INFORMATION SOCIETY TECHNOLOGIES (IST) PROGRAMME

AIDE
IST-1-507674-IP

Review of existing techniques and metrics for IVIS and ADAS assessment

Deliverable No. (use the number indicated on technical annex)		**D2.2.1**	
SubProject No.	**SP2**	SubProject Title	**Evaluation and assessment methodology**
Workpackage No.	**WP2.2**	Workpackage Title	**Driver workload and distraction and assessment methods and tools**
Activity No.	**A2.2.1**	Activity Title	**Review of existing techniques and metrics for IVIS and ADAS assessment**
Authors (per company, if more than one company provide it together)		**E. Johansson and J. Engström (VTEC), C. Cherri, E. Nodari and A. Toffetti (CRF), R. Schindhelm and C. Gelau (BASt)**	
Status (F: final; D: draft; RD: revised draft):		**F**	
File Name:		**Deliverable2.2.1_Final_v2.doc**	
Project start date and duration		**01 March 2004, 48 Months**	

Table of Content

List of Abbreviations and Glossary

ACC   Adaptive Cruise Control
ADAS   Advanced Driver Assistance System
CMT   Continuous memory task
DOD   US Department of Defense
ISO   International Organization for Standardization
IVIS   In-Vehicle Information System
LED   Light Emitting Diode
PASAT   Paced Auditory Serial Addition Test
PDT   Peripheral Detection Task
RNG   Random Number Generation
SA   Situation Awareness

List of Figures

List of Tables

Executive Summary

This report presents a review of existing methods and tools which are relevant to the off-line assessment of driver workload and distraction during use of IVIS and ADAS.

The deliverable should be seen as a complement to D2.1.1 produced in AIDE WP2.1 where D2.1.1 cover a wider range of assessment methods for both safety and usability evaluations of IVIS while D2.2.1 focuses on a more detailed review of existing general offline measurement techniques. The deliverable mainly focuses on the methods and techniques which will be used and further developed within the AIDE project. Therefore, some physiological measurement methods and techniques are only briefly described.

The following partners have contributed to the writing of this report:
- VTEC: Chapter 1-3 and 5
- BASt: Chapter 4 and 6
- CRF: Chapter 7 and 8

Chapter 1 presents some of the main resources which will be referred to later on in the deliverable. They are mostly larger EU projects either finished or on-going with the main focus of either method development or applying methods in the assessment of IVIS and ADAS.

In chapter 2 a range of metrics are presented related to driving performance. Driving performance in this chapter deals with the driver's ability to control the vehicle both in relation to lateral and longitudinal control as well as primary task event detection metrics.

Chapters 3 review the background to the assessment of IVIS with visual performance indicators as well as describe recent and on-going research in the field. The review covers both experiments where data has been collected the traditional way through manual transcription of video data and also by the range of more or less automatic head- and eye trackers.

Chapter 4 presents how the Occlusion technique can be used in order to measure driver visual distraction caused by IVIS. The Occlusion technique is considered to be a surrogate measure and the chapter discusses the technique in relation to its validity and reliability.

Chapter 5 very briefly describes some of the main physiological measurement techniques that can be used to assess IVIS and ADAS.

Chapter 6 presents the ideas behind the secondary task paradigm. The Peripheral Detection Task (PDT) is described as well as the Paced Auditory Serial Addition Test (PASAT) and different embedded secondary tasks.

Chapter 7 covers subjective assessment methods (one- and multidimensional). The chapter both describes both the methods which cover the drivers' self-reported measure of driving performance and workload as well as the reported driver performance as observed by an expert.

Finally in chapter 8 different methods and metrics for Situation Awareness are described and discussed.

1 Introduction

One of the main aims in AIDE is to (according to the 'Annex 1 – Description of work'):

> *[…] reduce the level of workload and distraction related to the interaction with individual and combined in-vehicle information and nomad devices […]*

A number of previous and on-going initiatives have focused on the development of, in most cases test batteries, which include techniques and metrics to measure workload and distraction. The intention has been to validate in-vehicle systems with respect to their potential negative effect on safety.

In AIDE, this work is continued but with the focus not only on single use of one specific IVIS and ADAS but on:
1. Multimodal HMI devices shared by different systems.
2. Systems which are coupled to a centralized intelligence for resolving conflicts between systems (e.g. information prioritization and scheduling).
3. Adaptivity of the HMI to the current driver state/driving context.

This means that the test methodology in SP2 WP2.1 and the techniques and metrics reviewed in this document, if later chosen to be further developed within SP2 WP 2.2 in AIDE ideally should be able to capture effects of workload and distraction for those systems.

The deliverable should be seen as a complement to D2.1.1 produced in AIDE WP2.1 where D2.1.1 cover a wider range of assessment methods for both safety and usability evaluations of IVIS and ADAS while D2.2.1 will focus on a more detailed review of existing general offline measurement techniques. The deliverable mainly concentrates on the methods and techniques which will be used and further developed within the AIDE projects. Therefore, some physiological measurement methods and techniques are only briefly described.

1.1 Structure of the deliverable

The present review is based on the actual techniques and metrics and not e.g. on concepts such as physical and mental workload etc. Each section describes the background behind the techniques and metrics. Also, the actual metrics are defined along with presentations of e.g. projects and specific experiments where the metrics have been applied. Each section also provides the reader with issues to further consider in future development within the field (e.g. within the different tasks in SP2 WP2.2).

More direct measurement techniques such as driving and visual performance is presented first and then followed by techniques within the secondary task paradigm. Then Subjective assessment methods are presented as well as some techniques and metrics with in the field of Situation Awareness. The Situation Awareness techniques could often be seen as combinations of the some of the techniques described in previous sections in this report.

1.2 Introduction to literature resources

Some of the references in the following review are from a set of finalized or on going projects. Below are brief descriptions of the some of those projects and resources. Some projects have a focus on method development whether some others mainly aimed at system development.

1.3 Method development

Below are descriptions of projects which have as main aims to develop methods in order to assess either ADAS or IVIS or sometimes both.

1.3.1 Humanist

The newly started Network of Excellence within the 6th framework program called HUMANIST (HUMAN centred design for Information Society Technologies) could be seen as an umbrella organization where e.g. results from projects such as the ones presented further below could feed into. Also, the main aim of HUMANIST is to gather most of the human factors and cognitive engineering competencies in Europe in order to contribute to the eSafety initiative by introducing human centred design for IVIS and ADAS.

The network begun in early 2004 and will end 2008.

1.3.2 HASTE

The main aim of HASTE (Human Machine Interface And the Safety of Traffic in Europe) is to develop methodologies and guidelines for the assessment of IVIS (in-vehicle information systems). For further overall project description see Roskam et al (2002).

The results of WP2 are presented in a public report (Östlund, 2004) which presented preliminary applications of scenarios, dependent variables and data analysis methods in simulator, laboratory and field experiments. This set up had been proven to be useful for the assessment of surrogate IVIS (one visual and one auditory/cognitive task) and in the ongoing work which will validate the results with real systems and tasks. The final result should be a recommendation for a test regime which should be:
- Technology independent;
- Has safety-related criteria;
- Is cost effective;
- Is appropriate for any system design; and
- Is validated through real-world testing.

The project started in 2002 and is scheduled to be finished at the end of 2004.

1.3.3 RoadSense

The main objectives of RoadSense (Road Awareness for Driving via a Strategy that Evaluates Numerous SystEms) are very similar to HASTE. The scientific objectives of RoadSense are (according to RoadSense web page):
- To develop driver behavioural indicators for safety, comfort and support assessment.
- To develop a framework for the integration of existing tools and techniques with regard to driver sensory and cognitive capabilities.
- To identify new tools and techniques derived from critical scenarios and technology case studies being proposed by industry.
- To develop a framework within which HVI (Human Vehicle Interactions) requirements validation techniques can be tested and co-ordinated - before vehicle technology is made available.

The project started in 2001 and has just recently been finished.

1.3.4 ADAM

The ADAM (Advanced Driver Attention Metrics) project aims like HASTE and RoadSense to investigate and suggest methods and tools to assess existing and future in-vehicle information systems.

The project partners are BMW and Daimler Chrysler. The project also has scientific board which consists of members from e.g. ISO groups, AAM, universities and TNO.

The ADAM project is an on-going collaboration with no fixed project end date.

1.3.5 CAMP

The Crash Avoidance Metrics Partnership is a joint government and industry program and was formed by Ford Motor Company and General Motors Corporation in 1995. The aim of the Driver Workload Metrics project which is one part of the program is to develop practical and reliable driver workload performance metrics and evaluation procedures in order to be able to assess which advance driver interface tasks are appropriate to perform while driving (Deering, 2002).

The metrics project has instrumented a set of vehicles with the possibility to provide data for lane keeping, car following, eye glance behaviour and object and event detection performance. Also, digital video on driver's face, the interior and the road scene around the vehicle could be captured.

A set of tasks intended for subjects to perform while driving were developed. The tasks were all natural tasks but chosen as to span a range of visual-manual, auditory-vocal, spatial and verbal task types.

The Driver Workload Metrics project begun in 2001 and is planned to be finished in the end of 2004.

1.3.6 Response

Response (The integrated Approach of User, System, and Legal Perspective), (Becker, 2000) dealt with advanced driver assistance systems and had the following objectives:
- Analysis of the aspects of system safety and safety of usage.
- Conceptual checklist for system development.
- Analysis with regard to standardization process and type approval.
- Methods to analyze system comprehensibility.
- Experimental methods to analyze risk to usage safety.
- Analysis of the legal implications of the testing of Driver Assistance Systems.
- Analysis of the legal implications of the market introduction of Driver Assistance Systems.
- Recommendations for the functional specification of Driver Assistance System on the basis of the integrated view approach used in RESPONSE.

The project was active between 1998 and 2000.

1.3.7 ADVISORS

ADVISORS (Action for advanced Driver assistance and Vehicle control systems Implementation, Standardisation, Optimum use of the Road network and Safety) aimed at develop a methodology to assess the effect and impact of different types of ADAS in terms of safety, efficiency and environmental performance of the road transport system. Several on test demonstrations were carried out. The aim was also to develop implementation scenarios in order to help introducing appropriate ADAS (Heijer et al., 2003).

The main achievements in the projects were:
- The development of an integrated and common ADAS assessment methodology.
- The results of assessment of a set of ADAS on road safety, driver comfort and network efficiency.

- A multi-criteria analysis (MCA) on a set of ADAS-types revealed a ranking of ADAS for which relevant criteria were considered most favorable.
- A risk analysis method based on FMEA was developed and applied on behavioral, legal and organizational risks of a set of ADAS (ISA, ACC, Fleet Management & Navigation Systems, and DMS).
- Identification of a set of multidimensional future priority scenarios for ADAS developments. ACC on the motorway, intervening ISA in urban areas, a warning type DMS for professional drivers and Integrated ADAS are chosen.
- Identification of major legal, institutional, socio-economic, financial, organizational and user acceptance ADAS implementation problems.
- Formulation of implementation strategies to overcome implementation barriers for priority future.
- Dissemination of the results through various channels and production of user-friendly terminology.

The project lasted between 2000 and 2003.

1.4 System development

The projects below have a quite different aim than pure assessment of ADAS or IVIS. The goal has been to develop a system or some sort of demonstrator vehicle. Within the system development processes assessment techniques have been deployed. Some projects have been developing both an assessment methodology as well as developed systems (e.g. ADVISORS).

1.4.1 SAVE-IT

SAfety VEhicle using adaptive Interface Technology (SAVE-IT) is partly sponsored by NHTSA. The partners involved are Delphi Delco Electronics (DDE) Inc., University of Michigan Transportation Research Institute (UMTRI), University of Iowa, Seeing Machines, Inc., Ford, and General Motors (GM).

The main objectives of the first funded part (Research and Concept Development) of the project are according to the SAVE-IT website:
- Conduct comprehensive human factors research to derive distraction and workload measures for use adaptive interfaces.
- Identify scalable system concepts and sensing technologies for further research to follow the SAVE-IT program.
- Advance the deployment of adaptive interface technology countermeasures for distraction related crashes.

The second phase (Data Fusion, System Integration and Evaluation) in SAVE-IT consist of the following objectives:
- Develop system operational performance requirements and guidelines for adaptive interface conventions.
- Develop and apply evaluation procedures for assessment of safety benefits.
- Enhance collision warning effectiveness by optimizing alarm onset based on driver's workload or distraction.
- Provide the public with documentation on human factors research findings for performance and standardization development .

The project begun in 2003 and is planned to end in 2005.

1.4.2 IN-ARTE

The aim of the IN-ARTE (Integration of Navigation and Anticollision for Rural Traffic Environment) project (1998-2001) was to improve traffic safety in rural environments by means of an integrated driver support system. For example, the project aimed to develop and evaluate this integrated system in order to identify the advantages and disadvantages of the system in terms of safety, usability and behavioral change (ref. IN-ARTE web page).

1.4.3 Comunicar

Comunicar started 2000 and was finished in 2003. The main aim of the project was to design, develop and test an easy-to-use on-vehicle multimedia HMI. The system monitored the workload of the driver, the traffic environment and managed the communicative exchange between system and driver. The HMI was integrated in two vehicle demonstrators (a city and an upper class car).

1.5 Standards and guidelines on Transport Telematics

Standardization work for transport telematics is conducted in a range of areas such as the international organizations ISO and IEC as well as the European organization CEN, CENELEC and ETSI. Below are some of the main initiatives:

- CEN/TC278 Road transport and traffic telematics.
- ISO/TC 204 Transport information and control systems.
- CENELEC/TC 214 Electro technical aspects of surface transport systems.
- CEN/TC224 Machine readable cards, related device interfaces and operations.
- ISO/TC22 Road vehicles.
- CENELEC/TC 9X Electrical and Electronic Applications for Railways.

For AIDE the Human-machine interface standards for road vehicles developed within ISO/TC22/SC13/WG8 will be of special interest.

