedometer…)
Time on any other areas	Time spent on all information except driving information and road perception information
Visual demand	Percentage of total time that one spends looking at an object or area such as the road or inside the vehicle
Decrease in glance frequency to mirror	Sensitive to WL variations. Decrease means lower WL

Longitudinal control

Mean speed	Mean speed during in-vehicle task accomplishment
Variance in longitudinal speed	Variance of speed when subjects have to maintain a given speed

Interaction with other vehicles

Time headway	Time for the host vehicle to arrive in place of leading vehicle without changing speed
Relative distance with other lateral vehicles	Distance between host vehicle and other vehicles
Following distance	Corresponds to headway in car following task
Duration of close following situations	Situations where the host vehicle follows a lead vehicle with closed velocities and short following distance (lower than safety distance)
Number of lane changing	Reflects an actual adaptation of the driver to increased WL

In the left column, the name of the metric; In the right column, a short description of the metric

Table 5.6 : Metrics for Performance Measures

6. Situation Awareness

In this chapter, an overview of Situation Awareness is given. For detailed information, please see AIDE deliverable 2.2.1, as well as [20, 61 pp163-178].

A general definition of Situation Awareness (SA) is as follows: [the state to which a person arrives through the process of] *the perception of the elements in the environment within a volume of time and space, the comprehension of their meaning and the projection of their status in the near future* [20,p97]. Generally speaking, this entails the clear understanding of what is going on in the environment and what is going to happen in the nearest future. SA has been found interesting and therefore studied in aviation, where it was important to understand what pilots perceived and how they were going to react at once. It has been successfully applied to other scopes, among which the automotive scope.

SA can be divided in three levels:

1) Perception of cues, where people receive the stimuli to create a clear and precise mental picture of the environment. If this phase is executed wrongly, due to lack of information or difficult cognitive process, there may be an error at further steps;
2) Comprehension of gathered stimuli, where people put all information pieces together and rank them according to their task goals; the meaning of information must be considered according the situation in which stimuli are collected (objective aspects), and according to subjective interpretation, the awareness;
3) Projection, that is the ability to forecast future situation events and dynamics, in order to execute Decision Making process on-time. Such phase is the one experts rely on mainly.

SA can be assessed in terms of Self-reported measures, and in terms of Performance Measures.

6.1. Self-reported measures

Assessing SA can be done by means of Probe Questions, which are asked to users with no pre-information, in order to get accurate measurements of SA for specific aspects. The issue of this technique is that after the first probe question, users might modify their behaviour in the expectance of other probe questions, thus cancelling spontaneousness. Another issue is the fact sometimes users are asked to stop the task they were running, to answer questions. This might lead to incoherent behaviour, thus giving inconsistent results. Moreover, asking questions after the test had completed might be too late to test specific aspects of the system or events, due to lack of Subject's memory of events that might have been registered subconsciously and therefore not remembered (lack of SA) or simply too technical and precise that the Subject does not remember them at the end of the test, but when they took place they were inside Subject' SA. Another issue is the possibility that Subjects give a self-judgement about their SA according to their Performance, saying they had good SA when Performance was high, or saying they had poor SA when Performance was poor, without minding what causes brought to poor Performance. [20, pp94,95].
Self-reported measures give an indication of self-awareness, not of what the Subjects should be aware of. In this sense, such measurements should be aided by other monitorings, e.g. Performance Measures.

To solve these problems, great care must be used when projecting assessment scales for SA, in order to split SA components in terms that are easily administrable by users and that allow to isolate the issues above mentioned.
We will discuss about SART and SAGAT as examples of Self-Report Measures.

6.1.1. SART

Situation Awareness Rating Technique (SART) is one of the best known and spreadly used scales to assess SA.
Characteristics of understanding by the users are:

- understanding of situations in making decisions;
- understanding is available to consciousness;
- understanding can readily be made explicit and quantifiable.

Two different scales are available with SART to assess SA:
- A 10-Dimensions Scale, each dimension ranging from 1 to 7;
- A synthetic scale, 3D-SART, which groups dimensions into three Domains. Each Domain is evaluated independently on a 100-mm scale. According to this second type, SA happens to be *SA = Understanding – (Demand – Supply)*. This formula is done on theoretical considerations rather than empirically or statistically. For this reason, such value can be used to compare systems together, but not as an absolute measure of SA.

PROs	CONs
• Measures directly derived from users (ecological validity); • General Construct allow scale to be used for different scopes other than aviation (i.e. automotive); • A certain level of Diagnosticity is provided. • Since SART takes into account Supply and Demand of attentional resources (generally considered WL constructs), it should provide some measure of how changes in WL affect SA [20 pp117-122]	• As a subjective measure, SART data should be interpreted along with performance data, because user' self-assessment might be inexact; • Caution must be used when using SART to measure WL, as they are different concepts • It is not clear whether the 10-dimensions division into three domains is sufficient and necessary [20 pp117-122]

Table 6.1: PROs and CONs of SART

6.1.2. SAGAT

The Situation Awareness General Assessment Technique (SAGAT) is perhaps the most used technique to assess SA. The technique is as follows:

During a simulation with a System being assessed,

- Simulation is frozen at random times
- Displays are blanked during SA assessment;
- Subjects are queried to describe their perception of the situation at that moment.

A set of queries is prepared in advance, and they are presented randomly to Subjects, with no particular notification.

Most important is the preparation of queries: they should be as similar as possible to Subject's way to think, in order to eliminate the need to elaborate the question to answer. Queries are usually determined by following a goal-directed task-analysis [20 p148].

Experimental results in aviation led to conclude that SAGAT has good sensitivity and reliability, intrusiveness due to freezing never influenced performance.

For safety reasons, SAGAT may not be applicable to all domains, especially real on-field applications. Two subjects should be used, in order to maintain safety: while the first is tested, the second takes control of system, to keep safety at a maximum. On automotive fields, it would mean a driver should stop the car and answer (with eyes and ears close) SAGAT questions, and then resume normal driving. Normal stopping situations, such as stopping at traffic lights, or anyway at low WL levels, should be advisable to probe Subjects. Sometimes recordings of scenes are used, and subjects are asked to view the replays. During the reproduction of the registration, normal SAGAT protocol was followed. This technique allowed to get some useful data about SA in some experiments [20].

PROs	CONs
• Subjective Measure • Good Sensitivity and Reliability • It is possible to query subjects during test, thus providing sound answers when the mental state of the Subject is still intact • Not proven to be intrusive, so other measurements can take place during test along with SAGAT	• Subjective Measure • Queries must be carefully identified, following a rigid protocol; this procedure might be expensive and time-consuming • For safety reasons, difficult, but not impossible, to administer in real traffic

Table 6.2: PROs and CONs of SAGAT

6.2. Performance measures

Performance Measures, opposite to Self-reported measures, rely only on Subject's interaction with the system. They are any measurement that infers subjects' SA from their observable actions or the effects these actions ultimately have on System Performance [20]. This allows experimenters to compare the desired and achieved performance of a System, and its weak points, that is those which report low SA. PM rely on Decision Making phase, and measure how well the Subject is able to react to variations in the environment with modifications in his behaviour. Performance measures allow to identify:

- The final Performance of Subjects;
- The related considerations in decision making and control actuation that require SA;
- The sufficiency of the operator' SA, especially in the presence of factors such as time pressure and uncertainty.

Performance based measures can assess the requirements placed on SA by the decision and control actuation strategies used by the operator. They can examine, in the context of the test situation, the relative impact of SA and other potential blocks to satisfactory performance [20].

Four types of Performance-based measures exist:

1. Global measures, which rely on the task accomplishment by subjects. This approach is criticized because good performance not necessarily means good SA and vice-versa;
2. Imbedded Task Measures, which consider specific measurements for the system being tested, such as deviation from reference values in steering wheel;

3. External Task measures, which examine subject reactions to changes to, or removal of, information relevant to the task at hand;
4. Testable responses, or Implicit measures, which serve to eliminate the ambiguity of Global Measures: they act in extremely controlled experimental conditions, and create a situation whose outcome is a set of pre-chosen actions: which action is taken is necessarily an index of SA: if the action is correct, then SA is good, because such action could be taken only if SA was good; if the action is wrong, then SA is poor, because correct actions were missed due to lack of SA. Those situations are not likely to block due to poor decision making, as experimentally proven.

Different measurements can be done at once, for they are not conflicting to each other.

7. Conclusions

In this Review, several techniques have been taken into account to evaluate IVIS / ADAS, according to complex and various aspects of Usability, Acceptance, Subjective Workload, and Situation Awareness. When installing such Systems, it is important to understand which benefits they bring to drivers, but even more important it is to clearly understand how much resources they require to drivers. Safety is the primary and fixed point to be achieved whatever other condition might take place.

- Usable Systems meet user requirements, needs, expectations. The most effective way to build Usable Systems is to design them according to Users, so that it is clear the User Mental Model. All interactions with the System, when carefully reflecting User Model, will be familiar and prone to few errors.
  - ❖ Focus Groups and Interviews are the best ways to understand and create the User's Mental Model.
- The communication between user and system is of primary importance: if the user cannot understand system's output, or the input is not intuitive, poor use can be made of the system in safe conditions when driving.
  - ❖ Card Sorting helps understand the correct terminology and visual metaphors to use.
  - ❖ To understand whether functions can be activated through a simple path, whose meaning is clear in the user's mind, Decision Trees help establish the flow of information important to place the right functions in the right place.
- When a simple interface is ready, it can be tested, either by experts and users.
  - ❖ Heuristic Evaluations and Checklist help find common Usability issues, by identifying common interaction problems or interface lack of functionalities.
  - ❖ Guidelines are used to build such tools, and must be done by skilled experts.
- Users are of great importance: they give a precise idea of how performing the system will be.
  - ❖ In preliminary phases of development, when a prototype does not exist yet, Observation is the most indicated technique: users interact with the interface (even on mock-ups, paper and so forth), and experts watch them.
  - ❖ The Number of errors and Total Task Time are early indicators of Usability issues. If users are often prone to make the same mistakes, then Design has failed somewhere. Before going on in development, it is important to correct such issues.
  - ❖ What the User thinks he can do with the system seldom is what the user actually can do with the System. For this reason, Thinking-Aloud and Co-Discovery Learning help understand mental processes and concrete short-term expectancies (I have to do this, I think I should do it this way).
  - ❖ When there exists a concrete interface, and the development is at a more advanced state, Questionnaires can be submitted to users, to probe their opinion about the system. Questionnaires are broadly used in IVIS / ADAS assessment, and though they are a very subjective source of data, they often are the only way to probe some aspects, for example Acceptance.
- Only users can say how much they like the system, and whether they are willing to pay to have the system installed in-car. The Semantic Differential Technique helps understand how useful, pleasant, effective, desirable… a system appears to users. This aspect is important for manufacturers, who use this information to decide whether to produce such system or not.

These techniques give good possibilities to create an usable product, familiar, easy to understand and learn. But as such system should be used while driving, it must be seen whether it interferes with safety.

- In order to do so, Workload has to be taken under control. Workload can be assessed either by Subjective means and by analytic means. Both approaches are important, as they reveal different aspects of Workload, and as such this data must be analysed together. For example, Subjective measures can capture how much effort a user has to use to maintain control of the vehicle;
  - ❖ Performance measures can express how much the use of the system deviate performance from that achieved with driving without the system.
  - ❖ Subjective Measures are taken by means of standard scales, where users are asked to rank different aspects of Workload from a minimum to a maximum. Aggregate results are done post-test, and they express globally how much users are loaded when interacting with the new system.
  - ❖ Several Observational Grids have been developed, in order to check whether users are prone to violate road laws when relying too much on a system, or to understand which tools would have introduced driving enhancements (e.g. users who drive too fast would be safer with an ISA-equipped car).

- Last but not least, the human perception, comprehension, and projection of the surrounding environment, the Situation Awareness, and its study, is an important component of IVIS / ADAS assessment. Some systems might lead to driver "deactivation" because relieve the driver from too much aspects of driving, thus paradoxically decrease safety. When users have high degrees of SA, they are likely to have less incidents.

The development of Usable and Safe IVIS / ADAS is therefore long and complex. It is not always relying on standard procedures is the best solution, and empirical experts opinions might be too long and expensive. It is important to keep clearly in mind what aspects must be verified, but those who verify them must have the right competencies.

- A procedure to assess such systems has been developed, along with a series of aspects to be taken into account. It can be the starting point for developing checklists, or used as is when identifying test requirements. The procedure has been developed bottom-up, based on existing tests done by European partners. In the appendices, several grids, whose structure is nearly the same, shortly but efficacely describe all important aspects of testing. All important aspects have been used to create the procedure, that has then been integrated with important topics which are not clear in many experiments.

8. References

1. AIDE – Annex1 –"Description of Work" 2003.
2. Bauer, A.(BASt), Tango F. (CRF) ADVISORS: An Evaluation of Lateral Support System (LSS),
3. Bellotti, F., De Gloria, A. (DIBE), (2000) State Of The Art Of Driving Support Systems And On-Vehicle Multimedia HMI, COMUNICAR Deliverable 2.1
4. Bekiaris, E. , Portouli, (2001) E. IN-ARTE Project TR 4014 Del. 7.4: System evaluation and impacts on traffic safety
5. Bengler, K., Praxenthaler, M., Theofanou, D., Eckstein, L., (2002) Investigation of Visual Demand in Different Driving Simulators within the ADAM project.
6. Bias, R.G. Software Usability Engineering (slides).
7. Brooke, J., SUS. A "quick and dirty" usability scale http://www.cee.hw.ac.uk/~ph/sus.html.
8. Brown, I.D. & Poulton, E.C. (1961). Measuring the spare 'mental' capacity of cardrivers by a subsidiary task. Ergonomics, 4, 35-40.
9. Bűscher S., Thorsten Frese (1998). User Needs Survey. IN-ARTE Project TR 4014.
10. Campbell et al, (2004), Comprehension Testing of Active Safety Symbols
11. Centre for International Economics Camberra & Sydney (2001) Review of Wiilingness to Pay Methodologies, http://www.ipart.nsw.gov.au/pdf/CIE.pdf
12. Charlton, S.G., and Baas P.H., Road User Interactions: Patterns of Road Use and Perceptions of Driving Risk, Transport Engineering Research New Zealand Ltd.
13. Chin, E., Nathan F. (2004). Roadsense D 2.1 part 1 Final version – state of the art on HMI metrics and target values.
14. COMUNICAR From Deliverable: Human Factor Tests on car Demonstrator – The Methodology
15. DeWaard, D., Brookhuis, K. et al; HASTE: Human Machine Interface And the Safety of Traffic in Europe – Development of Experimental Protocol
16. De Waard, D., (1996) The Measurement of Drivers' Mental Workload, Ph.D. Thesis, ISBN 90-6807-308-7
17. Dingus, T.A., Williges, R.C. Casali, J.G., Kleiner, B.M, Kiefer, R.J, The Use Of Speech Recognition Technology In Automotive Applications
18. Donk, V. (2004). Time Design: Workload Management in Automotive Situations. Workshop Time -Design CHI 2004.
19. Eggemeier, F.T. & Wilson, G.F. (1991). Performance-based and subjective assessment of workload in multi-task environments. In D.L. Damos (Ed.), Multiple-task performance. (pp. 217-278). London: Taylor & Francis.

20. Endsley, M.R., Garland, D.J. (2000) Situation Awareness Analysis and Measurement, Lawrence Erlbaum Associates, Publishers, London
21. European Statement of Principles on Human Machine Interface for In-Vehicle Information and Communication Systems, EC 5/98.
22. European Statement of Principles on Human Machine Interface for In-Vehicle Information and Communication Systems, Expansion of Principles, EC 11/98.
23. Green, P., Paelke, G. and Boreczky, J. The "Potato Head" method for identifying driver preferences for vehicle controls, UMTRI Human Factors Division
24. Guion, L.A. (2001), Conducting an In-Depth Interview
25. Gunn, R. et al, (2001) The In-Depth Interview
26. Hart, S.G. & Staveland, L. (1987). Development of a multi-dimensional workload rating scale: Results of empirical and theoretical research. In P.A. Hancock & N. Meshkati (Eds.), Human Mental Workload. Amsterdam, The Netherlands: Elsevier
27. Hjälmdahl, M. Department of Technology and Society, Lund University, Results From In-Car Observations In The Large Scale Trial With Active Gas Pedal In Lund, Sweden, , ICTCT workshop Nagoya [http://www.ictct.org/workshops/02-Nagoya/Hjalmdahl.pdf].
28. ISO13407: 1999 Human-centred design processes for interactive systems
29. ISO/IEC 9241-11: 1998 Guidance on Usability.
30. ISO 15005:2002 Road vehicles -- Ergonomic aspects of transport information and control systems -- Dialogue management principles and compliance procedures
31. ISO 15007-1:2002 Road vehicles -- Measurement of driver visual behaviour with respect to transport information and control systems -- Part 1: Definitions and parameters
32. ISO/TS 15007-2:2001 Road vehicles -- Measurement of driver visual behaviour with respect to transport information and control systems -- Part 2: Equipment and procedures
33. ISO 15008:2003 Road vehicles -- Ergonomic aspects of transport information and control systems -- Specifications and compliance procedures for in-vehicle visual presentation
34. ISO 17287:2003 Road vehicles -- Ergonomic aspects of transport information and control systems -- Procedure for assessing suitability for use while driving
35. ISO/TS 16951:2004 Road vehicles -- Ergonomic aspects of transport information and control systems (TICS) -- Procedures for determining priority of on-board messages presented to drivers
36. ISO 15006 Road vehicles -- Ergonomic aspects of transport information and control systems -- Specifications and compliance procedures for in-vehicle auditory presentation
37. ISO/CD 16673 Road vehicles -- Ergonomic aspects of transport information and control systems -- Occlusion method to assess visual distraction due to the use of in-vehicle information and communication systems