Also, the ISO 15005 'Dialogue management principles', PrEN ISO 15008-2 'Road vehicles - Ergonomic aspects - Part 2: Evaluation', PrEN ISO 16951 'Road - Procedure for determining priority of on board messages presented to drivers' and PrEN ISO 17287 'Road vehicles - Procedure for assessing suitability for use while driving' should be interesting for AIDE.

There are also a range of guidelines and checklists, both developed within international standardization organizations as well as within universities and national institutes etc. Below is a list of some examples of documents which are of special interest for AIDE:
- Commission of the European Communities (1999). Statement of Principles on Human Machine Interface (HMI) for In-Vehicle Information and Communication Systems ("EU Principles), (Annex 1 to Commission Recommendation of 21 December 1999 on safe and efficient in-vehicle information and communication systems: A European statement of principles on human machine interface) Brussels, Belgium: European Union.
- Alliance of Automobile Manufacturers (2003 June 17; Version 3). Statement of Principles, Criteria and Verification Procedures on Driver Interactions with Advanced In-Vehicle Information and Communication Systems, Washington, D.C.: Alliance of Automobile Manufacturers.
- Stevens, A., Board, P., A., and Quimby, A. (1999). A Safety Checklist for the Assessment of in-Vehicle Information Systems: Scoring Proforma (Project Report PA3536-A/99), Crowthorne, UK: Transport Research Laboratory.

Standardization work in the field of IVIS and ADAS assessment is presented in the separate AIDE Deliverable D 4.3.1 (Shindhelm et. al. 2004).

2   Driving performance

The following chapter covers driving performance metrics and methods.

2.1   Introduction

Driving performance metrics are used for a wide range of applications, including driver drowsiness and/or drug influence detection, driver training, road infrastructure evaluation and the assessment of effects of in-vehicle systems. The present review focuses exclusively on the latter application. As the common denominator of driving performance metrics is that they score the performance of the driving task, it is useful to start with a short discussion on how the driving task can be characterised and defined.

2.1.1   The driving task

The (primary) driving task can be described on different levels of abstraction. Based on Rasmussen's (e.g. 1986) hierarchical taxonomy of human behaviour, Michon (1985), proposed a widely adopted scheme where the driving task is considered on strategic, tactical and operational levels. The strategic level concerns behaviours directed towards more high-level goals, e.g. reaching a destination in time. The tactical level concerns behaviour on a shorter time frame, e.g. selecting headway and deciding when to change lane. Finally, the operational level concerns the moment to moment control of the vehicle. Table 2.1 below (loosely adopted from Green, 1995) summarises these three levels. The present review mainly focuses on metrics that quantify performance on the tactical and operation levels.

Table 2.1 The primary driving task level distinctions proposed by Michon (1985).

Driving task level (Michon)	Corresponding driver behaviour level (Rasmussen)	Example tasks	Example driver performance metrics
Strategic	Knowledge Based	Planning trip	Navigation errors
Tactical	Rule based	Selecting headway	Speed, headway
Operational	Skill based	Maintaining lateral control	Steering wheel movements, lane keeping metrics, pedal movements

2.1.2   Driving performance and safety

The basic assumption underlying the application of driver performance metrics to in-vehicle system evaluation is that they are directly related to accident risk. For example, it can be argued that even if degraded performance does not lead to a crash, it "does increase the likelihood of a crash by reducing safe driving tolerances and the ability to recover in the event of unexpected events" (Wierwille et al., 1996, p.1:12). However, the precise relation between specific performance metrics and accident probability is generally not well understood. A good discussion on this topic can be found in Dingus (1995).

2.1.3   Application of performance metrics to in-vehicle system evaluation

In-vehicle systems could be roughly divided into those that support the driving task (Advanced Driver Assistance Systems, or ADAS) and those that impose additional (secondary tasks) that may interfere with the driving task (In-vehicle Information Systems, or IVIS). However, these categories are seldom precisely defined and some systems, such as navigations aids, could be viewed as supporting the

driving task on high level but at the same time distracting the driver and thus interfering with lower-level driving control. For present purposes, ADAS and IVIS are defined as:

ADAS: Systems that interact with the driver with the main purpose of supporting the driving task on the tactical and operational levels (as defined in Table 2.1)

IVIS: Systems that interact with the driver and induce tasks that are not directly related to the driving task on the tactical and operational levels. Such additional tasks are called secondary tasks and may interfere with the primary task.

According to this taxonomy, a navigation aid would qualify as an IVIS, while Intelligent Speed Adaptation would be categorised as an ADAS.

As pointed out by Dingus (1995), the purpose of an in-vehicle system should determine the criteria and hypotheses guiding its evaluation. For example, these hypotheses can be expected to differ for the evaluation of potential safety benefits of an ADAS compared to the evaluation of safety risks of an IVIS. Based on the above definitions of ADAS and IVIS, two applications of driving performance metrics in vehicle system evaluation may be distinguished:

- Determining the potential performance enhancements and/or unexpected behavioural effects resulting from ADAS support.
- Determining the potential driving performance decrements resulting from the dual task of using IVIS use while driving.

A third type of application, of key relevance for the AIDE project, concerns the evaluation of integrated ADAS/IVIS solutions, such as the proposed AIDE system (to be developed in AIDE SP3). The general hypothesis tested when evaluating AIDE-type systems is that the integration and adaptation provided by the systems yields better driving performance, and hence increased safety, than non-integrated ADAS and IVIS solutions. Thus, performance metrics are key dependent variables in the evaluation of AIDE-type systems. For existing examples of this type of evaluation studies, see the evaluations of the GIDS (Michon, 1993) and the COMUNICAR (Hoedemaeker et al., 2004) systems.

2.1.4 Structure of this chapter

The present review of driving performance measures is structured based on the concrete physical and behavioural quantities that are measured (e.g. headway, lane keeping variation, etc.) rather than the theoretical constructs the metrics represent (e.g. workload, inattention, situation awareness). The metrics are sorted under the following main headlines:

- *Longitudinal control metrics*
  - Speed
  - Vehicle following
  - Pedal movement
- *Lateral control metrics*
  - Steering wheel movement
  - Lane keeping
  - Heading
- *Event detection metrics*
- *Combined control and event detection metrics*

For each type of metric a short introduction is provided, which explains the basic rationale for using it in ADAS/IVIS evaluation and how it can be measured, followed by more detailed definitions of the metrics. Finally, some examples of the application of the metrics in the context of ADAS and IVIS evaluation are reviewed.

The present review is partly based on existing similar reviews; in particular Green (1995), Wierwille et al. (1996) and the reviews performed in the RoadSense and HASTE FP5 EU projects (Nathan, 2004 and Roskam et. al. 2002 respectively).

2.2 Longitudinal control metrics

The longitudinal control metrics most commonly used in ADAS/IVIS evaluation can be grouped into three major categories: speed, vehicle following and pedal movement metrics.

2.2.1 Speed

There is abundant evidence of a strong relation between speed and safety on the collective (traffic) level (e.g. Finch et al. 1994; Nilsson, 1994). Speed is generally key dependent variable in all ADAS evaluation studies, especially the investigation of the potential safety benefits of intelligent speed adaptation (ISA) systems. However, speed metrics are often also commonly used in IVIS evaluation.

Speed metrics are generally simple and straightforward, both to measure and to compute, and can often be obtained directly from the vehicle's CAN-bus.

2.2.1.1 Metrics

A large number of speed metrics could be computed. The most commonly used are:

- *Mean speed:* The average of the longitudinal speed relative to the road surface.
- *Standard deviation/variance of speed*
- *Maximum speed:* The single maximum speed value.

Speed variation metrics, e.g. standard deviation, are strongly dependent on task duration. Thus, when tasks of different duration are compared, the metric has to be standardised. A method for this has been developed in HASTE and results will be reported in the WP3 deliverable by the end of 2004.

A slightly more complex speed metric, used e.g. in the HASTE project (Östlund et al., 2004), is *speed change*. Here, a linear function is fitted to the speed signal in a certain interval (corresponding e.g. to the operation of an IVIS) by means of least squares. The speed change value is then computed as the difference between the start and end value of the fitted line (see Figure 2.1). The purpose of this metric is to capture the speed change only while filtering out other variance in the speed signal. A basic problem with this metric is that it assumes that the speed change is approximately linear. An example when this does not apply is a fast speed drop followed by a slow speed increase which ends up on the initial speed. In this case, the speed change measure will indicate a speed increase.
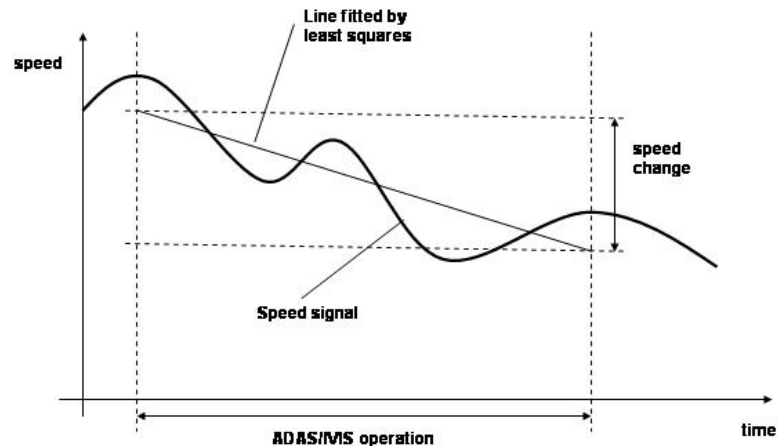
Figure 2.1 Illustration of the speed change metric

2.2.1.2 Application of speed metrics

In ADAS evaluation, a key hypothesis is whether support system changes speed during medium to long term use. Naturally, speed is the key metric when evaluating Intelligent Speed Adaptation (ISA) systems (e.g. Carsten and Comte, 1997). It has also been a common dependent measure in ACC evaluation studies. For example, results from the ADVISORS EU project (Nilsson et al., 2000) showed that the strongest effect of the inter-urban ACC is that it reduces the maximum speed, both in motorway and rural road environments. Speed variation was also affected (decreased by ACC). However, in general the speed effects of ACC are quite contrasting (Saad et al., 2004). For a more comprehensive review of these and other ADAS studies using speed as a dependent variable, see the AIDE SP1 Review of Behavioural Effects (Saad et al., 2004).

Speed measures have also been used for assessing the effect of secondary task activity, e.g. IVIS use. The effect usually found is a speed reduction during the secondary task (e.g. Curry, Hieatt and Wilde, 1975; Antin et al., 1990), which could be interpreted as a behavioural adaptation in order to reduce the primary task demand (and allocate resources to the secondary task). This effect has mainly been found for visually loading (as opposed to purely cognitive) tasks (Östlund et al., 2004).

Speed metrics have also been used in evaluation of adaptive integrated interfaces, e.g. the COMUNICAR Information Manager (IM) (Hoedemaeker et al., 2004) which prioritises and reschedules IVIS information in demanding situations. In this study, conducted in parallel in Sweden and Italy, it was found that Swedish drivers reduced their speed when driving without IM and maintained speed when the IM was on (no speed reduction was found for Italian drivers). This indicates that the IM did reduce drivers' workload, making speed compensation unnecessary. However, it does not follow that safety was improved, since at least the Swedish drivers, seemed to compensate for the high demands.

2.2.2 Vehicle following metrics

The headway to a lead vehicle is an indication of the longitudinal safety margins accepted by the driver. Thus, violation of these safety margins, as indicated e.g. by small following distances, could be interpreted as degraded driving performance with potentially negative safety consequences. Headway metrics are often included as primary dependent measure in the evaluation of long-term ADAS effects, especially Adaptive Cruise Control. They are less common, but nevertheless used, in IVIS evaluation.

In real-world experiments, headway could be measured by means of radar, laser or video sensors. There are a wide variety of different headway metrics. Most existing metrics can be divided into (1)

distance- and (2) time-based metrics, where the latter can be further divided into (a) time-headway metrics and (b) time-to-collision metrics.

The descriptions below are mainly based on definitions from the HASTE project (Östlund et al., 2004).

2.2.2.1 Metrics

2.2.2.1.1 Distance-based vehicle following metrics

Distance headway is defined as the average distance to the lead vehicle (from bumper to bumper). Large values (>~50 meters) are normally discarded. Possible summary statistic metrics include:

- *Mean distance headway:* The average distance headway.
- *Standard deviation of distance headway*
- *Minimum distance headway:* The minimum value of the distance headway signal. ´

Headway variation metrics are, like other variation metrics, dependent on task duration (see the corresponding discussion on speed variation metrics above).

2.2.2.1.2 Time-based vehicle following metrics

As mentioned above, the time-based metrics may either be based on (1) time headway or (2), time-to-collision.

Time headway is defined as the distance to the lead vehicle (from bumper to bumper) divided by the travel speed of the own vehicle. Large values (>~3 seconds) are discarded. Summary statistic metrics include:

- *Mean time headway*
- *Standard deviation of time headway*
- *Minimum time headway*

The second type of time-based vehicle following metric is based on time to collision (TTC). TTC represents the time until collision with an object (e.g. a lead vehicle) given the current trajectories and velocities of the own vehicle and the object. The TTC could thus be regarded as longitudinal time-based safety margins. There is abundant evidence that time-to-object information is used by humans and animals for guiding locomotion (Gibson, 1979), including driving (Gibson and Crooks, 1938; Lee, 1976). Time-to-collision has been suggested as a driving performance metric, e.g. by van der Horst and Godthelp (1989), who computed TTC from manual transcriptions of video recordings from traffic scenes. These authors suggested that only TTC values below 1.5 seconds should be regarded critical.

TTC to a lead vehicle is formally defined as the distance to the vehicle (bumper to bumper) divided by the speed difference to the lead vehicle. TTC is only defined if the distance between the vehicles decreases. Small (<~1 second) and large (>~15 seconds) values are discarded. Based on the TTC signal, a number of summary statistic metrics could be computed:

- *Minimum TTC:* The minimum values of the time-to-collision (TTC) signal.
- *Mean of TTC local minima:* The mean of the local minima in the TTC signal (defined above).
- *Time Exposed TTC (TET):* The proportion of time of which the TTC is less than X seconds (X=4 seconds was used in HASTE; Östlund et al., 2004).