38. ISO/CD TR 16352 Road vehicles -- Ergonomic aspects of in-vehicle presentation for transport information and control systems -- Warning systems ISO/CD TR 16352 Road vehicles -- Ergonomic aspects of in-vehicle presentation for transport information and control systems -- Warning systems
39. Jia Shen, XiaoJian Shen, "User Requirements in Mobile Systems", in Proceedings of the 2001 Americas Conference on Information Systems, page 1341-1344. (Boston, August 2-5, 2001).
40. Kieras, D.. Task Analysis and the Design of Functionality. In A. Tucker (Ed.) The Computer Science and Engineering Handbook, CRC Press Inc, 1997.
41. Kopf, M. (BMW), Allen, P (TRL), Becker, S. and Cieler, S (TUV), Dilger, E. and Krautter, W. (BOSCH), (1999) RESPONSE Del. 4.2. Checklist for Theoretical Assessment of Advanced Driver Assistance Systems: Methods, Results and Assessments of Applicability
42. Kirwan, B., and L. K. Ainsworth (1992). A Guide to Task Analysis. Taylor & Francis Inc..
43. Llaneras, R.E., and Singer, J.P., December 15, 2002, In-Vehicle Navigation Systems: Interface Characteristics and Industry Trends, Presented to Driving Assessment 2003, 2nd International Driving Symposium on Human Factors in Driver Assessment, Training and Vehicle Design
44. Luximon, A., Goonetilleke, R.S., (2001), Simplified Subjective Workload Assessment Technique, in Ergonomics ISSN 0014-0139
45. Mariani, M., Veltri, G. (UNISI), Montanari, R. (CRF), Peterson, D. Karlsson, B., (VOLVO), Amditis, A. (ICCS). (2000) COMUNICAR: Estabilishment of User Needs, Deliverable 2.2
46. Mattes, S. The Lane-Change-Task as a Tool for driver Distraction Evaluation, DaimlerChrysler AG, Research – Human-Machine-Interaction, Stuttgart, Germany
47. Martinetto, M. (JRC), CRF S.C.p.A. EUCLIDE: Enhanced human machine interface for on-vehicle integrated driving support system
48. Mauer, D., Warfel, T. Card sorting: a definitive guide, http://www.boxesandarrows.com/archives/card_sorting_a_definitive_guide.php.
49. Meister, D. (1976). Behavioral foundations of system development. New York: Wiley.
50. Meister, D. (1986). A survey of test and evaluation methods. Proceedings of the Human Factors Society 30th Annual Meeting, 123 9 - 1243.
51. Morgan, D., Krueger, R. (1998) The Focus Group Kit, Vol.1, Sage Publications
52. Mulder, L.J.M. (1988). Assessment of cardiovascular reactivity by means of spectral analysis. PhD Thesis. Groningen: University of Groningen.
53. NHTSA (April 2001 - March 2002) Crash Avoidance Metrics Partnership Annual Report
54. Nielsen, J., Mack, R.L, (1994) Usability Inspection Methods, John Wiley & Sons
55. Nilsson, L. and Harms, L. ADVISORS: Definition of assessment parameters: Road v. Simulator issues
56. Nodari, E., Toffetti, A., Zoldan, C. (2001) SENECA Speech control modules for Entertainment, Navigation and communication Equipment in Cars

57. Norman, D.A. (1981). Categorization of action slips. Psychological Review, 88, 1-15.
58. Nowakowsky et al (2002) An Experimental Evaluation of Using Automotive HUDs to reduce Driver Distraction While Answering Cell Phones, Proceeding of Human Factors And Erogonmics Society, 46th Annual Meeting, 2002
59. Nowakowsky, C., Green, P. Tsimhoni, O. Common Automotive Navigation System Usability Problems and a Standard Test Protocol to Identify Them
60. O'Donnell, R.D. & Eggemeier, F.T. (1986). Workload assessment methodology. In K.R. Boff, L. Kaufman & J.P. Thomas (Eds.), Handbook of perception and human performance. Volume II,cognitive processes and performance. (pp 42/1-42/49). New York: Wiley.
61. O'Brien, T.G., Charlton, S.G., (1996) Handbook of Human Factors Testing and Evaluation, Lawrence Erlbaum Associates
62. Parker, D., Reason, J., Manstead A.S.R., Stradling, S.G., (1995) Driving Errors, Driving violations and accidents involvement. Ergonomics, 38, 1036-1048
63. Pauzié, A., Pachiaudi, G. , (1996) Subjective Evaluation of the Mental Workload in the Driving Context, Laboratory Ergonomics Health Comfort, INRETS / LESCO, Interational Conference on Traffic and Trasport Psychology
64. Pedon, A., Metodologia per le Scienze del Comportamento, Ed. Il Mulino.
65. Peterson, D. Piamonte, P. (Volvo), Gelau, C. (BASt), van Winsum, W., Hoedemaeker, M. (TNO), Dangelmaier, M. (IAO), Hess, M., Kuhn, F. (DC), Mariani, M. (UNISI), (2000) COMUNICAR Validation Plan, Deliverable 6.1
66. PSA Peugeot Citroen Subjective Assessment Methods; PSA – Proposition of a new Method (slides).
67. Rambaldini, A., Alessandretti, G., Irato, G., Nodari, E., Toffetti, A., Zoldan, C. (Centro Ricerche FIAT S.c.p.a, Italy) Motetti, P., FIAT Auto S.p.a., Italy, (2003), Iterative Design Of A New On-Board System For Public Vehicle: From The Idea To The Prototype
68. Rasmussen, J. (1982). Human errors: A taxonomy for describing human malfunction in industrial installations. Journal of Occupational Accidents, 4, 311-333.
69. Rattazzi, Il Questionario, Cleup Editrice Padova, 1990.
70. Reason, J. (1990) Human Error, Cambridge University Press, Cambridge
71. Reason, J.T., Manstead, A.S.R., Stradling, S.G., Baxter, J.S. and Campbell, K.A. (1990) Errors and violations on the roads: A real distinction? Ergonomics, 33, 1315-1335.
72. Reid, G. B. And Nygren, T. E. (1988), The subjective workload assessment technique: a scaling procedure for measuring mental workload. In P. A. Hancock and N. Meshkati (eds), Human Mental Workload (Amsterdam: North-Holland), 185 - 218.
73. Risser, R. (1985) Behaviour in traffic conflict situations. Accident Analysis and Prevention Vol. 17, No. 2, pp 179-197.

74. Rothengatter, T. (1997). Errors and violations as factors in accident causation. In T. Rothengatter and E. Carbonell Vaya (Eds.) Traffic and Transport Psychology: Theory and Application. Amsterdam: Pergamon.
75. SAE Recommended Practice - Navigation and Route Guidance Function Accessibility While Driving (SAE J2364)
76. Sirevaag, E.J., Kramer, A.F., Wickens, C.D., Reisweber, M., Strayer, D.L. & Grenell, J.F. (1993). Assessment of pilot performance and mental workload in rotary wing aircraft. Ergonomics, 36, 1121-1140.
77. Statement of Principles, Criteria and Verification Procedures on Driver Interactions with Advanced In-Vehicle Information and Communication Systems, Driver Focus-Telematics Working Group (2003)
78. Stradling, S. G., and Meadows, M. L. (2000). Highway Code and Aggressive Violations in UK Drivers. Proceedings of Aggressive Driving Issues Conference. Ontario Ministry of Transportation.
79. The psychology of error, http://www.humanfactorsmd.com/hfandmedicine_reducerror_nature.html
80. TRL. VTI, RUG, CRF, ADVISORS: Traffic Safety Assessment
81. UMTRI website, Usability Guidelines, http://www.umich.edu/~driving/guidelines
82. Usability Evaluation Methods and Guidelines, www.pages.drexel.edu/~zwz22/UsabilityHome.htm
83. User/ Observation field studies, http://www.usabilitynet.org/tools/userobservation.htm
84. Usability Evaluation Methods and Guidelines, http://www.usabilitynet.org
85. Van Der Laan et al, (1997) A Simple Procedure for the assessment of Acceptance of Advanced Transport Telematics
86. Várhelyi A. et al Department of Technology and Society, Lund University A Large Scale Trial With Intelligent Speed Adaptation In Lund, Sweden
87. Vidulich, Ward and Schueren (1991), Using the Subjective Workload Dominance (SWORD) technique for projective workload assessment in Human Factors, 33(6), 677-692
88. Wickens, C.D. (1984). Processing resources in attention. In R. Parasuraman and D.R. Davies (Eds.). Varieties of attention. (pp.63-102). London: Academic Press.
89. Wickens, C.D. (1992). Engineering psychology and human performance. New York: HarperCollins.
90. Wierwille, W.W. & Casali, J.G. (1983). A validated rating scale for global mental workload measurement application. In Proceedings of the Human Factors Society 27th Annual Meeting (pp. 129-133). Santa Monica, CA: Human Factors Society.
91. Willingness to Pay http://www.pitt.edu/~super1/lecture/lec11871/006.htm
92. Wise, J.A. Hopkin, J.A. (2000), Human Factors in Certification, Lawrence Erlbaum Associates

93. Zhang, X., Kompfner, P., White, C., ERTICO (1998) CONVERGE: Guidebook for Assessment of Transport Telematics Applications: Updated Version, Project Co-ordinator Paul Kompfner.
94. Zijstra & Van Doorn, (1985), The construction of a scale to measure perceived effort. Department of Philosophy and Social Sciences, Delft University of Technology.

9. Appendices

9.1. Case Studies

This section contains a list of Case Studies taken from existing literature. The studies listed here regard only Usability and acceptance aspects. In the 2.2.1 deliverable it will be possible to find a description of Workload and Situational Awareness researches. For each Project, most important information has been captured, and according to a bottom-up approach, a common experimental descripting grid has been identified. The common grid has been applied to Case Studies, in order to represent them in a standardized view. The Grid is based on the Step-by-step Procedure (see § 9.2).

Comunicar: From Deliverable: Human factor tests on car demonstrator – The Methodology [14]

This deliverable explains the realisation of an ADAS System Test. Below the phases of the test are summarized.

Purpose	Testing of in-vehicle ADAS system *Information Manager (**IM**)*
Technology used	ADAS, I-HMI with haptic knob, vocal system and two LCD displays
Place of Testing	Italy, Sweden
Type of Driving test	Real Traffic
Subjects	Two groups of 32 drivers; 16 male and 16 female for Italian Testing, 32 male for Swedish Testing;
Requirements for Subjects	Age 25-50 years, Owners of a valid Driving Licence for at least 10 years
Experimental Design	Within-Subjects
Test Modality	Half subjects began the test with IM on, half with IM off. Subjects were not informed about the status of IM; each subject had to do the test twice, once with IM on, once with IM off.
Company of Subjects	Test Leader
Actions of Accompaniment Tester	Instruct subjects what tasks to perform and when, trigger incoming information (only Italian study) and make notes
Test Complexity Levels	High/Medium/Low Traffic Situations (Indipendent Variables)
Test Complexity Scenarios	Turning manoeuvres in inner-city traffic for High Traffic; Passing a cloverleaf for Medium Traffic, and Driving on an highway for Low Traffic
Tasks to be performed	Reading of incoming information activated by the system or test leader, displayed to the subject (navigation, mail, SMS, phone); driver-initiated activities under request of the test leader (e.g. changing radio tuning)
Tasks scheduling	Predefined landmarks over the track
Feedback from Subject	Short vocal comment after reading the displayed information
Aspects to be tested	Workload, Acceptance, Usability, Driving Performance/Safety
Indicators	See table below for complete listing of indicators
Data Collection methodology*	Indicators for every Aspect, Type of Indicator and Measuring means

Data collection typology	PDT recorded continuously during the test drive in a log file along with measures of driving performance; RSME and DQS rating collected once with IM on and once with IM off; subjective measures (Acceptance and Usability) taken once by interviews at the end of the session
Time of Testing	Day hours, with care to avoid rush hours. Time of testing 150', divided in Car Familiarisation (~20'), IM on driving session (~60'), IM off driving session (~60'), Interview and debriefing (~10')
Interview target	RSME (Rating Scale of Mental Effort)
Contents of Briefing	Description of the System to be used in the test Telematic services (navigation system, traffic information, messaging, Internet, mobile phone), ADAS (FCW and LDWS); "Intelligent Filter" (managed by the system) to decide which messages are to be postponed according to the present driving conditions Description of the test itself.