2.2.2.1.3 Other vehicle following metrics

Brookhuis, de Ward and Mulder (1994) has suggested a number of alternative vehicle following performance metrics, based on the ability to follow the speed profile of a lead car. They propose the following metrics:

- *Coherence:* A measure of the correlation between the speed profiles of the own and the lead car
- *Phase shift:* An index of the delay between the speed profiles
- *Modulus:* This quantifies the amplification between the two signals and identifies, for example, "overreactions" of the following vehicle.

2.2.2.2 Application of vehicle following metrics

Vehicle following metrics are typically applied when evaluating medium- to long-term effects of ADAS, in particular ACC. For example, Saad et. al. (1996) found that ACC reduces the frequency of short time headways. For ISA, Hjälmdahl et. al. (2004) found that long-term use of ISA increased time headway. For a more comprehensive review of these results, see the AIDE SP1 Review of Behavioural Effects (Saad et al., 2004). In a naturalistic study on behavioural effects of antilock brakes (ABS), Fosser et. al. (1997) showed that drivers with ABS adopted shorter headways than drivers without the system.

Vehicle following metrics have also been used for IVIS evaluation. Greenberg et al. (2003) found increased distance headway as a result of strongly visual secondary tasks (but not for more cognitive tasks), although this effect was found mainly for older drivers. Brookhuis et al. (1994) found effects of mobile phone use on one of their vehicle following metrics, the "phase shift" measure (see above). Lamble et al (1998), investigated vehicle following behavior as a function of display position, and found that time-to-collision in headway-closing situations decreased with display angle, particularly for vertical angles.

In the HASTE project, vehicle following metrics were applied to evaluation of "surrogate" IVIS (S-IVIS, visual and cognitive) (Östlund et. al., 2004). The general result from this study was that both distance and time headway increased during visual, but not cognitive, S-IVIS operation. In a few studies, indication of reduced headway control was found as well.

2.2.3 Pedal movement metrics

A few examples of performance metrics based on brake and accelerator pedal behaviour can be found in the literature. The assumption behind these measures is that reduced longitudinal control is reflected in abrupt and/or jerky braking/deceleration (Nygård, 1999). Brake pedal metrics have so far mainly been applied in conflict studies in the incident/accident analysis field, but could possibly be applied to ADAS/IVIS evaluation as well. Brake reaction time metrics are, however, very common. These are here classified as event detection metrics and reviewed in section 2.4.

A number of accelerator pedal metrics are listed by Wierwille et al., (1996), including accelerator standard deviation, reversals, accelerator releases and others. However, these are rarely used in actual studies and, as the authors point out, their safety relevance is quite unclear.

2.3 Lateral control indicators

The lateral control metrics reviewed here are grouped into three main categories: steering wheel metrics, lane keeping metrics and heading metrics.

2.3.1 Steering wheel metrics

Metrics based on steering wheel movement are very common in all types of driver performance assessment. The analysis and modelling of steering behaviour has a long history with strong links to research on manual tracking performance (e.g. Kelley, 1969; Wickens and Gopher, 1977). In the 1960's and 70's several influential steering models, based on control theory, were proposed (e.g. Weir and McRuer, 1968; Donges, 1978). In these types of models, lateral control is modelled in terms of a feedback control mechanism where steering wheel movement represents the driver's input which, via the vehicle dynamics, gets translated into lateral position and heading values that are controlled by the driver (the same types of models may be applied to longitudinal control). Although the models differ in detail, most of them model the steering task on two levels. The first level concerns the pre-view control based on predictions a few seconds ahead while the second level concerns moment-to-moment corrections to small disturbances, e.g. wind gusts. According to Weir and McRuer (1968), the first level is associated with lateral position control and the second with heading control. In a frequency analysis of steering movements during normal driving, McLean and Hoffman (1971) found two peaks in the frequency spectrum, which were hypothesised to correspond to these control levels. The largest (primary) peak occurred in the 0.1-0.2 Hz region for straight sections and at 0.15-0.3 Hz in circular sections. The smaller (secondary) peak occurred in the 0.35-0.6 Hz region.

When visual attention is diverted, e.g. during IVIS operation, heading errors build up, which are corrected by means of large and disruptive steering wheel movements. These patterns can be interpreted as a violation of the driver's subjectively chosen safety margins, and thus as reduced control ability.

An increase in small corrective steering wheel movements may also indicate increased *effort* spent on the lateral control task (e.g. McDonald and Hoffman, 1980), and often occur when the primary task demand increases (for example, the amount of small steering corrections generally increases when driving fast on a narrow road). In general, steering performance is strongly influenced by the nature of the primary task, e.g. speed, curvature and lane width (McLean and Hoffman, 1971, 1972), but also individual factors such as driving experience (e.g. Greenshields, 1963) and age (Liu et. al., 1999). Thus, the relation between driving demand, secondary task demand and steering performance is complex and still not fully understood.

Numerous steering performance metrics have been suggested to quantify the effects of attentional demand on steering behaviour, ranging from relatively simple computations such as standard deviation, to more advanced approaches like spectral analysis and steering entropy.

Steering wheel angle is relatively easy to measure by means of sensor mounted on the steering column. Most metrics require a rather high spatial resolution, normally < 0.5 degrees.

2.3.1.1 Metrics

The most common steering wheel movement metrics are

Standard deviation/variance of steering wheel angle
This is a simple, but still very common, evaluation metric. Due to its simplicity, it is included in almost all IVIS evaluation studies where steering angle is measured (e.g. Liu, Schreiner and Dingus, 1999). A basic problem with the measure is that it is sensitive to variation in the steering wheel data not related to IVIS demand, especially lower frequencies that are associated with road curvature and manoeuvres. Thus, a variety of more sophisticated measures have been suggested, and further reviewed below.

High frequency component (HFC) of steering wheel angle
A detailed analysis of steering performance can be achieved by means of spectral analysis. This involves transforming the steering signal to the frequency domain (by means of Fourier transform) and

analysing which frequency bands that are affected by different factors (e.g. IVIS load). McLean and Hoffman, 1975 found that the frequency content in the 0.35-0.6 Hz band is sensitive to variations in both primary and secondary task load. Thus, the power spectral density, i.e. the area under the spectral curve in the relevant frequency region, could be used as a steering performance metric. A somewhat simpler approach to compute the high-frequency component was adopted in HASTE, where the 0.3-0.6 range was filtered out using a band pass filter and the final metric was obtained by computing the standard deviation of the remaining signal (Östlund et al., 2004).

In most studies on steering frequency, focus has been on the 0-0.6 Hz area of the steering angle spectrum, which indeed has been found to be dominant frequency band for steering activity. If, however, the steering signal is studied in detail you find that reversals often are within higher frequencies. The amplitude is however rather small (often < 2 degrees), which may be the reason for higher frequencies not to be found in the frequency analysis of the steering activity. The human bandwidth in tracking tasks are higher than 0.6 Hz, and at least 2 Hz (e.g. Jex & McDonell, 1966), supporting that 0.6 Hz would be a too narrow upper limit. It should thus be considered to filter the steering signal at 2Hz instead of the 0.6 Hz used in HASTE.

Steering wheel reversal rate (SRR)
SSR is one of the most commonly used driving performance metrics. The metric represents the number of times that the steering wheel is reversed by a magnitude larger than a specific angle, or gap. The gap sizes reported in the literature varies between 0.5-10 degrees. McLean and Hoffman (1975) demonstrated that SSR (with gap sizes between 0.5 and 5 deg.) correlated strongly with the high frequency component measure (HFC) described above (specifically with the power spectral density at frequencies higher than 0.4 Hz). Thus, due to its simpler computation, it is more commonly used than the HFC. Different ways of computing the SSR are reported in the literature. For example, in HASTE local minima and maxima were identified by means of signal processing algorithms, and the differences between adjacent minima and maxima calculated. If the difference was larger or equal to the gap, the reversal (or more correctly, the peak) is counted. In HASTE, gap values of 1, 3, 5 and 7 degrees were compared. This method is further illustrated in Figure 2.2 below (see Östlund et al., 2004 for details).
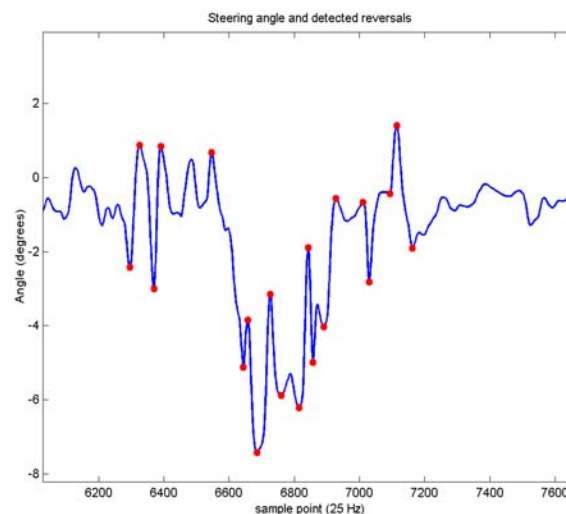

Figure 2.2 Example of steering wheel reversal rate calculation in HASTE (gap = 2 degrees).

An alternative, investigated by Malaterre (1994), is to count the number of reversals per distance unit rather than per time unit. This study showed that the distance-based version of SSR is more sensitive to driver distraction than SRR per time unit.

Steering wheel action rate (SAR)
This measure, used e.g. by Verwey (1991), is related to steering wheel reversal rate, but uses a velocity- rather than a position threshold. SAR is defined as the number of steering wheel movements per second faster than a threshold velocity. A similar measure, Rapid Steering Wheel Turns (RSWT) was used in the HASTE project (Östlund et al., 2004).

Steering entropy
The steering entropy metric, described in Nakayama et al. (1999) and Boer (2000), is based on the notion of the driver as a satisficer rather than an optimiser. That is, rather than minimising lateral position and/or heading angle, as suggested by the early steering models (e.g. Weir and McRuer), the driver seeks to keep within subjectively chosen safety margins. When the attentional demand increases, e.g. due to execution of a secondary task, corrective steering manoeuvres are required to stay within the safe boundaries. This results in less predictable steering behaviour, or equivalently, higher entropy. The steering entropy is calculated on the basis of prediction errors of steering signals. The predictions are obtained by applying a predictive filter on the steering signal. In Nakayama et al. (1999), the predictions were obtained by performing a second-order Taylor expansion using the samples at the three previous time steps, while in Boer (2000), an averaging filter was used for the prediction. In The entropy of the signal is then calculated on the basis of the distribution of these errors. This involves dividing the errors into a finite number of bins (ten bins were used in Boer (2000). The bin-ranges are obtained by calculating the error value α at the 90:th percentile of the *null distribution*, i.e. the error distribution obtained from a baseline condition (driving without a secondary task). The bin edges are then chosen as +/-(0, 0.5α, α, 2.5α and 5α). Normally, individual null-distributions are calculated for each subject, which implies a normalisation whereby the inter-individual variance is reduced. The proportions p_i , i=1, 2, ..., I, where I is the number of bins, is then calculated for the experimental data (driving with a secondary task). The entropy h of the signal for a given time-period is finally given by:

$$h = \sum_i p_i{}^I \log p_i .$$

Experimental results reported in Boer (2000) show that the metric is sensitive to both visual/manual and cognitive load. The metric also correlated strongly with the PDT (the Peripheral Detection Task) and subjective workload ratings.

Note that the baseline data used for computing the null distribution cannot be used for statistical comparison to experimental data since the data sets will then be dependent (which violates the assumption of e.g. ANOVA). Thus, the baseline data must be drawn from a different data set.

Other steering wheel metrics
Other steering performance metrics, less commonly used, include:
  o Peak (max) steering deflection;
  o Steering velocity standard deviation;
  o Number of steering holds (number of times that the steering velocity is close to zero);
  o Steering zero crossings.

See Wierwille et al. (1996) for an extensive list of possible steering performance metrics.

Another type of steering wheel metric, which not really fit into the lateral control category, is steering grip metrics. The driver's steering wheel gripping behaviour can be viewed as a driving performance metric, based on the assumption that a driver with both hands on the wheel has a higher degree of control on the vehicle and better prepared to cope with unexpected events.

Steering grip can be measured by existing sensors that have, for example, been used in the context of drowsiness detection.

Wierwille et al. (1996) propose the following steering grip metrics:
- *Hands-on-wheel occurrences:* The number of times that the driver places both hands on the wheel without changing hand positions.
- *Mean hands-on-wheel duration:* The mean length of time that the driver places both hands on the steering wheel without changing hand positions.
- *Total hands-on-wheel time:* The total time that both of the driver's hands are in contact with the rim or spokes of the steering wheel.

Moreover, *steering grip force* and *grip force variations* are two simple measures that have not yet been investigated. It may however be hypothesised that these indicators, similar to steering reversal rate, reflect steering effort and control. A technical issue that has to be solved for these measures to be feasible is that it has to be assured that the measured steering grip force is reliable; i.e. fast response, correct level, no signal drift. This issue was partly investigated for a specific steering grip force sensor within the AWAKE project. It was found that there was a signal drift due to heating. Also force variations higher than 3 Hz were mechanically filtered out in the sensor, but this is not a problem as hand pressure at the wheel are not changed rapidly in normal driving conditions.

2.3.1.2 Application of steering performance metrics

Steering wheel metrics are mainly used for investigating the effects of secondary task load induced by IVIS. Application to ADAS evaluation is uncommon (and it is not entirely clear what effects to expect). The most popular metrics are the standard deviation/variance and reversal rate.