- See next table for detailed information about Data Collection methodology

Criteria	Indicator	Type; Measuring
Workload	• Missed PDT signals • Average value of PDT reaction times	Objective; log file
	• Average value of RSME ratings	Subjective ; questionnaire
Acceptance	• Standardised attitude scale (van der Laan et al., 1997)	Subjective ; questionnaire
Usability	• Specific usability questions	Subjective, questionnaire
Driving performanc e, Safety	• Average vehicle speed • Standard deviation of vehicle speed • Standard deviation of yaw rate • Maximum of yaw rate • Standard deviation of steering angle • Average distance to the left and right border of the lane • Standard deviation of distance to the left and right border of the lane • Frequency of lane warnings • Maximum of risk level (risk level estimated by the IM) • Maximum duration of risk levels	Objective; log file
	• Average value of DQS ratings	Subjective, questionnaire

The following table is a description of the various test phases :

#	Test Phase	Description
1	Introduction (5')	Brief explanation about the objectives of the study and description of the phases and the structure of the COMUNICAR field test
2	Briefing on the COMUNICAR System and experience with the system (10')	Description of the COMUNICAR System, the Information Manager and the main functions included on the vehicle demonstrator. Description of the I-HMI and explanation on how to operate it Familiarisation with guided tasks (same tasks that are asked during the test drive): ❖ to activate the navigation system from the address book ❖ to perform a Telephone Call ❖ to refuse a Telephone Call ❖ to switch on and off the Radio
3	Experience with the automatic gearshift (5')	Familiarisation with the automatic gearshift of the vehicle demonstrator
4	Experience with the PDT (5')	Explanation of the PDT and how to handle the PDT device
5	Field Test (120')	1. First test track (40 – 50') 2. Completion of Rating Scale Mental Effort and Driving Quality Scale (5') 3. Second test track (40 – 50') 4. Completion of Rating Scale Mental Effort and Driving Quality Scale (5') 5. Post – Test Interview (5') 6. Post –Test Questionnaire (10')
6	Debriefing (5')	Clarifying any questions and gathering eventual opinions concerning the COMUNICAR System and/or the COMUNICAR test. Thank to the subject.

Table 9.1: Description of LSS Test Phases

Advisors: An Evaluation Study of the Lateral Support System [2]
A. Bauer (BASt); F. Tango (CRF)

This Report, developed by CRF, has the aim to offer a case study relative to LSS (Lateral Control Support), which has two main operative modalities: *Lane Warning (LW)* and *Lane Change Support (LCS)-Blind-Spot (BS)*

The benefits of LSS concern two main aspects:
- The reduction of lateral vehicle Collision and
- The improvement of driver safety and comfort.

Purpose	Testing of LSS: LW, LCS, BS
Technology used	Lancia K 3.0 V6 equipped with • 3 Cameras (1 frontal for LW, and 2 lateral for BS); • 3 ECU for the image processing of these video-sensors; • 1 Camera for collecting driver's images • CAN-C bus (chosen as the protocol communication for the data exchange and for the communication among all the components); • 1 PC used for Data-Logging and strategies implementation (with all the modifications necessary for the tests inside the ADVISORS project); • Devices for HMI (towards the driver). • Haptic warning at steering wheel level • Two beepers, one for each car' side
Place and characteristics of Testing	Highways and railroads in the surroundings of Turin, 2-3 Lanes for each carriage, with the same width foreseen by Italian law; Lane Marking: White lanes – dashed or continue.
Type of Driving Test	Real Roads
Subjects	24 Subjects, 12 Male, 12 Female
Requirements for Subjects	Age: 25-38 years old, at least 5000 km/year and 5 years of driving experience.
Experimental Design	High Traffic, Low Traffic
Test Modality	Two separate sessions, the first in rush hours (high traffic concentration), the second in low traffic. Each subject tested both sessions. • Phase 1: subject drives the vehicle with the whole system OFF (9.5 km) • Phase 2: subject drives the vehicle with LW ON (10.5 km) • Phase 3: subject drives the vehicle with LW + BS ON (17 km – the track with both systems ON is the longest one)
Aspects to be tested	Usability, Workload, Acceptance, Willingness to pay

<table>
<tr><td>Indicators</td><td colspan="2">❖ Usability
◆ Open-structure interview on system functionality
◆ “Usability Scale”
◆ “Driving Quality Scale” (See Annex 2)
❖ Workload
◆ “Rating Scale Mental Effort” (See Annex 3)
◆ Subjective measurement: “NASA-TLX” – Simplified version
❖ Acceptance
◆ Semantic Differential technique
◆ Willingness to pay
◆ Importance ranking
❖ Primary Task Performance
◆ Number of critical interactions with other traffic participants and number of warnings (recorder and rated by the test leader during the test, together with a brief description of the corresponding traffic situation)
◆ Mean speed
◆ Number of lane change and overtaking manoeuvres
◆ Time to cross lane
◆ Others (i.e.: indicators of steering angle behaviour, …)</td></tr>
<tr><td>Data Collection methodology (how data is collected)</td><td colspan="2">Logging from Cameras, sensors, ECU, transmitted via CAN Bus</td></tr>
<tr><td rowspan="5">Data collection typology (Which data is collected)</td><td>Vehicle Information</td><td>Velocity, Left/Right Direction Indicator, Vehicle Lights;</td></tr>
<tr><td>LW HMI Control</td><td>LW Left/Right Buzzer, Steering Vibration;</td></tr>
<tr><td>BS HMI Control</td><td>BS Left/Right Buzzer, BS Left/Right Led, Magna Left/Right Led;</td></tr>
<tr><td>ECU Lane Position Sensor</td><td>Left/Right Border Visible, Left/Right Border Marker Type, Left/Right Border Line Type, Left/Right Border Distance, Lane Heading Angle;</td></tr>
<tr><td>ECU BS</td><td>Distance of the Object, Alarm Flag, Obstacle Presence</td></tr>
<tr><td>Time of Testing</td><td colspan="2">~120’ for each subject. For each session (two sessions) the timings are
Briefing, 10’ for phase 1, 10-15’ for phase 2, 10-15’ for phase 3, debriefing</td></tr>
<tr><td>Usability Test Used</td><td colspan="2">Brooke Questionnaire</td></tr>
<tr><td>Questionnaire Result (Descriptive)</td><td colspan="2">LSS was rated as quite useful, satisfying, and user friendly with no difference between the two experimental conditions</td></tr>
</table>

Road User Interactions: Patterns of Road Use and Perceptions of Driving Risk [12]
Samuel G. Charlton, Peter H. Baas,
Transport Engineering Research New Zealand Ltd. & University of Waikato

Purpose	better understanding of the human factors of local road transport system: road user demographics, risk perceptions of road users, and the driving attitudes of various road user groups.
Technology used	Pictures
Place and characteristics of Testing	Auckland, Hamilton, Tauranga, Gisborne, New Plymouth, and Palmerston North
Subjects	327 Subjects, 158 men and 154 women (15 not specifying sex)
Requirements for Subjects	Unrestricted driving licence
Subject Notes	General Recruitment (Schools, Sports Clubs, Senior Citizens Associations and Church groups Age from 15 to 78 (mean 41.63, SD 15.35) 35% from main urban area, 41% in secondary urban area, 24% rural areas. Grouping from 5 to 15 people
Test Modality	Rate the relative risk in a series of photographs of common driving situations, their willingness to accept that risk, their own driving skill, and the skill of the other drivers in the situation, using a 100-point scale for each picture.
Test Complexity Scenarios	8 different urban, rural, and motorway driving situations and each driving situation photograph was digitally edited to contain one of several vehicle types (e.g., motorcycle, compact car (coupe), sedan, van or ute, or large rigid truck).
Tasks to be performed	Express options about photographs depicting particular driving situations, according to perceived risk and willing behaviour
Feedback from Subject	Comments about photographs
Aspects to be tested	Human Errors and Risk Perception
Indicators	Personal information (Age, Gender, Ethnographical provenience and so forth) Given a particular situation depicted in a picture, 1) Degree of driving risk in the situation, 2) Willingness to accept the risk in that situation, 3) Degree of control over their own vehicle in that situation, 4) Driving skills of the other driver(s) depicted in the situations.
Test Results	Results about DBQ • Young males much more likely to violate traffic rules and display aggressive acts towards other drivers. • Decrease in scores with driver age, young women did not display the high levels of violations reported by young men. • Statistical analysis indicates significant age and gender differences for violations ($F(2, 304) = 17.11$, $p < .01$ and $F(1, 304) = 19.47$, $p<.01$, respectively), and aggressive violations ($F(2, 304) = 16.31$, $p < .01$ and $F(1,304) = 10.18$, $p<.01$). • Significant positive correlations between these scores and amount of weekly driving reported; $r = .173$ and $r = .144$, $p <.01$ for violations and aggressive violations respectively. Rural residents reported high rates of violations, (marginally significant; $F(2, 321) = 2.62$, $p < .07$.)