In a series of studies, reviewed in McDonald and Hoffman (1980), Hoffman and colleagues used steering wheel reversal rate to study the effects of secondary tasks (the nature of the tasks, e.g. visual or cognitive, is not stated in the review) on steering behaviour. As mentioned above, the authors found quite complex relationships between primary task demands, secondary task demand and driver characteristics. In short, they found that the addition of a secondary task may lead to increased SSR when driving demand is low. On the other hand, when task demands are high (e.g. when driving fast in a narrow lane), the SSR decreases with the addition of the secondary task.

The authors hypothesise that these seemingly contradictory results can be explained by the following reasoning:

1. When primary driving task demands are low, the drivers use their spare resources to invest more effort in steering. The purpose of this is to compensate for the driving performance degradation imposed by the secondary task. This is reflected by increased SSR.

2. When primary driving task demands are high, and match or exceed performance capacity, drivers cannot invest more effort to cope with the additional task. In this, case, less attention is spent on the steering task and thee SSR is decreased.

Based on the above argument, McDonald and Hoffman suggest that frequency measures such as SSR "represent control effort, rather than an absolute measure of tracking performance" (p. 735).

Liu et al. (1999) applied steering wheel angle variance and reversal rate to the evaluation of different types of displays (visual, auditory and multimodal). Based on the reasoning of McDonald and Hoffman, they hypothesise that increased workload would lead to a *decrease* of the small, continuous, steering wheel movements used to correct for e.g. wind and roadway disturbances, but increase large steering corrections. Thus, they used a large gap value (10 degrees) for the SSR. The results showed that this SSR metric is sensitive to the complexity of visual displays, especially for older drivers.

In the HASTE project (Östlund et al, 2004), the effects of two artificial (or "surrogate" S-) IVIS tasks, one purely visual and one purely auditory/cognitive (further described in Merat, 2003), on a number of steering wheel metrics (standard deviation, high frequency component, SRR with gap sizes 1, 3, 5 and 7, RSWT and steering entropy) were investigated. The general result was that the *visual* S-IVIS had strong effects on *all* metrics (including all gap size variants of the SRR), which was interpreted as an indication of reduced control performance. The auditory/cognitive task, on the other hand, generally only led to an increase in small corrective steering wheel movements (SRR with gap size 1-3, high frequency component at 0.3-0.6 Hz, steering entropy and rapid steering wheel movement). This was also accompanied by *reduced* lateral position variance and increased gaze concentration towards the road centre.

As the above examples show, increased steering activity can be associated with both increased and reduced lane keeping performance. Further research is needed to clarify these issues and provide a stronger theoretical basis for inclusion of steering wheel metrics in IVIS evaluation regimes.

2.3.2 Lane keeping metrics

Together with steering wheel movement, lane keeping metrics are the most commonly used lateral control performance metrics. The rationale is that increased lane weaving and/or lane exceedences indicate degraded control and, hence, increased accident probability. Like the headway metrics reviewed above, the lane keeping metrics could be either distance-based or time-based. The distance-based measures (e.g. standard deviation of lane position), are generally easier to compute and, thus, more commonly used than time-based metrics (such as Time to Line Crossing, TLC).

Lane position measures are strongly influenced by traffic conditions, e.g. road curvature and overtaking (Östlund et al., 2004) and require strict control of experimental conditions.

In the real world lane position is normally measured by means of video-based lane-tracking systems, many of which are commercially available. For present purposes, a minimum accuracy of about +/- 5 cm is required.

2.3.2.1 Metrics

2.3.2.1.1 Distance-based lane keeping metrics

The most common distance-based lane keeping metrics are:
- *Mean lane position:* The mean lane position is defined as the mean distance between a reference point on the vehicle and an arbitrary position in the lane (normally one of the lane boundaries or the lane centre).
- *Standard deviation/variance of lane position:* This is one of the most common performance metrics. Its popularity is probably due to its high face validity and computational simplicity. The metric has been shown to be relatively independent of speed (Godthelp, Milgram and Blaauw, 1984). As with other variation metrics, it is strongly dependent on task duration (see the comment in section 2.2.1.1 above).
- *Lane exceedences:* Several related lane exceedence metrics exist. The most common metric seems to be LANEX, defined as the proportion of a time any part of the vehicle is outside the lane boundary (Östlund et al., 2004). An alternative is to count the number of *times* that the vehicle exceeds lane boundaries (Wierwille et al. 1996). Lane exceedence metrics have higher face validity than lane variation measures (e.g. standard deviation) as safety indicators. However, they may be insensitive to small shifts in workload/distraction (due to the implicit threshold). Another variant is to count the number of *major* lane deviations, where a major

lane deviation is defined as a situation where any part of the vehicle exceeds the lane by more than half of the vehicle width. This metric was employed by Liu et al. (1999) and in RoadSense (Nathan, 2004).

Wierwille et al. (1996) lists a number of other possible distance-based lane keeping metrics, including:
- Lane RMS deviation;
- Peak lane deviation;
- Mean lane exceedence duration.

2.3.2.1.2 Time-based lane keeping metrics

The lateral counterpart to the time-to-collision (TTC) metric (described above in the section on vehicle following metrics) is the time-to-line crossing (TLC), first developed by Godthelp and Konings (1981). TLC is defined as the time to reach the lane marking assuming fixed steering angle and constant speed. As mentioned above, time-to-object information has been demonstrated to be the main perceptual factor guiding locomotion in humans and animals. In driving, TLC could be regarded as reflecting the driving strategy, or more precisely, the time-based lateral safety margins adopted by the driver. This interpretation is supported by results in Godthelp, Milgram and Blaauw (1984), demonstrating that the TLC correlates strongly to driver's self-chosen occlusion time. Too small TLC values are thus strong indicators of reduced lateral control (where "too small" is determined relative the subjectively chosen safety margins).

Compared to the distance-based metrics described above, TLC is substantially more difficult to compute, especially in the real world. The calculation, involves calculation of the predicted path which can be computed from vehicle speed, steering wheel angle, heading angle and lateral position. The method described in Godthelp et al. (1984) assumed a straight road. An alternative method, taking road curvature into account was proposed by van Winsum and Godthelp (1996). In the real world, approximations are often used. In the HASTE project, an approximation was used where a linear path towards the lane boundary was assumed (Östlund et al., 2004).

Based on the TLC computation, several different summary statistic metrics can be computed, e.g.:
- *Median TLC:* Used in Godthelp et al., (1984).
- *15% level TLC:* The 15 percentile of the TLC values, i.e. 15% of the TLC values are below this value (Godthelp et al., 1984).

In HASTE (Östlund et al., 2004), the TLC metrics were based on minimum TLC values, defined as the min TLC within a TLC waveform. TLC values higher than 20 seconds were ignored, as well as TLC waveforms of duration less than one second. The identification of TLC min values is illustrated in Figure 2.3 below.
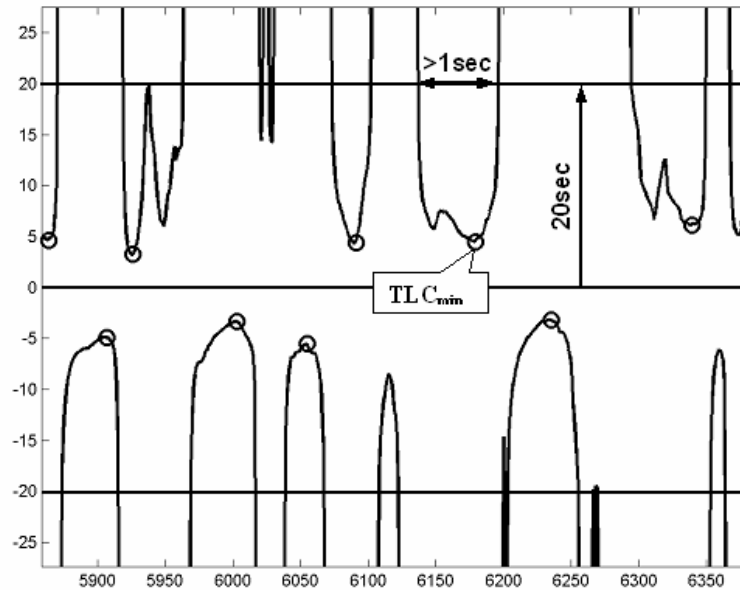
Figure 2.3 Principles for identifying relevant TLC_{min} values in HASTE (adopted from Östlund et al., 2004)

The TLC metrics used in HASTE were:
- *Minimum TLC:* The minimum TLC value in the relevant data segment
- *Mean value of the min TLC values:* Average of the identified TLC minima
- *The proportion of TLC min values < X s:* In HASTE, a threshold of 1 s. was used.

2.3.2.2 Application of lane keeping metrics

Lane keeping metrics have mainly been used for two purposes:

(1) Assessing potential positive/negative effects of lane departure warning and lane keeping aid (active steering) systems.

(2) Evaluating the potential lateral control decrement induced by secondary tasks (e.g. IVIS)

An example of the former was the evaluation of the lane-departure warning sub-function in the IN-ARTE driving support system (developed in the IN-ARTE EU-funded project). Using standard deviation of lane position as the main dependent measure, it was found that lane keeping was significantly improved by the system (Burns, 2001).

Lane keeping has also been found to be influenced by other ADAS. For example, in ADVISORS, it was found that the use of an ACC/Stop&Go system increased lane position standard deviation (Nilsson et al., 2000). The tentative explanation is a decrease in attention on the driving task resulting from the automation.

Lane keeping metrics are almost always included in IVIS evaluation studies, especially lane position standard deviation/variance and number/proportion of lane exists. Many studies have demonstrated a strong relationship between visual demand and lane keeping performance (e.g. Farber et al., 2000). Greenberg et al. (2003) found that a strongly visual task (retrieving a voice mail on a hand-held phone) led to a relatively high number of lane violations. In HASTE (Östlund et al., 2004), it was found that the visual surrogate IVIS task induced strongly degraded lane keeping, as indicated by several measures. The sensitivity of these measures was strongly influenced by lane width.

For cognitive load, many studies have found a null effect on lateral control performance (e.g. Alm and Nilsson, 1995, who investigated the effect of mobile phone use). However, in HASTE, it was found that the auditory/cognitive S-IVIS task significantly *improved* lane keeping performance, i.e. standard deviation of lateral position decreased during cognitive load (Östlund et al., 2004).

2.3.3 Heading metrics

Heading metrics are quite rare, but have been used in some studies. Heading angle is easy to obtain in a simulator but more difficult to measure in the real world, at least relative to real-world coordinates. This can be achieved by means of some video-based lane trackers.

Greenberg et al. (2003) used the 90 percentile of the heading error (HE90) as a measure of lane keeping performance when investigating the effects of using in-vehicle systems e.g. hand-held and hands-free phones). They found the measure to be influenced by secondary tasks with a visual component. However, cognitive tasks did not seem to influence heading error in any direction.

2.4 *Event detection metrics*

Signal detection performance metrics are abundant in psychological research. In the driving domain, a wide variety of event detection metrics and techniques have been developed, which recently have gained increased popularity as ADAS/IVIS evaluation metrics. Event detection is obviously strongly related to crash probability, and thus perhaps the performance metric class with the strongest *prima facie* safety relevance.

Detection performance can be measured to stimuli that are more or less relevant to the primary task. A typical example of a driving-related detection task is the detection of braking lead vehicles or suddenly appearing pedestrians. On the other side of the spectrum there are metrics that measure detection of artificial stimuli like blinking LEDs inside the vehicle (such as in the Peripheral Detection Task, reviewed in 6.2) which could be viewed as secondary detection tasks. In the literature, the distinction between primary and secondary event detection tasks is not always clear. This is of course a matter of terminology, and boils down to how one defines the primary and secondary tasks. For present purposes, the distinction by Wierwille et al. (1996) between *embedded* and *un-embedded* tasks will be used. The former refers to a task "which is a naturally occurring component of the primary driving task" while an un-embedded task refers to "an extraneous task that has no natural place in the driver's task ensemble" (p.1:15). The present section only deals with embedded detection tasks, while un-embedded tasks are reviewed in chapter 6. Embedded detection tasks are naturally difficult to implement in the real world and are generally constrained to simulated driving.

Event detection performance is influenced by a variety of factors, including the stimulus modality (visual stimuli are most common in embedded tasks) and intensity, response modality, stimulus-response compatibility and expectancy (see Wickens (1992) for a review of relevant literature). With respect to ADAS/IVIS evaluation, the last factor is clearly the most difficult to control for.

There exists a wide variety of embedded event detection metrics, using different combinations of stimuli and response types. Traditionally, the most common approach is to use lead vehicles or other on-road obstacles as stimuli and braking as the response modality. See Green (2000) for a thorough review of the effects of attentional demand on brake reaction time to external events.

Other possibilities are to use accelerator release (Wierwille et al. 1996) or steering avoidance (e.g. Summala, 1981) as response modalities. The metrics listed below, based on Wierwille et al. (1996), include only the general types of metrics that can be used in event detection evaluation methods (regardless of the stimulus and response types).

Another approach to event detection performance measurement is the change blindness paradigm. This is based on the observation that, under a wide variety of conditions, humans can be amazingly blind to changes and failing to see changes even when they are large, repeatedly made and anticipated (Simons, 2000). Change detection can be severely impaired if the changes occur simultaneously with a brief visual disruption such as saccades, blinks, splats (mud splashes), gaps, shifts, occlusions, or cuts (op. cit). Thus, change detection could be used as a performance metric in IVIS evaluation.

2.4.1 Metrics

2.4.1.1 Response time

Wierwille et al. (1996) distinguishes two types of response time calculations. The first type is used when the onset of the stimulus is clearly defined. The metric is defined as the time from presentation of a specified stimulus (with specified start time) to the time that the driver responds correctly, either verbally or with appropriate hand or foot motion.

The second type can be used when the stimulus is at a fixed location in the road scene and is defined as the time from correct driver response (verbal, manual, or pedal) until the drive's vehicle is alongside the stimulus.

2.4.1.2 Response distance

The distance of the driver from the stimulus when the driver responds correctly, either verbally or with appropriate hand or foot motion.

2.4.1.3 Errors of omission

The number of times that the driver fails to respond to a specified stimulus presentation.

2.4.1.4 Errors of commission

The number of times that the driver responds incorrectly to a specified stimulus presentation.

2.4.2 Application of event detection metrics

Event detection metrics are commonly used for evaluating the effects of visual and cognitive secondary tasks. It seems to be one of the few performance metrics that is reliably sensitive to both types of distraction. When evaluating cognitive tasks, event detection metrics are usually the primary dependent measure. For example, Alm and Nilsson (1995) found significantly reduced detection performance as a result of talking in the mobile phone. Similarly, McKnight and McKnight (1993) found that talking on the phone reduced detection rate and response time to a number of naturally occurring events (e.g. red lights, braking vehicles etc.). Lee et al. (2001) investigated the effects of a hands-free voice-based email system, using lead vehicle braking as the detection event (and driver braking as the response measure), and found that the use of the email system increased reaction time by 30%. The change blindness approach has recently been used to evaluate the effects of mobile phone use (Strayer, in press).

Event detection measures are also commonly applied to the evaluation of visual load. For example, when evaluating display positioning, Summala et al (1998) found that brake reaction times to lead vehicle stimuli were substantially impaired when gaze was directed to the speedometer and centre console positions. In a study in the Ford VIRTEX moving base simulator, Greenberg et al. (2003) used severe lane deviations by front and rear vehicles found as the events to which the drivers should

respond using the turn indicator. They found that both visual and cognitive tasks (e.g. using hand held and hands-free phone functions) significantly reduced the ability to detect these events.

Finally, event detection metrics have been used to assess the potentially negative effects of ADAS automation on driver behaviour. One example is Nilsson (1995), who found, in a simulator study, that use of ACC lead to later braking reactions to a stationary traffic queue (in fact, 40% of the subjects with ACC crashed into the queue).

2.5 *Combined control and event detection metrics: The Lane Change Task*

The Lane Change Task is a relatively new approach to IVIS evaluation, developed in the German national project ADAM. The basic task given to the subjects is to repeatedly perform lane changes on command (from roadside signs) while at the same time operating the IVIS to be assessed. The effect of the IVIS is evaluated in terms of the level of degradation of the lane change quality (where lane change performance is scored relative to a normative model). The difference between the actual and the optimal lane change path is influenced by the ability to detect and respond to the posted lane change commands (signs) as well as the ability to maintain lateral control. The metric could thus be regarded as a combined lateral control and event detection metric. A general description can be found in Mattes (2003). The LCT can be implemented in different set-ups, from relatively simple desktop simulations to more hi-fidelity driving simulators.

Factors associated with the driving simulation setup, the test track and analysis methods are carefully specified, in order to enable cross-site comparison of results. Thus, the LCT should rather be viewed as a general test methodology than merely a performance metric. The LCT method is designed to be cost-effective and easy to use, in order to enable wide scale adoption by the automotive industry. The method is currently subject to standardisation efforts within ISO TC22/SC13 WG8 (ISO, 2004).

2.5.1 Details of the LCT performance metric

The present description is based on Mattes (2003). However, these specifications are still under discussion in the ISO task force and may be amended in a possible future standard.

The LCT simulated road is 3000 m long with three lanes. The subjects are instructed by signs on the roadside to change lane on average every 150 meter (see Figure 2.4). The subjects perform a specific secondary task (or different variants of the task) throughout the 3000 meter drive. In order to avoid any effects of speed, the speed is fixed to 60 km/h (reached when the pedal is pressed down to maximum).

Figure 2.4 The simulated road with the signs commanding a lane change.

The LCT performance metric is computed from the deviation between a normative model (ideal path) and the actual path, as illustrated in Figure 2.5.
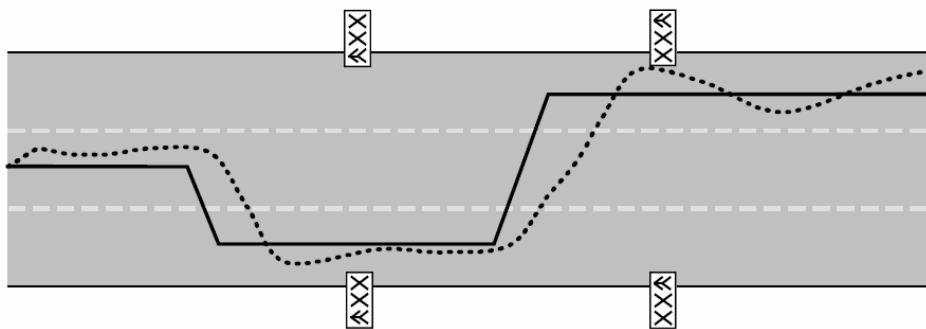
Figure 2.5 Illustration of LCT scoring based on the normative model.

2.5.2 Application

The LCT measures workload and distraction and is thus intended for IVIS evaluation only. So far, studies using the LCT have mainly focused on validating the method and comparing it against other workload/distraction metrics. As reported by Mattes (2003), the LCT is sensitive to a variety of secondary tasks, both visual and cognitive, including talking on the phone, unwrapping sweets, adjusting stereo sound volume and operate the navigation system. The metric correlated with longitudinal control performance (r=0.54), lateral control performance (0.68), glance behaviour (r=0.58) and subjective workload (r=0.87; however, it is not stated exactly which metrics were used in these comparisons).

2.6 Summary and conclusions

2.6.1 Summary

This section reviewed the most common driver performance metrics applied to ADAS and IVIS evaluation. The metrics were grouped into the four main categories longitudinal control, lateral control, event detection and combined metrics.

The most common types of longitudinal control metrics used for ADAS and IVIS evaluation are speed and headway. The headway metrics can be divided into distance based (e.g. distance headway) and time-based (e.g. time headway and TTC). Most of the longitudinal control metrics are computed from summary statistics such as the mean, standard deviation, percentile, max, min, etc. A main application is the assessment of medium to long-term behavioural effects of ADAS with the goal of supporting speed and headway maintenance (ISA and ACC). In this case, reduced speed and increased headway is generally interpreted in terms of increased safety. Longitudinal control metrics are also commonly applied in IVIS evaluation, where the most common effects are reduced speed and increased headway. However, in this case, this can be interpreted as a compensation for increased attentional demand (which seems to be found mainly for visually loading IVIS). It is not entirely clear how these compensatory effects should be interpreted in terms of safety; although it is positive that the driver is able to actively manage the increased load, a radical speed reduction may lead to traffic conflicts e.g. with rear vehicles. In sum, longitudinal control metrics are more naturally applied to the ADAS than IVIS evaluation. In some IVIS evaluation approaches, e.g. the LCT, speed is not included as a dependent measure, but rather controlled for in the experimental set-up.

Like headway metrics, lane keeping metrics are either distance-based (e.g. standard deviation of lane position) or time-based (TLC-based metrics). Most metrics are computed from summary statistics .Lane keeping metrics are most common in IVIS evaluation, where reduced control performance has been demonstrated to correlate strongly with visual IVIS load (Farber et al., 2000). Although, TLC metrics are attractive from a theoretical standpoint, the simpler distance-based metrics (e.g. lane position standard deviation and LANEX) are more feasible in practice, and seem to give similar results. Lane keeping metrics, however, do not appear to be very useful for assessment of purely cognitively loading tasks, such as talking on the mobile phone. Rather, there is strong evidence that cognitive tasks lead to increased lane keeping performance (reduced variance; Östlund et al., 2004). This effect is probably related to an increased gaze concentration towards the road centre although a detailed explanation for this phenomenon requires further research. For ADAS, lane keeping metrics are naturally used to assess the benefits of lane keeping support systems.

Steering wheel metrics are generally more involved than the other control metrics, and are based on signal processing (e.g. filtering and spectral analysis) rather than summary statistics. Steering wheel metrics are generally considered part of the "standard" toolkit in IVIS evaluation. However, as the present review shows, increased steering activity could be associated with both increased and reduced lateral control (and hence safety). For example, in HASTE, cognitive loading tasks led to increased steering activity but reduced lane position variation, while visually loading tasks led to increased steering activity and increased lane position variation (Östlund et. al., 2004). Thus, the safety interpretation of the steering wheel metrics is not always straightforward. One way to distinguish these effects is to filter out the large steering corrections e.g. using a reversal threshold larger than 10 degrees (e.g. Liu, Schreiner and Dingus, 1999). However, there does not seem to be any consensus in the literature on the details of this (e.g. reversal thresholds).

Event detection metrics have been shown to be sensitive to both visual and cognitive load and are strongly valid safety indicators. In fact, this seems to the only type of performance metric suitable to assess safety effects of purely cognitive tasks. Event detection metrics have also been used to test the hypothesis that ADAS (e.g. ACC) may reduce drivers' ability to react to unexpected hazards (Nilsson, 1995). A key problem with these metrics is that they are naturally difficult to implement in real world settings (at least embedded detection tasks). Another problem is to control for expectancy effects.

All the performance metrics reviewed, especially the control metrics, seem to be sensitive to scenario parameters (e.g. speed, road width, vehicle type, traffic density etc.), but also to driver characteristics (in particular age). Thus, these parameters must be carefully controlled in any IVIS/ADAS validation experiment using driving performance as dependent variables.

The Lane Change Task (LCT), which represents a combined lateral control and event detection metric, is a promising approach to IVIS evaluation. Early results show that it is sensitive to both visual and cognitive distraction. An important advantage of the LCT is the standardized test scenarios, which will enable direct comparison between studies and repeatability of results. Moreover, its relative simplicity and low cost will facilitate wide adoption by the industry. However, as with other control performance metrics, the safety interpretation of the LCT results is not as straightforward as for "pure" event detection metrics. Moreover, in the current set-up, it is not possible to determine whether a decrement in LCT performance is due to reduced detection performance, reduced control performance or both.

2.6.2 Conclusions and recommendations for further research

As is clear from the review, ADAS and IVIS may have a wide range of behavioural effects and there exist many different methods and metrics that can be used to quantify them. There is a great diversity in specific driving performance metrics that can be used for IVIS/ADAS evaluation, and there does not seem to be any consensus on which metrics are best suited to different evaluation purposes. The selection of metrics is seldom guided by a clear set of hypotheses. Rather, so far the goal has been to identify any effects of an ADAS/IVIS and, thus, a large set of metrics are generally used. Moreover, in many cases new performance metrics are invented for the specific study in question. This makes it difficult to assess the repeatability of results from different studies.

Moreover, the effects of scenario-related factors (curvature, speed, road width, test environment etc.) on the metrics have not been thoroughly investigated. Also, the effects of compensatory behaviours are often neglected; for example reducing speed in response to a secondary task may influence other measures, e.g. lane keeping.

An exception of the above is the LCT, which represents the main current effort towards a standardised, cost efficient, test regime for IVIS. However, the method is only applicable to IVIS evaluation of distraction/workload evaluation.

Based on the current review, the following general suggestions for further research in driving performance metrics within AIDE T2.2.5 could be given:

1. Basic empirical research to identify and explain the effects of ADAS/IVIS and scenario-related parameters on driving performance.

2. Identification of a set of valid and cost efficient metrics and scenarios that can be used for different types of ADAS/IVIS evaluation, and guidelines for selecting metrics to specific assessment purposes (hypotheses). In the case of IVIS, the LCT is a natural starting point.

3. Analysis on how performance metrics are best interpreted in terms of accident risk (as input to WP2.3)

3   Visual performance

Visual load due to in-vehicle tasks is of special interest especially since most information required by the driver, in order to manage the driving task, is visual (Wierwille et al, 1996). There are, two forms of visual resources: foveal and peripheral vision where foveal vision provides the high resolution vision while the peripheral vision provide motion cues (ibid.).

When secondary tasks require visual resources the amount of visual resource allocated to the driving task may decrease (Rumar, 1988). In order to cope with this situation the driver must use some sort of multitasking or time-sharing. Time sharing behaviour was studied by Rockwell (1988) where he looked at how his subjects shared the time between the road ahead and a smaller set of in-vehicle tasks. Rockwell (ibid.) found that individual glance times usually clustered at approximately 1.25 seconds and that operating the radio normally demanded 4-5 number of glances. Thus, the time sharing behaviour for carrying out the tasks was consistent while driving. The time of individual glances was fairly consistent and the number of glances increased for more complex tasks (ibid.).

Bhise et al (1986) conducted similar experiments as the ones by Rockwell. However, a wider range of tasks were used. The results were similar to the ones by Rockwell. E.g. the number of glances varied greatly between different tasks. However, Bhise and colleagues found some indications that the single glance durations did vary somewhat with different tasks (ibid.).

Thus, the driver copes with secondary tasks which compete for visual capacity also required by the primary task of driving by time sharing. However, there might be a problem, with regard to time sharing, when a driver chooses to monitor too many secondary stimuli. Rockwell (1988) describes this as a problem which occurs when e.g. complex displays require glance durations beyond the 90th percentile. In those cases the strategy most of the time is to make a set of glances of 1.25 seconds duration of each until the task is completed. However, if the display is too difficult to comprehend within 1.25 sec. of glance duration the driver might be tempted to increase the average duration of glances and thus the safety might be compromised (ibid.).

After Rockwell, numerous experiments have thus been conducted. However, according to Serafin (1993), there exist great inconsistencies among different studies which make the results difficult to interpret and compare. Serafin (ibid.) presents a detailed overview of experiments made where eye movements has been measured as a function of age, with regard to straight and curved roads and as a function of driver behaviour and workload.

Even though there still are a lot of inconsistencies between the definitions within the field of visual performance measurement a general distinction can be made between:
- Glance-based measures
- Non-glance based measures

Both groups are the outcome of ocular segmentation where *fixations*, *saccades* and *eye-closure* are identified (described in Larsson, 2002). 'Fixations' includes smooth pursuits and is defined as (ibid.):
> *"the alignment of the eyes so that the image of the fixated target falls on the fovea for a given time period"*

'Saccades' are defined as (ibid):
> "*the rapid movements changing the point of regard*.