	❖ Women: more lapses, more frequently trying to start in the wrong gear, getting into the wrong lane approaching a roundabout, etc.; F (1, 308) = 6.16, p < .01. ❖ Young drivers reported more errors than older drivers (F (2, 322) = 2.59, p < .07). Older drivers and women drivers in the UK/US rate dangerous driving situations as being higher in risk whereas young drivers and men drivers are more willing to accept the risk in these situations and rate their own driving abilities higher (Groeger & Chapman, 1996; Lerner & Rabinovich, 1997). Young drivers in the UK have higher rates of violations and aggressive violations as measured with the DBQ. NZ drivers, (particularly our young males) appear to have much greater propensity for some types of aggressive violations such as racing other drivers at stop lights. This finding is perhaps indicative of the high level of acceptance of speeding by New Zealand drivers. Driving fast in rural roads is more common among NZ young male drivers; Young drivers most likely to report receiving speeding ticket in the preceding year; 26% as compared to 17% of all drivers. **Conflicts & crashes** About Potential for conflicts between road user groups resulting from their differing patterns of road use, attitudes, perceptions, and driving behaviours: High crash rates, particularly for older drivers, at weekdays 14-16 and weekends at 9-11. Older drivers had a disproportionately high number of crossing, turning, and manoeuvring crashes on 50 kph roads during these times of day; Young drivers had high rates of collisions with turning vehicles as well as crossing and manoeuvring crashes in 50 kph zones. Crashes drivers between the ages of 20 and 64 at these times involved higher speed roads and were predominantly overtaking, head on, and rear end crashes. On rural roads, elderly drivers have a very smooth and uniform driving style with fewer accelerations and braking actions than young and middle aged drivers (Schlag, 1993). On inner city roads, the elderly are much more likely than other drivers to ignore red lights at controlled intersections (although less likely to drive through on amber lights), fail to follow give way rules at intersections, and failed to reduce speed at road-level railway crossings (Schlag, 1993). Young drivers in these situations display generally higher speeds, more overtaking manoeuvres, acceptance of smaller gaps between vehicles when turning at intersections, and a more dynamic driving style (rapid acceleration and deceleration, sharp braking, etc.). Although older drivers bring a wealth of advantages to the driving task in terms of experience and knowledge, they generally have greater difficulty perceiving, interpreting, and judging the movements and intentions of other drivers (Schmidt, 1987). When the behaviour of those other drivers is prone to rapid changes and higher velocities, as it is with young drivers, conflicts and crashes are perhaps a predictable result.

EUCLIDE: Enhanced human machine interface for on-vehicle integrated driving support system [45]

PROJECT CO-ORDINATOR : Centro Ricerche Fiat S.C.p.A. (I)

PARTNERS : CEDIP Infrared Systems (F), Volvo Car Corporation (SE), DaimlerChrysler (D), University of Stuttgart IAT (D), Chemnitz University of Technology (D), Robert Bosch GmbH (D), ICCS/NTUA (EL), EC-JRC-ISIS (I)

AUTHORS: M. Martinetto (JRC)

Purpose	Test of Collision Warning and Visual Enhancement ADAS in condition of reduced view (dark, night) and adverse weather conditions (fog) System Usability, System Functionality, Safety Impact	
Technology used	Car equipped with CW and VE system, infrared frontal camera	
Place and characteristics of Testing	Göteborg Test Track (43 KM) ❖ Low traffic (1000 cars/h) ❖ Extra urban: 70 km/h, 2-3 lanes ❖ Highway: 90 km/h, 2 lanes ❖ Highway: 110 km/h, 2 lanes	Turin extraurban track (56 KM) (45') ❖ Mid-High Traffic (by 7.00 PM) ❖ Extra urban: 70 km/h, 2 lanes ❖ Highway: 110 km/h, 2 lanes ❖ Highway: 130 km/h, 3 lanes
Type of Driving Test	Real Traffic	
Subjects	12 in ITALY + 12 in SWEDEN Between 25-35 in ITALY (young), between 50-65 in SWEDEN (elderly) 50% Male, 50% Female	
Requirements for Subjects	At least 5 years of driving experience At least 10.000 Km/year	
Experimental Design	Within-Subjects	
Test Modality	Half subjects begin with the System turned ON, half with the System turned OFF. They are mixed to avoid learning effects. Every Subject drives twice, once for each System State	
Company of Subjects	A technical Expert, who records information about system performance; A human factors Expert, who notes information about driver performance.	
Actions of Accompanyment Tester	Instructs Subjects what to do next, collects real-time data	
Test Complexity Levels	Low Traffic, High Traffic Both good and bad weather No Obstacles artificially placed	
Test Complexity Scenarios	Night Highway Extraurban	
Data Collection methodology (how data is collected)	Pre-Questionnaire (with system off), Post-Questionnaire (with system on), Observation, Log File	
Data collection typology	A-prori/A-posteriori	Log File contents

(Which data is collected)	Acceptance (Questionnaire) Driving quality (Questionnaire) System Performance (Real Time) Scenario Variables (Real Time) Description of Critical Events	1. Time/Scan Number 2. Vehicle Speed 3. Pmd 4. Tpmd (Information) 5. Ttc (Imminent) 6. Time Headway 7. Number Of Information Obstacles 8. Warning Mode
Time of Testing	1,5- 2 hours per subject, including time for filling questionnaires in	

CAMPBELL%202004-01-0450.pdf
Comprehension Testing of Active Safety Symbols [10]
Campbell, J.L., Hoffmeister, D., H., Kiefer, R., J., Selke, D., J., Green, P., Richman, J., B.

Project whose aim is to describe a methodology to identify the best icons for automotive warning and information

Purpose	• Develop a valid and reliable process for comprehension testing of candidate automotive symbols; • Comprehension testing on a set of new symbols for in-vehicle active safety systems.
Technology used	Schemes, Paper
General Tasks	❖ Identify comprehension of given icons ❖ Rank icons which were symbolizing the same aspect according to real icon meaning to the extent of ➢ Identify best icons for given meaning ➢ Identify worst (potentially dangerous) icons for given meaning
Place and characteristics of Testing	U.S., Sweden, Germany, Japan
Subjects	77 Subjects
Requirements for Subjects	- Driver Licence, - at least two years of driving experience, - at least 18 years old, - matching desired combination age/gender, - not working in automotive field
Type of Test	Partial Examination of data
Test Modality	• Subjects had to write down the meaning of the icon they were presented with • Then, they had to Rank similar icons to the intended meaning for the icon group, given in a second time
Test Complexity Scenarios	Each icon presented on a different page
Tasks to be performed	Write meaning of icon Rank icon meaning adherence to real intended meaning
Tasks scheduling	Test Leader
Feedback from Subject	Written response
Aspects to be tested	Correctness of meaning identification
Indicators	Ranking scales
Test Results	• Ranking of icons that represented how well people understand them, and therefore can be safely included into instrumentation; • Ranking of icons that leaded to wrong judgement and that possibly could lower safety, and therefore should be avoided when designing instrumentation

Cyber-Cars Iterative Design Of A New On-Board System For Public Vehicle: From The Idea To The Prototype [67]

Centro Ricerche FIAT S.c.p.a, Italy
Rambaldini Amon, Alessandretti Giancarlo, Irato Giorgio, Nodari Elisabetta, Toffetti Antonella, Zoldan Cristina — FIAT Auto S.p.a., Italy
Motetti Paola

In this project, three tests have been conducted, dealing with different aspects of the project, from mock-up to prototype. The main goal is the Project of a *Cyber Car*, vehicle for public transporting able to a set of functionalities to bring the user where he wants without the need of a driver, autonomous driving and door-to-door service. The project points to realize an User-friendly, self-explainative, usable interface.

First Test: Keyboard Scheme Choice via Performance Evaluation of different schemes

Purpose	Understand which of the designed alphanumerical keyboards is the most suitable to use in the context of CyberCar interface
Alternatives to choose from	Different Types of Keyboard: Base(line), Base +Arrows, Square and Circular
Technology used	Mock-ups
Place and characteristics of Testing	CRF
Subjects	12 Subjects, 8 male, 4 female
Subject Notes	Age 25-33, All CRF internals.
Test Modality	❖ People had to insert the destination “SEBASTOPOLI”; ❖ to fulfil their task, people could either rotate a knob right or left, and press the knob as well ❖ The Test Leader looked at the knob movements and reproduced them by moving a small transparent paper cursor placed over the paper keyboard, simulating the behaviour of the system
Company of Subjects	Test Leader
Actions of Accompaniment Tester	Monitoring of subjects’ behaviour and reproduction of movements
Test Complexity Scenarios	Interact with a spinning knob to insert destinations
Tasks to be performed	Insert the destination “SEBASTOPOLI”
Tasks scheduling	Previously stated
Feedback from Subject	Comments regarding interaction with the knob
Aspects to be tested	Usability, Acceptance
Indicators	HMI task performance (Speed and Errors, Observation)
Data Collection methodology (how data is collected)	Analytical Observation, Interviews

<table>
<tr><td>Data collection typology (Which data is collected)</td><td colspan="2">❖ Differences in
◆ Performance Timings among different types of keyboards;
◆ Performance Timings among different types of keyboards and Baseline (the best obtainable performance with each specific keyboard)
◆ Total number of steps among different types of keyboards;
◆ Total number of steps among different types of keyboards and Baseline
◆ Usage of Arrows among different types of keyboards and Baseline
◆ Number of times each Arrow is used (only for the device that scored best, to check for learning effects)
❖ People’s subjective preference for a specific device
❖ Personal Comments</td></tr>
<tr><td rowspan="3">Test Results</td><td colspan="2">Best Keyboard in all tests: Circular Scheme</td></tr>
<tr><td>Type</td><td>Worst</td></tr>
<tr><td>Performance Timings and Number of Steps</td><td>Baseline</td><td>Square</td></tr>
<tr><td>Usage of Arrows</td><td>Users</td><td>Base + Arrows</td></tr>
<tr><td></td><td colspan="2">Description</td></tr>
<tr><td>Number of Times each arrow is used</td><td colspan="2">Even people not attracted by arrows begin, after a period of time, to make consistent use of them</td></tr>
<tr><td>Subjective Preference</td><td colspan="2">Worst impact with Square format</td></tr>
<tr><td>Personal Comments</td><td colspan="2">Incoherence between user’s prediction of next letter (or arrow) reached by interaction with knob, either by rotating and pressing it, more remarked on the square model</td></tr>
</table>

Second Test: Evaluation of Stops and Music Menu

Purpose	Evaluate Efficacy and Efficiency of proposed solutions in Stops and Music Menus
Alternatives to choose from	Different Types of Keyboard: Base(line), Base +Arrows, Square and Circular
Technology used	DENIM 0.1 free by University of California Berkeley; Mid-Reliable prototype simulating Target and Music menu interaction logic in a neutral context (no colors, particular forms or specific fonts)
Place and characteristics of Testing	CRF
Subjects	12 Subjects, 8 male, 4 female
Subject Notes	Age 25-33, All CRF internals.
Test Modality	Interview of Subjects when interacting with prototype. They were asked ❖ Before specific selection ◆ The meaning of selected label ◆ How they expected the procedure to go on ❖ After specific selection ◆ Which options among those showed were superfluous ◆ Which options were missing
Company of Subjects	Test Leader
Actions of Accompanyment Tester	Recording of user’s comments

Test Complexity Scenarios	Interact with a prototype and make menu choices
Tasks to be performed	Interaction with menu: browsing, selection
Tasks scheduling	Previously stated
Feedback from Subject	Comments regarding interaction with the menu: what is missing, what is superfluous, what is to be reorganized
Aspects to be tested	Usability, Acceptance
Indicators	Specific Usability Questions
Data Collection methodology (how data is collected)	Observation, Interviews
Data collection typology (Which data is collected)	Personal Comments
Test Results	List of topic to be reviewed, deleted, reorganized, joined, inserted. Among all: ❖ Add Map Presentation to Stops Listing ❖ Select a sector of a Map ❖ Timings of routes ❖ Pointers "You are Here"-Destination" over the Map ❖ Different names for some functions ❖ Useful Categories ❖ No sequencing possible in Music section ❖ Alphabetic songs listing

Third Test: Final Prototype

Purpose	Define the weakness of the CyberCars HMI high-fidelity prototype, by consulting a group of "future users" to collect opinions, errors of use, time performance.
Technology used	Rotating Knobs, One button for Operator's call, one button for Emergency call
Topics to test	Adequacy of prototype according to different aspects
Physical Interface	1. Input Device: Knob and Buttons 1.1. Test knob adequacy for user target and type of task 1.2. Test Self-explanatory use of other knobs 2. Mapping between Input Device and Graphical Organization 2.1. Allocation of graphical objects matching with rotary knob movement (e.g. circular organization of graphical objects related with knob rotational movement)

Virtual Interface	1) Information Readability: Format of Information displayed easily readable without effort by users (e.g. character dimensions, character contrast) 2) Graphical Design: aesthetical appearance pleasant or color choice is inconsistent with stereotype 3) 2nd evaluation of contents: evaluate for the second time (the 1st time was during low-fidelity prototype evaluation) and more precisely the adequacy of services offered by the system, in order to understand whether they are satisfying user's needs (adequacy of contents) 4) 2nd evaluation of contents meaning: understand if labels of different functions are consistent with users expectations and easy understandability by users (adequacy of used words) 5) 2nd evaluation of contents logical organization : understand if Information Categorization (functions), displayed on the interface, is consistent with expectation and stereotype of users (in which menu a specific service is categorized) 6) System Navigational Performance Efficiency 7) System interaction Learnability: understand whether people improve performance by training
Transport Service	Adequacy of CyberCars Transport Service: collect people expectation about transport service
Place and characteristics of Testing	CRF
Subjects	24 Subjects, 20 male and 4 female
Subject Notes	Average age: 27.9 years (youngest 24, oldest 35). Graduated: 21 Secondary school diploma: 3
Test Modality	Co-Discovery Learning Technique 1) Pre-Questionnaire 2) Battery of 7 Tasks to be accomplished in pairs. After each task, both subjects had to express aloud their comments 3) Extensive use of system; Comments aloud 4) Repetition of 7-Task Battery 5) Post-test Questionnaire 6) Final Questionnaire after one week of test, by means of email.
Company of Subjects	Test Leader
Tasks to be performed	Exhaustive use of System
Feedback from Subjects	Comments about system, Pre-Post-Final Questionnaire
Aspects to be tested	Usability, Acceptance

<table>
<tr><td rowspan="6">Indicators</td><td colspan="4"></td></tr>
<tr><td colspan="4">Usability</td></tr>
<tr><td colspan="2">Perceived</td><td colspan="2">Measured</td></tr>
<tr><td>What</td><td>How</td><td>What</td><td>How</td></tr>
<tr><td>Comments</td><td>Semi-structured interview during tests</td><td>Deviation from Ideal Performance (efficiency)</td><td>Confrontation of Performance time of each subject with a baseline time (a repeated measure of experimenter performance)</td></tr>
<tr><td colspan="2"></td><td>Deviation from First Time Performance and another (learning)</td><td>Confrontation between performance time when the user never tried the task before and performance time after a long exploration of system</td></tr>
<tr><td colspan="2"></td><td>Number of Navigation Errors</td><td>Log files</td></tr>
<tr><td colspan="2"></td><td>User Satisfaction</td><td>Two Post-Task Questionnaires</td></tr>
<tr><td>Test Results</td><td colspan="2">Perceived Usability Issues

❖ Things to modify (labels not understood, current selection not clearly highlighted)
❖ Things to Add: short description of selections
❖ Things to remove: System status display</td><td colspan="2">Measured Usability Results
❖ Efficiency: People performance significantly different from baseline in almost all cases. It can be due to the deep knowledge of system by the baseline, but it is observable that after a period of use, user’s performance converges to baseline’s.
❖ Learning: not many errors, so it is possible to conclude that the interface is self-explanatory and needs no specific training to be used. Anyway, it has been observed that First Time Performance is lower than performance after training
❖ Number of Navigation Errors: not many errors. Anyway, it was observed a misunderstanding of radio tuning functions. A solution has been integrated.
❖ User Satisfaction: measured using two questionnaire. One was administered immediately after the interaction with system and the other was administered after a week. The Semantic Differential Technique was used. The information collected with both are here organised as follows:
◆ Habits in moving in the urban context: The need for urban transport system is needed, although the current one is valued usually late on service;
◆ Opinions about CyberCars Transport Services: considered Useful, reliable, safe, more punctual even if pointed as more expensive than bus.
◆ Evaluation of the CyberCars Interface: very innovative and consistent with expectations, easy to use and learn
◆ Consideration about the use of rotary knob: adequate for specific context. Problematic topics such long lists and writing speed were adjusted</td></tr>
</table>

E-Safety - LAVIA: the French project on intelligent speed adaptation
test in progress in INRETS
LAVIA: Limiteur s'Adaptant à la VItesse Autorisée (Adaptable Limiter to Permitted Speed)

Purpose	❖ Test the Acceptance of the System by the drivers ❖ Test the Driver Behaviour and impact of system on driving The System adapts speed according to the mandatory limits, by decreasing it when needed or by notifying the driver the high speed
Technology used	❖ Two Fully instrumented prototypes with two synchronized acquisition systems : video and data ❖ A fleet of twenty vehicles ❖ Reduced instrumentation (no video) ❖ Video recorder and video playback ◆ 3 synchronized video channels, 25 frames per second ◆ 3 hours of autonomy ◆ 3 views restitution on a replay 3 screens workstation with data incrustation ❖ Data acquisition system ◆ 8 weeks of autonomy ◆ Auto-diagnostic and reports transmitted daily using GSM and SMS. ◆ CAN bus interface ❖ GPS
General Tasks	Test Advisory System (when ON, notification to driver of speed exceeding) Voluntary Active System (System automatically controls speed, which cannot exceed maximum limit; can be turned OFF) Mandatory Active System (same as above, cannot be turned OFF)
Place and characteristics of Testing	- Large trial area close to Paris Including the city of Versailles, Saint Quentin en Yvelines, Velizy - Representative of actual routes covered by drivers in suburb of Paris - Broad variety of road situations : urban road, urban motorway, inter-urban road (up to 110 km/h speed limit) - Possibility to get full route from origin to destination
Subjects	100 drivers who can use the vehicle freely during 8 weeks
Subject Notes	
Accompaniment of Subject	Psychologist or Ergonomists
Action of Accompanying Person	Observe driver's behaviour while driving and interacting with other drivers

Test Modality	❖ Step 1 : Drivers recruitment ◆ Attitude questionnaire to investigate driver's opinions towards speed, safety, speed limiter… ◆ Drivers recruitment with respect of various selection criteria : attitudes, social groups, age, gender, etc. ❖ Step 2 : In-car observations by observers with a selection of 16 drivers on 2 vehicles prototypes ◆ Study focuses on specific driver behaviour and other vehicles behaviour on a reference course ◆ Detection of potential ergonomics problems ❖ Step 3 : Middle-scale evaluation ◆ Middle-scale in-field experiment ◆ 20 vehicles loaned to 100 drivers during one year ◆ 5 waves of 8 weeks with 20 drivers per wave ◆ Each driver tests 4 modes : neutral, advisory only, voluntary, mandatory ❖ Acceptance and behaviour evaluation ◆ Interview or questionnaire after each variant ◆ Data logging to acquire information about driver behaviour, driving context, current speed limit, actual speed, driver identifier, travel motive etc. ◆ Data processing for quantitative behaviour evaluation
Aspects to be tested	Acceptance, Workload
Indicators	Pre-Questionnaire, Current and Authorized speed, acceleration, followed path, System Switching

Subjective Assessment Methods – PSA: Proposition of a new Method [66]

PSA introduces a new technique to evaluate Workload, by using a multidimensional scale: PSA-TLX. This test grid reports an experiment whose results were captured by this new method.