Short eye-closures are the same as 'blinks' while longer eye-closures could be an indication of drowsiness (ibid) even though long eye-closures are still classified as blinks.

Thus, from *raw data* of eye movements *fixations and saccades* are the main building blocks. Then *dwell time* (i.e. the sum of all the individual fixations within a certain target) and *transitions* (i.e. the

move from one target to another target) are calculated in order to in a next step e.g. cluster these into *glances*.

The traditional way of measuring the entities and groups of measures above has been through manual transcription of video data (e.g. as performed and described in Victor et al., 2001). However, today there exist a range of more or less automatic head- and eye tracker sensors where at least two will be used within the AIDE project (i.e. FaceLAB (www.seeingmachines.com), Pertech system). There is also a need for usable tools for the actual analysis which will be addressed within AIDE (Wp2.2. Task 2.2.2).

The Occlusion method could be said to be more of a surrogate measure *reflecting* visual behaviour while performing an in-vehicle task and thus this measure will be covered in a separate section in this review.

3.1 Glance-based metrics

Glance based metrics are, within traffic research, often used in the assessment of glance behaviour towards certain pre-defined areas of interest (in vehicle or outside). Metrics which indicated visual distraction are being investigated and many studies in the literature explore the glance behaviour in relation to specific in-vehicle tasks.

According to the ISO 15007-1 a glance is defined as *a series of fixations at a target area until the eye is directed at a new area*.

The most commonly used glance-based metrics as defined in the ISO 15007-1 document is presented below.

3.1.1 Metrics

The most commonly cited glance based metrics as well as recommended in the ISO 15007-1 and ISO 15007-2 are:

1) 'Glance frequency'
2) 'Total Glance Time'
3) 'Single Glance Duration'

4) 'Time off road scene ahead'
5) 'Transition times between areas of interest'
6) 'Percentage of time spent on different areas'
7) 'Fixation probabilities' and
8) 'Link value probabilities'

Below the definitions of the metrics are given.

3.1.1.1 Glance frequency

The number of glances to a target within a pre-defined time period, or during a pre-defined task, where each glance is separated by at least one glance to a different target.

3.1.1.2 Glance duration and Total glance time

The duration of a glance refers to the time from the moment at which the direction of gaze moves towards a target to the moment it moves away from it. Total glance time to a target is the time which is

associated with a target (e.g. an IVIS) and provides a measure of the visual demand posed by that location.

3.1.1.3 Glance location probability

The Glance location probability refers to the probability that the eyes are fixated at a given target (location) during a sample interval.

3.1.1.4 Time off road scene ahead

Time off road scene ahead is defined as the sum of glance durations, over a sample period, for glances to all targets other than the road ahead (e.g. left and right side-view mirror, displays and instrument panel (e.g. radio, speedometer).

3.1.1.5 Transition times between areas of interest

A transition time is defined as a more or less linear function of the distance from one target to another. The duration between the end of the last fixation on a target and the start of the first fixation on another target is often defined as the transition time.

3.1.1.6 Percentage of time spent on different areas

'Percentage of time spent on different areas' is not presented in the ISO documents listed previously. However, this measure has been proved useful in a range of experiments (e.g. in Harbluk et al, 2002; Victor et al. in press). In the experiment by Harbluk the forward view was divided into a matrix with a set of cells, in order to calculate what percentage of time each subject spent looking in each cell.

3.1.1.7 Fixation probabilities

The fixation probability to a target reflects the relative attentional demand associated with that target. The example ISO 15007-2 states:

> *"if device use were to induce a relative decrease in the fixation probabilities associated with the driving scene, such as road curvature or rear-view mirrors, this would be considered indicative of the visual demand associated with the device".*

3.1.1.8 Link value probabilities

The link value probability is defined as the probability of a glance transition between two different locations. A transition is a change in eye fixation location from one defined target location to a different location.

3.2 General description of Non-glance based measures

Non-glance based measures are measures which do not for example calculate glance distribution between two areas of interest but rather are derived from fixations and saccades. Non-glance measures could for example be used when assessing the effect from cognitive and auditory challenging tasks (e.g. in Recarte et al., 2000, 2003, in the HASTE and CAMP project and in Victor et al, in press).

3.2.1 Metrics

Some examples of non-glance metrics are based on the fixations and saccades and are presented below.

3.2.1.1 Fixation based metrics

The duration, frequency/rate of fixations as well as gaze variation can be measured.

3.2.1.2 Saccade-based metrics

The duration, frequency/rate and amplitude/size of saccades can be measures. Also, the peak velocity and peak acceleration of saccades are metrics used.

Other non-glance based measures are pupil, eyelid, and head-movement characteristics however; they are rarely used in order to assess in-vehicle systems.

3.3 *Application of visual performance metrics*

The work by Serafin (1994) thoroughly examined eye pattern on straight versus curved roads. Eye patterns and transition probabilities between road features, car mirrors and in-vehicle objects were measured. However, there was no focus on assessing the impact of in-vehicle information systems on the visual behaviour but rather 'just plain driving'. The main conclusions were that some differences in eye fixations are due to road curvature. Also, on all road segments, drivers tended to fixate as far down the road as they could. Eye fixations durations were fairly consistent between different curvatures.

Rockwell conducted research on eye movement both with regard to perceptual search and scan pattern development in novice drivers (1972) and general time sharing between road ahead and in-vehicle tasks (1988). Rockwell also presented ideas and issues to solve within the more technical field of measuring eye movements in traffic research and in driving conditions (Rockwell, 1972).

Chapman et al. (1998) has studied eye movements in relation to differences in scene complexity as well as in relation to driving experience. The authors found that scanning increases with the complexity of the driving scene and that the frequency of eye movement at the same time increases along with a decrement in fixation length.

Dingus et al. (1989) conducted an on-road experiment in order to examine visual glance times etc. Tasks such as conventional in-vehicle tasks and navigation tasks were compared with regard to glance times and number of glances. Total visual demand (sum of individual glance lengths into the car) varied with the tasks and the *total* glance times varied between 0.78 second for checking the speedometer and 10.63 seconds for one of the navigation tasks. Single glance lengths varied somewhat between different long tasks even though the overall results were in accordance with earlier work.

In a review on measures and methods used to assess the safety and usability of Driver Information Systems, Green (1993), present summaries on experimental findings on eye movement data. Work by Zwahlen is reviewed by Green (ibid.). One of Zwahlen's ideas was to look at the time of occluded vision in relation to lane keeping performance. Thus, Zwahlen measured the time it took for subjects, who was not allowed to look at the road while performing an in-vehicle task, and before he/she started to have degradation in lane keeping performance. Based on the results from this method Zwahlen could hypothesize that no in-vehicle task should exceed a certain length in time, a certain amount of glances or require glance times above a certain number of seconds. Thus, even though the main focus was not on the actual measures of glance behaviour during the drive the work has served as a basis for legislation for the use of IVIS while driving. Green also presents work conducted by Wierwille and students in his research group. The main work reviewed was conducted in on-road experiments and with an ETAK navigation system. Subjects performed a range of tasks (e.g. checking speedometer, adjusting fan, reading name on cross street). Numbers of glances (per task), mean glance length along with the number of lane exceedences were measured. One experiment conducted by Verwey in 1991 was examined in the review by Green. The objective of the experiment was to assess factors which contributed to driving workload. A visual detection task, a visual addition task and an auditory addition task were used in the experiment. Also, the type of road differed (motorway, rural) as well as traffic density. A dashboard mounted camera recorded eye movements. Glance frequencies and duration in relation to four locations (right, left, interior mirror, display) were analyzed. The different driving conditions were reflected in the eye movements. Drivers looked more to the left for the left

turns and more to the right for right turns. Also, there were differences in the number of glances to the mirror and display as a function of the driving situation. The number of mirror glances was sensitive to workload, that is, subjects looked more in the mirror for easier driving situation compared to the more challenging. Also, the easier the driving task, the longer the average fixation. Glance frequency appeared to be more sensitive to workload than glance duration.

Recarte and Nunes have examined the effect of mental workload on visual search and decision making in several experiments (2000, 2003). More precisely the aim of the research was to examine if mental activity affect visual-detection and response-selection capacities. The traditional research (previously reviewed) has mostly focused on the phenomenon "did not look and therefore did not see" (i.e. exogenous distraction) where visual tasks have been explored features. However, what Recarte and Nunes (ibid.) wanted to explain was the further problem "looked but did not see" (i.e. endogenous distraction). In one field experiment (Recarte et al, 2000) the main aim was to examine the consequences of performing two verbal and two spatial-imagery tasks on visual search. The main effect showed that the pupil size increased while performing the each of the four tasks. Also, the glance frequency at mirrors and speedometer decreased during the spatial-imagery tasks. In a second field experiment (Recarte et al. 2003) subjects performed a range of tasks which were verbal acquisition tasks, production tasks and complex communication tasks either by phone or with a passenger. Also, a simultaneous visual-detection and discrimination test was used as performance criteria. The main findings were that the mental tasks produced spatial gaze concentration and visual-detection impairment. Consistently with the result in the previous experiment the frequency of glances towards mirrors and speedometer was reduced (ibid.).

The impact of cognitive tasks on driver visual behaviour was also examined by Harbluk and Noy (2002). The tasks performed by the subjects were all mobile phone tasks but with different level of cognitive complexity. All tasks were performed with a hands free phone. Saccadic movements (i.e. high-speed ballistic eye movements) were analyzed as well as the percentage of time the subjects spent looking at the road ahead (road ahead was defined as the central 15° of the windshield), the left and right peripheral areas and on in-vehicle objects (instruments, mirrors etc). In addition to this analysis, density plots were created in order to further visualize how subjects looked in the forward view.

The results showed that saccadic eye movement was sensitive to cognitive workload. The number of saccades decreased along with more and more complex tasks. It was also shown that the percentage of time spent looking at the central region increased as the task demands increased (see Table 3.1).

Table 3.1 Percentage of time spent looking at central and peripheral regions.

	Left Periphery	Central 15°	Right Periphery
No Task	0.73	78.63	2.09
Easy Task	0.65	80.84	2.19
Difficult Task	0.55	82.68	1.56

The forward view was divided into a matrix of cells (11 in the horizontal and 5 in the vertical plane) in order to show more in detail how the gaze behaviour of drivers changed. Then, summaries were made on what the percentage of time each driver looked in each cell of the matrix during No Task and Difficult Task conditions. The results showed that even though drivers looked more into the central during the Difficult Task condition they seemed to have different strategies. Some drivers looked more into the middle and up when performing the difficult task while some others looked more down while performing the task (see Figure 3.1).
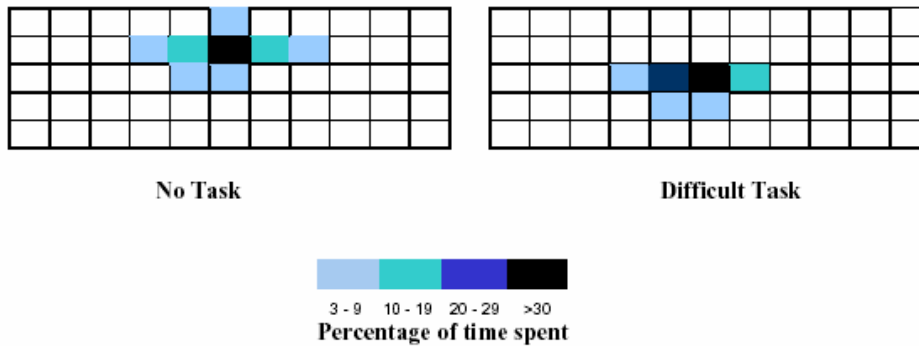
Figure 3.1 Example of a participant from the look down group.

Victor and Johansson (in press) performed experiments where subjects performed visually, cognitively, or manually demanding in-vehicle tasks. Glance analysis (according to the ISO 15007-1 document) was conducted for the visual and/or manual tasks. The analysis showed, among other things, a strong correlation between the glance measures. Analysis of gaze concentration was made for 'normal driving' (i.e. no task – baseline) and the tasks with an auditory, verbal and /or cognitive component. For the auditory, verbal and/or cognitive tasks, fixation density maps were calculated for each task. This experiment had more difficult calculation tasks compared to Harbluk and Noy (see above) and thus had a stronger effect of gaze concentration. One of the tasks in the experiment by Victor and Johansson (in press) was to have conversations in a hand held and a hands free phone. A third task included conversation with the passenger. The results showed that speaking with the passenger caused a similar amount of gaze concentration as hands free conversation.

The HASTE project has conducted research on visual performance. The following section is based on Deliverable 2 in HASTE. For further details see Östlund (2004).

Two different eye tracker systems were used. Transport Canada (TC) carried out this research using the University of Calgary Driving Simulator (UCDS) and a head mounted eye tracker by ASL. Volvo Technology (VTEC) carried out a second experiment in the VTEC simulator together with the Face LAB system developed by Seeing Machines. The Face LAB system has a pair of cameras placed on the instrument panel.

The main independent factors investigated in the project were task complexity for two surrogate tasks (one of which was mainly auditory and the second more visual task) and road complexity (e.g. straight and curved sections). The principal variable analyzed for the auditory tasks was variation in gaze angle while glance-based measures were analyzed for the visual task.

In the VTEC experiment the following measures were calculated

1. for the *visual task*:
- Glance frequency;
- Glance duration;
- Glance duration variation;
- IVIS glance duration proportion (i.e. the proportion of IVIS glance time of the total task time);
- Percentage road center (prc; i.e. the proportion of glance time spent towards the road center);
- The number of glances longer than 2 seconds.

2. for the *auditory task*:
- Gaze angle variation (st_ga; i.e. the standard deviation of gaze angle);
- Percentage of road center (prc).

For the visual task in the VTEC experiment the comparison of mean single glance durations show that mean glance duration increases almost linearly as a function of difficulty level of the visual task. An effect of road level (curves and straight segments on a rural road) was also shown where the curves led to significantly shorter glances. The interpretation for this presented in the HASTE deliverable is that higher visual demand imposed in the curves compared to the straight sections, leaves less visual resources for the visual secondary task.