Purpose	Assess Driving Workload when using night-driving support system
Subjects	10 drivers
Subject Requirements	- right-handed - experienced - used to night-driving and to type of road familiar with itinerary
Company of Subjects	Test Leader
Type of Testing	Within-Subjects
Test Complexity Scenarios	0. Familiarisation with the vehicle + night vision system: 5 min 1. Baseline condition : Driving (no PDT, no night vision system): 10 min 2. Driving with the PDT (without the night vision system): 10 min 3 Driving with the NV system (continuous mode without alert or alert mode) + PDT: 10 min 4. Driving with NV system (continuous mode with alert): 10 min
Tasks to be performed	Two levels of Task Difficulty: 1) Changing display settings (colour and brightness) 2) Destination entry with « map-pointing » (zoom + scrolling)
Aspects to be tested	Workload
Indicators	PSA-TLX

Data Collection methodology (how data is collected)	Questionnaire
Data collection typology (Which data is collected)	Subjective Measures
Test Results	Results were expressed with graphs for Axes Effort and Disruption. Histograms with coloured display of Weight of Task Dimensions was traced. Each graph has two columns, one depicting reference, without use of Night Vision System, and the other using the Night Vision System. Results express good impact of system on 4 dimensions out of seven (lateral and longitudinal control, reactivity to static environment, itinerary following) about Disruption, while the system need more effort for lateral control, and performs good for other dimensions.

SENECA Project: Users Evaluation [56]
Toffetti, A., Nodari, E., Zoldan, C. (CRF)

Purpose	Evaluation of Usability of SENECA system, whose purpose is to enable the command of a considerable number of devices via vocal means, compared with one system controlled by manual means; Evaluation of SENECA impact on driving Analysis of SENECA recognition capabilities
Technology used	Seneca prototype installed on vehicle (Lancia Lybra) Seneca system constituted by COMAND multifunctional system and by a PTT little lever mounted under the system, Three cameras have been mounted for audio-recording each interaction and video-recording of road, participants glances and system display
Subjects	16 Male Subjects
Requirements for Subjects	• Active driving licence for at least 6 years • Drive at least 6000 kilometres per year • 10/10 sight, with correction as well. • At least a high school degree • Familiar with mobile phones and computers.
Subject Notes	8 Subjects with age 25-35, 8 Subjects with age 55-65
Type of Test	Within-Subjects
Test Modality	Description of Basic interaction with the system
Accompanyment of Subjects	Two experimenters, one to instruct Subjects, one to monitor subjects what to do
Test Complexity Scenarios	Vocal interaction compared to manual interaction
Tasks to be performed	Users were asked to accomplish these tasks: 1) Enter pin code to switch system on 2) Call a person whose number is stored in phone agenda 3) Store the radio station subject is listening to 4) Ask the system a list of stored stations 5) Call a telephone number at Subject's choice 6) Switch to CD and select given track 7) Use the information coming from the Navigator to reach a given destination 8) Select a given Radio Station 9) Switch to tape and change the side of the cassette 10) Erase one destination from the Navigator memory
Tasks scheduling	Subjects are instructed by experimenters
Feedback from Subject	Vocal Comments, Questionnaires
Aspects to be tested	Usability, Acceptance, Performance
Indicators	Error Number, Subject Monitoring
Data Collection methodology	Questionnaire, Semi-structured interview, Observation
Data collection typology	Subjective Measures

Test Results	Seneca efficiency is not high, when compared with an ideal performance, contrarily to effectiveness that is high: both elderly and young participants solve a high percentage of tasks correctly. Regarding to Subjective evaluations, it has emerged that both young and elderly expectations on vocal system are very good and after interacting with Seneca, participants maintain their positive judgement. Comparison between manual and input system has shown that manual input modality is more effective and also efficient (in this case only for young users) than vocal one, even if Seneca system has been subjectively evaluated more positively than the manual one. Observational data and subjective perception point out that Seneca system has had a minor impact on driving respect to the manual one. Seneca seems to need some improvements in its recognition capabilities and in its interface in order to allow an increase of effectiveness and efficacy in the interaction.

9.2. Step-By-Step (SBS) Procedure

By generalization of case studies previously seen, it is possible to extract the following Step-By-Step procedure for Test Design. In the Procedure, highly structured, the following symbols have been used, with the corresponding meaning:

⊗	The topic is required for testing
⊕	Subtopics can stack (e.g. different elements of the sublist can be used together)
∅	Subtopics are mutually exclusive (e.g. just one element of the sublist can be used at one time)
	No symbol means that the topic can be omitted in testing, even though inserting it should improve test description
⊗⊕	This means that the principal topic (the one that resides on the same line of the symbol) is required, and is structured in subtopics, from which a subset of topics can be specified
⊗∅	This means that the principal topic (the one that resides on the same line of the symbol) is required, and is structured in subtopics, from which just one subtopic can be specified at one time
⊕∅	Has no meaning, as the two symbols cannot go together

Note: Best attribution of symbols to tree structure is done bottom-up (from leaves to root, that is from deeper levels e.g. 1.a.i(1).(a) towards 1) in order to capture correct group nidification.

SBS Procedure

1) **⊗ Definition of the object of the test, that is what we are going to test**
   a) Brief description of Test Target
2) **⊗ ⊕ Definition of the technology the test will be conducted on**
   a) Main device
   b) Auxiliary devices used in the test (haptic devices, sound, video, text)
3) **⊗ Definition of General Tasks to be completed in order to test the aspects of the device**
4) **⊗ Definition of Main Test Phases**
   a) Definition of Time for accomplishing each phase
5) **⊗ ⊕ Place of Testing (list of Test Places Scenario)**
   a) ⊗ Choice of Test Drive Type (Simulator, Test Track, Real Traffic)
   b) Specification of Lane characteristics (Number of Lanes, dimensions, carriages, Lane Markings) where applicable
6) **⊗ Identification of Subjects**
   a) ⊗ Choice of Number and Gender distribution
   b) ⊕ Identification of Subject Constraints
      i) Age
      ii) Driving Licence time ownership
      iii) Driving Experience
      iv) Special attitudes
      v) Familiarity with PC devices (mouse, keyboard)

vi) Level of driving expertise inside specific range [this is to have all subjects at the same level of expertise]

c) ⊕ Subject Notes (Normal vision, Age, Study titles)

7) ⊗ ∅ Experimental Design (within-subjects, between-subjects, Mixed)

a) ⊗ Define Test Modalities and initial settings for groups (which groups are going to use what, rules for changing orders…)

8) ∅ Define accompaniment of subjects

a) Subjects are alone with instrumentation

b) There is an active test leader: he instructs the user on what to do next, collects data, answers questions

c) There is a passive test leader: he just monitors user's behaviour and collects data

9) ⊗ Identification of Test Complexity levels (High/medium/low traffic)

a) ⊕ For each level, Identification of applicative Scenarios

i) Hazards

ii) Particular climatic conditions (Good, wet, dry)

iii) Light Conditions

b) For each Scenario, identification of a set of Tasks to be performed in relation with General Tasks

10) ∅ Specification of Task Scheduling

a) The Subject follows a pre-made scheduling which specifies the time to accomplishment for every task

b) The Subject must accomplish all tasks in a list, in any order

c) The Subject is free to act, and he is monitored by a test leader

11) ∅ Identification of Feedback from the Subject

a) Simple Notification of task initiation and end

b) Information extraction from data sent to the Subject

c) Continuous speech about what the subject is expecting to happen, what happens, what impact has it had on him (*Thinking-Aloud*)

12) ⊗ Identification of Test Variables

a) ⊗⊕ Identification of Independent Variables

i) Age

ii) Sex

iii) Driving qualities

iv) Discriminant qualities

b) ⊗ Identification of Dependant Variables

i) ⊗ Identification of Relation between Dependant Variable and a subset of Independent Variables

ii) ⊗ Description of Relations (functions) identified

13) ⊗ ⊕ Identification of Factors to be tested (Workload, Acceptance, Usability, Driving Performance/Safety)

a) For each Factor, Identification of

i) ⊗⊕ Indicators, using standard scales or measurements

(1) ⊕ Usability

(a) *Usability Scale*

(b) *Driving Quality Scale*

(c) *Open-structure interview on system functionality*

(d) *Specific Usability Questions*

(e) *Brooke Questionnaire*

(f) *Other Standardised Questionnaire*

(g) *HMI task performance (speed and errors)*

(h) *HMI task performance (observation and verbal measures)*

(i) *Verbal description of barriers to use*

(ii) Behavioural signs of uncertainty how to proceed

(iii) Failing trials in task fulfilment

(iv) Comprehension of HMI Output

(i) Co-Discovery Learning

(j) Thinking-Aloud

(k) Self-reporting Diaries

(2) ⊕ *Acceptance*

(a) *Semantic Differential technique (or Standardised attitude scale, Van der Laan et al, 1997)*