Glance frequency was shown to be less sensitive to task complexity in the curved sections. However, the measure still shows a rather linear effect of task complexity in the straight sections where the number of glances increase along with task difficulty.

The percentage of time spent towards the road center decreased as a function with difficulty level. For the PRC measure significant differences were found between all complexity levels of the visual task in the rural road.

Results from the VTEC experiment show a strong gaze angle concentration towards the road center (see Figure 3.2), while the subjects performed the *auditory/cognitive* task.
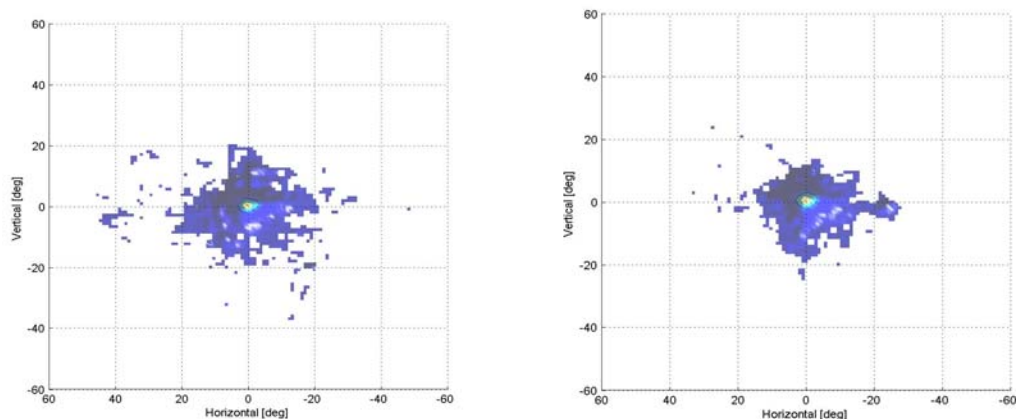
Figure 3.2 Spatial density plots for gaze data during baseline driving (left) compared to auditory task on the rural road (the three levels of task complexity are aggregated).

These results are thus consistent with the experiments reviewed above (Recarte et al. (2000, 2003) and Harbluk et al., 2002).

The effect in both prc and gaze angle variation indicate that increased task performance results in reduced gaze variance (i.e. higher gaze concentration towards the road centre).

The following measures were calculated in the TC experiment:

1. for the *visual task*:
   - Glance frequency;
   - Glance duration;
   - Total glance times to IVIS.

2. for the *auditory task*:
   - Percentage of road center

The glance frequency to the display with the visual task was significantly higher for the more difficult level compared to the easier task. It is argued in the deliverable that drivers dealt with the increasing

complexity of the visual task by looking at the display more often. However, the drivers did not alter their mean glance duration as no significant differences observed for mean glance duration as a function of task difficulty or road complexity were found. Results from 'Total Glance Duration' indicate that as task difficulty increased, drivers spent greater amounts of the total task time looking at the display. Thus, drivers spend greater amounts of time looking away from the road when they are trying to do more demanding visual in-vehicle tasks.

For the visual task the percentage of time that drivers spent looking at the area of "road center" decreased as the level of visual task difficulty increased.

Percentage of time spent to the road center was measured for the auditory task. However, this measure did not yield significant results as a function of task difficulty.

The ADAM project has conducted research on visual performance measures throughout a range of experiments. For most of the experiments conducted in the project the following set of secondary tasks has been used:
- Radio tuning: Set radio to a certain frequency.
- Sound adjustment: Set treble to maximum.
- Change Cassette: Change audio cassette and put in case.
- Navigation speller: Enter street name with rotary push button.
- Navigation map: Set target cross of navigation system to a certain point on map.
- Cell phone: Enter 4 digit PIN in cellular phone.
- Sweets: Unwrap a sweet and put it into the ashtray.
- Talk on telephone: Answer some simple questions in hands free mode.
- Kleenex: Unfold a Kleenex and put it onto passenger seat.
- Address Book: Lookup phone number in small paper address book.
- Map book: Open map book on page X and decide which of two towns Y and Z in further North.
- Coins: Get a 20 and a 10 cent coin out of a purse, close the purse and place back onto the passenger seat.

The tasks were considered to be 'mainly visual demanding', 'mainly manual demanding', 'visual and manual demanding' or 'neither visual nor manual demanding' (Bengler et al., 2004).

A first experiment (ibid.) which included eye movement data was conducted in within the ADAM project in the BMW static driving simulator. Except for driving performance and other objective measures visual performance was collected. Dependent variables related to visual performance were: Total task time, Number of glances, Glance duration and Perceived distraction. The visual data were collected by one camera, which recorded the face of the driver.

8 tasks from the list above were analyzed with respect to eye movement. In the ADAM report (ibid.) the results on Total Glance Time (i.e. the summed duration of all glances needed to complete the task) is presented and shows high variances and a significant effect of secondary task.

The results indicates that the Navigation task with map created the most number of glances followed by the Navigation – Spelling task and Changing frequency on the radio.

The number of glances was also correlated with the subjects own subjective ranking of visual distraction. The Spearman correlation test show a correlation of 0.74 which can be compared to car following, mean distance (0.62), lane deviation (0.58), and task duration (0.34).

The second experiment was conducted in the Daimler Chrysler Driving Simulator which has a moving base. Recorded data of visual performance was captured by the use of Seeing Machines Face LAB system. Also, subjective data of subjective opinions from the subjects on 'distraction potential' was

collected. Total Glance Time was calculated and the results again show large variance within tasks and differences between different tasks. A comparison was also made between the two experiments and for the glance data there was a high correlation, which were not the case for all vehicle measures.

According to SAVE-IT, prior research clearly indicates that visual distraction degrades driving performance and increases the likelihood of crashes (Witt et al, 2004). SAVE-IT also concludes that prior research rarely used automatic eye tracking systems to measure visual behaviours in real time and with a focus on task-based (e.g. radio tuning) rather than time-based visual behaviour. Real-time measurement of time-based visual behaviours is critical to SAVE-IT rather than task-based. That is, there is no need for the system to identity the exact task performed when distraction is detected.

The main aim of the part of SAVE-IT which concern visual performance is to:

1. Identify eye glance measures that are diagnostic of visual distraction and that can be used in real-time, adaptive interface technology systems.
2. Determine performance (including RT) effects of visual distraction. For example, RT = f (glance duration, glance frequency, etc.).

It is important to note that the first aim here is not the same as in Work package 2.2 of AIDE where the focus lays on off-line measurement of workload. That is, no real time identification of eye glances is of crucial importance. Within AIDE this will be done in WP 3.3.

Current experiments in SAVE-IT phase 1 measured visual behaviour in real time using Seeing Machines eye tracking system and the results indicates that visual distraction degraded performance (e.g., RT).

The project has so far in the field of 'visual distraction' produced one literature review by Delphi and conducted two experiments where visual data was collected. The second experiment along with the data analysis and algorithm development was planned to take place in September-December 2003.

The following description of the two experiments is based on a presentation at the Phase I SAVE-IT Briefing (Zhang, 2003).

In the first experiment the following independent variables were included:

1. Road Type: Rural Roads vs. Freeways;
2. Visual Distraction: Distraction Levels and Display Eccentricity.

A within-subjects design was used with 2 road types along with 7 *visual* distraction conditions balanced across 14 subjects.
The dependent variables were:
- Variables that are common across SAVE-IT experiments: brake reaction time, foot-off-accelerator reaction time, steering reaction time, steering entropy
- Additional variables: SDLP, mean lane position, lane departures: number and duration, crashes: number and severity, steering SD, mean steering, speed SD, mean speed, secondary task: accuracy and completion time, head and eye movement measures such as, *Eye gaze (XY) coordinates, Peak glance duration, Mean glance duration, Glance frequency, Total glance duration, and Time-based combinations of eye glance measures, Head orientations (XYZ coordinates), Blinks, eye closures*. (No definition of peak glance duration was given in the document).

The objective of the second experiment was to determine the performance impact of momentary off-road glances. For example look at the off-road glance duration at the moment of lead vehicle braking. The main change to the first experiment is that the lead vehicle braking is tied to the off-road glances.

The main findings so far in this specific part of the SAVE-IT project are:
- Crash predicted from measures of visual distraction.
- Driving performance predicted from measures of visual distraction.
- Reaction time predicted from measures of visual distraction.
- Performance effects of gaze eccentricity.
- Visual glance behaviour during performance of conventional and navigation tasks.
- Effects of task complexity and display characteristics on eye glance behaviours.
- Effects of driving task demand on eye glance behaviours.
- Effects of driver age, experience and route familiarity on eye glance behaviours.

It is important to acknowledge that the overall objectives of SAVE-IT is not to come up with a test battery for evaluation of IVIS and ADAS but rather create a workload and distraction manager that assesses the driver's attention allocation based on the relative demands of the outside and inside vehicle tasks (Delphi, 2004). Thus, the measures still reflect the demand the in-vehicle tasks impose on the driver but the results will be used in real time in the prototypes build in the second phase of the SAVE-IT project. An example of a future condition would be that the adaptive frontal collision warning sensitivity and ACC headway adjustment would be based on the level of distraction detected. Also, if the driver was visually distracted the warning could instead be presented in an auditory mode.

The CAMP project has recorded data on visual performance both with an eye tracker as well as with video analysis of the face of the driver. Tasks evaluated have been natural in-vehicle tasks of different modalities. Brief description of plans and preparation of experiments can be found in the two annual official reports from CAMP (Deering, 2002; Shulman et al. 2004). However, by the time of writing this review, no official results from the experiments were published.

The COMUNICAR project performed laboratory test, driving simulator tests as well as on-road tests throughout the development of the 'Comunicar prototypes'. The only measure in the evaluation concerning visual performance was the assessment of viewing behaviour in the laboratory phase where the results indicated that operating particularly the radio while driving is performed faster with the visual comunicar prototype, leading to shorter distraction from the road scene (Hoedemaeker, 2003).

The HUMANIST network of excellence intends, among other things to, according to the official web site, to develop integrated tools and methods in order to investigate safety issues of communicating technologies in the context of driver/environment/systems. However, no official information on more specific tools is yet to be found (e.g. tools which relate to visual performance).

The RoadSense project conducted a literature review on measures for the assessment of in-vehicle systems (Nathan, 2004). One of the goals for this review was to recommend measures to incorporate in the D-BITE, further to be used in real on-road tests.

Visual performance measures which have been recommended to be used in the real on-road tests within the project have been separated into recommended measures of 'visual management and 'system usability/ suitability' (the definitions in brackets are from the RoadSense document).

Visual management:
- 'Time on road perception information inside the vehicle' (defined as the time spent looking at mirrors and at a system giving visual road information).
- 'Time on driver information inside the vehicle' (defined as the time spent looking at the dashboard.
- 'Time on any other information inside the vehicle' (defined as the time spent on all information except for driver information and road perception information).
- 'Visual demand ISO' (defined as the percentage of the total time that the driver spends looking at an object or area such as the road or inside the vehicle).
- 'Decrease in glance frequency to the mirror'.

System usability/ suitability:
- 'Dwell time' (defined as the sum of consecutive individual fixation and saccade times to a target in a single glance. Definition partly refers to the ISO 15007 document).
- 'Number (or percentage) of eye fixations in an area' (defined as the number of time the point of gaze is located in a defined area).
- 'Glance duration' (defined as the time from the moment at which the direction of faze moves towards a target to the moment it moves away from it which includes the transition time to that target. Definition partly refers to the ISO 15007 document).
- 'Glance frequency' (defined as the number of glances to a target within a pre-defined time period or pre-defined task where each glance is separated by at least one glance to a different target. Definition partly refers to the ISO 15007 document).
- 'Length of eye fixations' (defined as how long a driver visually fixates an object/scene).

By the time of writing this review, no official information on the results from the experiments was available.

3.4 *Demand for further research*

From the research conducted above it can be concluded that visual performance indicators are important in the assessment of IVIS especially since the effect from cognitive load often is more evident in this particular data than in many other measures (e.g. vehicle performance data). The latter is evident in e.g. the research in Recarte (2000, 2003), Harbluk (2002), Victor et al (in press) and in results from the HASTE project.

Even though it seems like many researchers have focused on the ISO metrics described in the beginning of this chapter (mainly glance frequency, total glance time and single glance duration) there exist a need to clearly define new introduced measures as well as describe how and if previous ISO metrics have been altered. How well a definition is followed can for example depend on the eye tracker used for the collection of data.

In many papers which describe research on visual behaviour the technical aspects of measurement are mentioned. Even though the equipment has been refined since the earlier experiments e.g. conducted by Rockwell in 1972, there exists a great need today of improvement of the eye tracking sensors as well as of the analysis process of the data. Even though manual video transcription may not be as used as before, the semi- automatic analysis of data obtained from advanced eye tracking systems can still be very time consuming.

The equipment need to be easier to use by an inexperienced person. The set up needs to be easier with a faster calibration process of the sensor and with a quick and easy way of building e.g. face models. There is also a great need of improvement in the classification of fixation and glances in order to reduce the amount of data discarded due to tracking errors. The process needs to be more automatic with no need of manual double check to ensure that data validity and reliability.

Most eye trackers currently on the markers share the problem with field of view. There is a trade off effect for the remote sensors where one has to zoom out the cameras in order to fit subjects into the view, causing reduction of accuracy. For most of the head mounted sensors there is instead a problem that the field of view of the camera mounted on the subject's head is too limited compared to the field of view of the subject. Thus, the subject can e.g. glance outside what the camera sees and therefore one have to guess to what object the glance is directed. Also, there is a trade off between visual angle and resolution.