(b) *Willingness to use*

(c) *Willingness to buy*

(d) *Importance Ranking*

(e) *Standardised Questionnaire*

(3) ⊕ *Driving Performance, Safety*

(a) Average Vehicle Speed

(b) Standard deviation of vehicle speed

(c) Standard deviation of yaw rate

(d) Maximum of yaw rate

(e) Standard deviation of steering angle

(f) Average distance to the left and right border of the lane

(g) Standard deviation of distance to the left and right border of the lane

(h) Frequency of lane warnings

(i) Maximum of risk level (risk level estimated by the IM)

(j) Maximum duration of risk levels

(k) Average value of DQS ratings

(4) ⊕ Workload

(a) ⊕ Objective Workload

(i) Missed PDT signals

(ii) Average value of PDT reaction times

(iii) Lane Departure

(iv) Standard Deviation of Lane Position

(v) Response Time to an in-vehicle prompt

(b) ⊕ Subjective Workload

(i) RSME

(ii) NASA-TLX

(iii) PSA-TLX

(iv) DALI

(v) MCH

(vi) SWAT

(vii)SWORD

(viii) Bedford Scale

(ix) Test of Estocolmo University

(x) Modified Cooper-Harper Scale

ii) ⊕ Type of Indicator (Objective, Subjective)

iii) ⊗ Measuring Means (log file, questionnaire, interview)

14) ⊗ Definition of Data Collection Methodologies (specifies how data is collected)

a) ⊕ from Cameras, from speech, from logging, from observation

15) ⊗ Definition of Data Collection Typology (specifies what data is collected)

16) ⊗ Test Results

17) ⊕ Test timings

a) Overall test duration per Subject

b) Projection on how long the entire test will take for all Subjects

18) Definition of Interview Contents

19) ⊕ Definition of Questionnaire Contents, related to the Subject who takes it

a) Grouping of similar questions, with specification of reading all of the before answering

b) Results of Questionnaire (Numerical)

c) Results of Questionnaire (Descriptive)

20) ⊕ Identification of aspect better performed by Subjects

a) Numerical Tables, indices according to indicators chosen
b) Calculation of Standard Error, means and other statistical parameters
c) Comment of most performant values

21) ⊕ Identification of aspects badly performed by Subjects

a) Numerical Tables, indices according to indicators chosen
b) Calculation of Standard Error, means and other statistical parameters
c) Comment of less performant values

22) ⊕ Definition of Briefing

a) Description of the System used in the test
b) Description of details to be tested
c) List of tasks to be accomplished, if the subject is left alone with the instrumentation
d) Time for each section of test, if relevant for the test itself

23) ⊗ Conclusive section, based on numerical results, interviews, questionnaires

a) Specification of what results have been achieved
b) Specification of what results are yet to be achieved
c) Specification of what results were disappointed by experimental testing, and why

9.3. Aspects to be taken into account when assessing IVIS / ADAS

Below is a listing of topics to deal with when evaluating IVIS / ADAS Usability, Acceptance, and Workload. These topics have been identified throughout all examined literature. Topics can be divided according to:

- Visual Aspects
  - Visibility and Legibility, Readability of Textual Information

Font size, contrast with background

  - Chromatic Contrast between active items and Background
    - Chromatic problems with color-blindness people
  - Luminance Contrast vs. Ambient Illumination
    - Legibility of display according to outside illumination (dark, reflection)
  - Lexicon Content of Textual Information
    - Content's choice of terms and language
  - Icons Characteristics
    - Icons Dimension, Size, Location when
      - The user is not driving
      - The user is driving
    - Icons Contrast with Background
    - Affordance of Icons
      - Perceived meaning in relation with real meaning
    - Textual description of Icons
    - Metaphorical usage in active Content
    - Active Selection correctly highlighted (in lists, in buttons, in icon groups)
    - Density of Information

Too much density brings to difficult or impossible understanding of meaning

  - Semantic Content
    - Icons, Textual information well-grouped according to some semantic criterion
  - Space Organization
    - Display of much information (e.g. Web Pages) is easily readable on screen
- Auditive Aspects
  - Low SNR (Signal-To-Noise) ratio
    - The user does not understand the synthetic voice due to outside traffic noise
  - Intrusive warning messages
    - Beeping continuously can irritate the user
  - Too loud sound or too high frequency

        - Sudden Warnings at loud or too high frequencies can fear the user or irritate it
    - Directional Warning understanding
        - Does the user understand what side the warning refers to?
    - False Warning intrusion degree
        - False Warnings generated too frequently irritate the user
    - Too dense information
        - The user can forget key parts of message
        - Persistent data should not be rendered as Auditive, but visual
    - Vague Terminology
- Tactile Aspects
    - Too much / Too few resistance
    - Checking of controls resistance and / or sensitivity when wearing gloves
- Interaction with System
    - Too much complexity in interaction
        - Too many steps or too many widgets to interact with
    - Commands or output not understood
        - Terms used are too technical, user can not understand promptly the output
        - Terms use background different from user's. Possible mapping incoherences
    - Just-in-time services
        - Informative messages arrive on time (e.g. Navigator TURN LEFT enough before the turning time)
        - Warning Messages Completeness (e.g. lateral warnings specify what side the warning concerns)
    - Information Categorisation precision (topics in the wrong category e.g. Hotels in 'traffic' instead of 'services' or 'leisure')
    - Too much time to do something
        - Time subtracted from driving
    - Sensation to be controlled by the system
        - The system follows a protocol too much inflexible ; the user has to adapt to the system
    - Input Device
        - Too much sensitivity can end in wrong selections
        - Few sensitivity can delay task execution
        - Device position and exercise of it
    - Help System
        - Users use it and do not find what they are looking for
        - Users do not use it (Why ?)

- Interface self-location
  - If the user can not easily and promptly find his/her position in menus, he/she can get frustrated
- Consistency of multimodal information
  - Different modalities, when providing simultaneous feedback, must be synergistically accorded and not give contradictory information (e.g. visual and auditive modalities)
- Retention
  - How easy is to retain modal interactions between driving sessions (e.g. the user learns to use the system and remembers how to use it)

❖ Performance with the System
- System's ability to compensate Workload when acting on it regarding driving safety
- System Efficiency
- System Use Satisfaction
- Willingness to use System
- System Reliability

❖ Adaptability of System
- Number and variability of different Visualization / Interaction possibilities
- Precise aspects / functions can be hidden in order to save Time for selection and Space on the display
- Detail information suitable for User Type (e.g. more info about monuments for tourist, more info about viability for everyday use)

9.4. Summary of Techniques and Projects

Below is a table which summarizes the Choice of Evaluation Methods in the Projects described so far.

Project	*Usability Techniques*	*Acceptance Techniques*	*Subjective Workload Techniques*
LACOS	• Functional Behaviour	• Willingness to pay via verbal report	
SAVE-IT	• Comprehensibility and User Errors / Problems by Observation • Interviews • Questionnaires	• Surveys • Questionnaires	
IN-ARTE	• Standardized Questionnaire	• Standardized Questionnaire	
UDC	• HMI Task Performance by Observation and Verbal Measures	• Willingness to Pay • Willingness to Purchase	
COMUNICAR	• Observation of HMI-Task Performance (Speed and Errors) • Questionnaires	• Semantic Differential • Questionnaire	• RSME • Average of DQS • Questionnaire
ADVISORS	• Open-structure interview • Driving Quality Scale • Usability Scale	• Semantic Differential • Willingness to Pay • Ranking	• NASA-TLX • RSME
EUCLIDE	• Driving Quality questionnaire	• A Priori / A Posteriori Questionnaire	
CAMPBELL	• Questionnaire • Ranking	•	
CYBERCARS	• Analytical Observation • Interview • Performance Compared with baselines • Error Numbers	• Interviews • User Satisfaction	
ADAM			• Lateral Position • Number/Duration of Glance • Perceived Distraction
HFES			• SD of Lane Position • Line Crossing Rate • Speed Loss
PSA, VODIS			• PSA-TLX
SENECA	Number of Errors Questionnaires Semi-structured Interviews	Differential Scale Willingness to Purchase	•
INRETS (DALI)	Number of Errors Usability Questionnaires		• DALI

Projects using Techniques

Technique	Type	Used in																		
The listing is far to be exhaustive and partitioning	*The projects were also investigating other aspects that are beyond the scope of this paper*	ADVISORS [2]	CYBERCARS[67]	COMUNICAR	EUCLIDE [47]	ADAM [5]	UDC	SAE [10]	SAVE	IN-ARTE [4, 9]	LACOS	HFES	SRS	PSA [66]	VODIS	ISA Wales	ISA Sweden [27]	SENECA [56]	INRETS DALI
Questionnaires	Usability, Acceptance	✓		✓	✓			✓	✓	✓								✓	
Surveys	Usability, Acceptance	✓	✓						✓									✓	
Driver Behaviour Questionnaire													✓						
Focus Groups	Usability, Acceptance		✓																
Observation	Usability, Acceptance		✓															✓	
Card Sorting	Usability, Acceptance		✓					✓											
User Errors Monitoring	Usability		✓						✓									✓	✓
Comparison with Baselines	Usability		✓																✓
HMI-Task Performance	Usability			✓			✓											✓	
Usability Scale					✓														
Checklists	Usability, Acceptance, WL				✓														
Ranking	Usability, Acceptance, WL	✓	✓					✓											
Willingness to pay	Acceptance	✓					✓				✓								
Willingness to purchase	Acceptance						✓											✓	
Semantic Differential	Acceptance	✓		✓														✓	
Driving Quality Scale	Usability	✓			✓														
RSME	Workload	✓		✓															
PSA-TLX	Workload													✓	✓				
NASA-TLX	Workload	✓		✓															
Perceived Distraction	Workload					✓													
Lateral / Lane Position	Workload					✓						✓							
Verbal Measures	Usability						✓											✓	
Wiener Fahrprobe	Workload, Usability															✓	✓		
DALI	Workload, less Usability																		✓