It is recommended that, based on this review, when assessing IVIS to primarily use the ISO measures described in the beginning of the chapter in order to collect data for later comparisons. It is also

important to distinguish between visual and auditory/cognitive load, i.e. if the IVIS task is auditory, gaze concentration should be looked at as well as e.g. percent road center. The main focus of eye movement has lately been on the assessment of IVIS. However, ADAS might need a different approach in the analysis where visual performance might be regarded in more "semantic" terms such as whether the driver looks at signs, oncoming vehicles, lead vehicles, pedestrians etc.

3.5 Technical overview

Traditionally, manual transcription of eye movement data has been made based on video recording of the subject's head and eye movements. However, today there exist a great number of different devices in order to sample data more automatically. A wide range of methods can be used in order to track the eye:

1) Pupil Center, Corneal Reflection of Light Emitting Diodes.
2) Image processing using template matching and feature tracking.
3) EOG (measuring the movement of facial muscles).

Some systems are head mounted systems (eye ware versions, VR helmets) while others are so called remote systems.

Both Seeing Machines' FaceLAB 2.0™ and Smart Eye's Smart Eye Pro 1.2™ are two non-intrusive and passive. Each system uses one set of cameras. Information below has been gathered via www.smarteye.se, www.seeingmachines.com and through personal communication.

Smart Eye Pro 1.2 runs in real-time at a rate of 30 Hz (in recent versions up to 130 Hz). IR illumination is used (IR flashes) to illuminate the face of the driver independent on surrounding light conditions. According to the Smart Eye Pro 1.2 manual the two cameras can be positioned totally independent from each other. However in order to e.g. get good tracking quality it is advantageous to place the cameras in order to get the driver in the center of the camera image. The cameras also need to be placed in certain positions to minimize sun glare straight in the lenses of the cameras. The cameras also need to be adjusted if driver is using glasses (to avoid glare in the lenses of the glasses).

FaceLAB 2.0 runs in real-time at a rate of 60 Hz. The system makes use of surrounding day light to illuminate the face of the driver. At dusk and night condition IR illumination is used. Two different driver models need to be created for NIGHT and DAY mode. The system often looses tracking when in DAY mode and when the driver is passing through e.g. a tunnel. An optimal camera configuration for the FaceLAB system is when the cameras are placed symmetrically in front of the head of the user with plates under the cameras lying in a common place (according to FaceLAB manual). The FaceLAB system is, due to the small tracking volume, largely dependent on the height of the driver. It is often necessary to adjust e.g. the tilt angle of the cameras for different drivers to place the face of the driver optimally in the camera view. FaceLAB has been validated in one experiment (Victor et al. 2001).

The Applied Science Laboratories (ASL) works with the technology and systems for eye tracking and product information can be found at the ASL web page. A range of different eye tracker products are available. The products are based on both IROG (limbus tracker) and VOG (video) techniques. The ASL model ETS-PC II is a non-contact measurement instrument specifically designed for eye tracking in moving vehicles. The EST 501 system has head-mounted optics and is designed to measure the eye line of gaze with respect to the head. The 504 system has remote optics and the system combines a magnetic head tracker with the pan tilt. Further information can be found on the ASL web site (www.a-s-l.com).

LC technologies has developed an unobtrusive "eyelid closure and visual point of regards measurement system" which has been used in e.g. simulator experiments. The systems are un-obtrusive and remote and can be used to track a user's gaze point.

The Pertech system used by Renault in AIDE is developed by Pertech which is a company initiated in March 2004 (based on collaboration between e.g. Renault, Tekano, Universities). The main idea behind the Pertech sensor is image processing. There are two versions of head mounted devices available (one where the sensor is attached to eye glasses and one where the sensor in head mounted with a "strap").

4   The occlusion technique

The measurement of driver visual distraction induced by in-vehicle tasks has become a major issue in Human Factors research (c.f. Gelau, 2004). One possibility to study effects of drivers' interaction with new on-board systems while driving could be to perform experiments in real traffic or in a driving simulator. Because this approach is very demanding and expensive, there is a need for a method that is easy to use and applicable in the very early stages of system development (e.g. Krems et al., 2004). The *occlusion technique* has recently come under consideration as an assessment tool that fulfils these requirements (e.g. Gelau & Krems, 2004). The goal of this chapter is to briefly describe the application of the occlusion technique and discuss its validity and reliability as a method for HMI assessment of IVIS.

The primary task of driving relies heavily on the visual channel which is also the case for many in-vehicle tasks (see previous chapter). Thus, methods and tools are needed to assess the visual demands of the primary driving task and/or those demands imposed upon the driver by tasks performed additionally while the vehicle is in motion. As a method for HMI assessment of IVIS the occlusion technique is applicable for both assessing the visual demands of driving and simulating the interruption caused by doing some other task while driving (see the section on time sharing in the previous chapter). The occlusion technique is defined as a *"measurement method involving periodic/intermittent physical obscuration of the participant's vision or the obscuration of visual information under investigation"* (ISO/ TC22/ SC13/ N763R). The occlusion technique is based on the systematic control of the permitted time intervals for a subject to view at an object or perform a task.

In general, applications of the occlusion technique are well established in research on driving behaviour. A frequently quoted study was published by Senders, Kristofferson, Levison, Dietrich and Ward (1967), who introduced the occlusion technique as a procedure to quantify the visual demands of the primary driving task. Major progress was made by the development of so-called PLATO spectacles (Milgram & van der Horst, 1986), which made it technically very easy to control occlusion and non-occlusion sequences and eliminated interfering effects of readapting the eye after the occlusion interval. More recently the occlusion method was implemented on PCs in order to be able to use it in a laboratory context without the need of additional devices (e.g. Keinath, Baumann, Gelau, Bengler & Krems, 2001). A more detailed technical description will be given in the next paragraph. A review of the development and usage of the method in other areas than HMI assessment can be found in van der Horst (2004).

4.1 Description of the method

Studies which explored the occlusion technique as a tool for the assessment of the in-vehicle HMI investigated the application in the two different areas: First, there is evidence that the occlusion techniques reliably discriminates between visual displays of different complexity (e.g. Gelau et al., 1999; Keinath et al., 2001). This means that it can provide valid results when purely visual tasks have to be assessed. With the second area which has already been mentioned in the introduction also visual-manual tasks are addressed. For these tasks, which are perhaps more representative for the majority of in-vehicle tasks, the occlusion technique is applied to assess them under the aspect of their *interruptability* (e.g. Keinath et al., 2004; Krems et al., 2004) or *chunkability* (e.g. Noy et al., 2004; Stevens et al., 2004). The remainder of this chapter will focus on the latter area.

As already mentioned there are basically two different methods for obtaining the systematic control of the time intervals permitted for vision. The majority of published research studies have used goggles or spectacles to achieve the occluded intervals. Another means of occlusion in a number of studies is a PC that has periodic screen blanking (see Stevens et al., 2004 for a review). There is evidence which suggests that PLATO goggles or screen blanking methods are equally preferred by subjects (Weir et al, 2003). However, the PLATO goggles provide a more realistic environment for assessing the visual workload of IVIS as the driver is unable to view the vehicle display or IVIS controls whilst focusing

on the road ahead. When using the screen blanking method, subjects are still able to view the touch buttons, whereas during real driving both the screen and the manual controls are not visible when the driver is viewing the road scene (Stevens et al., 2004).

The logic behind the application of the occlusion technique to assess in-vehicle tasks with respect to their interruptability is simple and straightforward. The aforementioned "obscuration of vision" when performing an HMI task is applied, in order to simulate the interruption caused by the driver's view back to the road scene during conditions of real driving. Interruptability means the degree to which subject performance on the HMI task suffers from these obscuration of vision. Basically this is determined by comparing subject performance on the task under unoccluded conditions with the performance under conditions of occlusion. In practice this means that the following parameters are calculated from the data of an occlusion experiment:

TTT (total task time): The time it takes to perform the HMI task under investigation, under unoccluded conditions.

OCCLT (occlusion time): The time for which the scene is occluded during the trial where the task has to be performed. This can be a constant value, such as 3 sec; or it can follow a certain distribution, like a normal-distribution with AM= 2 sec and SD = .3 sec.

INSPT (inspection time): The viewing time where the shutter glasses are open during the trial where the task has to be performed. Values between 1.5 and 2 sec are often used. Also INSPT can be generated by means of a distribution.

TSOT (total shutter open time): The total time for which the scene is visible. If goggles are used this is the sum of all sequences where the goggles are open.

TSCT (total shutter closed time): The total time for which the scene is not visible. If goggles are used this is the sum of all sequences where the goggles are closed.

OCCLT and INSPT are the essential parameters of the occlusion method. However, for the aforementioned comparison of task performance under occluded and unoccluded conditions the index *R* is calculated. This index is defined as follows (c.f. ISO/ TC22/ SC13/ N763R):
R_{jk} = Mean $TSOT_{jk}$ / Mean TTT_{jk} for the jth subject an kth task

From this ratio it can be derived that task performance suffers from the interruption caused by the obscuration of vision when R is greater than 1, i.e. time of vision required under conditions of occlusion is greater than under unoccluded conditions. Up to now no "critical value" of R has been defined which could be applied for a decision if a certain HMI task is "sufficiently interruptable" in order to not to interfere with the primary task of driving. The definition of such a value needs to be justified by research. Thus, the next paragraph gives a brief review on recent studies on the validity and reliability of the occlusion method.

4.2 Research on the validity and reliability of the occlusion technique

A set of studies was performed in recent years to assess occlusion technique and to show experimental validation of the method. The studies had to prove whether occlusion is able to distinguish in-vehicle dialog concepts which are tolerable concerning visual demands and chunkability, from those which are not (Krems et al., 2004).

In a study (ibid.) reading tasks of different complexity were presented to the subjects. In the simple task the subjects had to find a route from city A to city B on a map, whereas the shortest route had to be identified in the difficult task. The presentation time of the map varied in eight steps between 0.2

sec. and 1.2 sec. Results show that the probability of error was higher for the complex version across all times of the presentation. Values started to converge at a vision interval of 1.2 sec. In a second part of the study, the stimuli were subject-paced. Similar results were found in this case compared to those of the first part.

To test occlusion with regard of interuptibility of dialogs subjects were required to find a given name in simplified phone directories displayed on the PC. Subjects were interrupted at specific steps of the dialog and were given time as long as necessary to identify the correct entry. Vision interval was manipulated (0.6, 0.9 and 1.2 sec.). Furthermore, task complexity was varied (predictable place; reordered list). The number of errors was measured. It was found that probability of errors decreased with rising vision interval and was lower for the simple task.

Another study on the effects of task interruption was carried out where subjects had to find the phone number in a displayed text of 5 lines. The 'total task time' to solve the problem and 'task errors' were measured for a dual-task situation (with secondary task) and a single task situation (without secondary task). The total task time was longest for the uninterrupted situation, but no difference was found for errors. It was concluded that time of occlusion was also used for task completion. Furthermore, criteria which are only based on uninterrupted total task time (e.g. 15-second rule as proposed by Green, 1999) seem to be not sufficient for the evaluation of real driving situations."

Summarizing the experiments it was concluded that:
- occlusion technique was able to discriminate between displays and dialogs of different complexity,
- occlusion technique was able to discriminate between different conditions of task resumption and to show which of the given conditions affected an additional secondary task,
- The occlusion technique can be rated as a method that is able to evaluate the HMI design of IVIS with respect to their suitability while driving.

Whereas the validity and sensitivity of the occlusion technique has been demonstrated in numerous experiments, there is only rare evidence on its reliability. This is a clear gap in research since reliability can be interpreted as a precondition of validity. Moreover the reliability of the occlusion technique has recently been questioned in the discussions in ISO/TC 22/SC 13/WG 8. To fill this gap a project has been started by BASt where data from four occlusion experiments performed within BASt projects are re-analysed under the aspect of reliability. Results will be available by the end of 2004.

4.3 Conclusions

The present chapter provided some basic information on the application of the occlusion technique and a review of studies demonstrating its validity. However, it should be stressed that the considerations taken in this chapter are confined to a certain class of situations. In general, the occlusion method can be used as a single task procedure as well as a procedure to simulate a multi-task situation. If a subject is interrupted without any additional task to be solved during the occlusion phase there is a single task situation. If there is an additional task that has to be worked on during occlusion and if the first task has to be suspended or abandoned after occluding the scene there is a task-switching situation (Gelau & Krems, 2004). To explore this extension of the "scope" of the occlusion technique would be a worthwhile goal for further research.

5   Physiological measures of workload and stress

A brief presentation of a selection of physiological measures sometimes used in order to measure workload follows below.

5.1   Electroencephalogram (EEG)

Electroencephalogram (EEG) is a technique for studying the electrical current within the brain. Electrodes are attached to the scalp and wires connect these electrodes to a machine which records the electrical impulses. According to a previous review of psychophysical measurement techniques in HASTE (Roskam et al, 2002) the EEG has a low frequency and high amplitude under sleeping conditions. With increasing mental activity, the frequency increases and the amplitude decrease. De Waard (1996) distinguishes between background EEG, where certain frequency analyses are performed on certain ranges or bands, and event related potentials. During event related potentials (ERPs) the EEG is related to a specific stimulus and during the events the P300 is important. The P300 is a specific peak with positive amplitude found about 300 msec after the presentation of a stimulus. The amplitude of the P300 is used as a measure for the amount of attention that was required in order to evaluate the stimulus and can therefore be used as a measure for mental workload.

According to de Waard (1996) background EEG is mainly used to measure low vigilance state while not many experiments exist where background EEG is used for high workload. However, event related potentials have been used in a range of workload experiments (de Waard, ibid.).

5.1.1   Examples of the application of EEG

De Waard (ibid.) refers to work by Sirevaag on background EEG who found a decrease in alpha waves (8 to 13 Hz) activity and an increase in theta (4 to less than 8 Hz) when the subjects performed two tasks instead of one.

5.2   Electrodermal activity (EDA) and Galvanic Skin Resistance (GSR)

Electrodermal Activity (EDA), a technique very similar to Galvanic Skin Resistance (GSR), measures the level of autonomic system activity by measuring the electrical resistance of the tissue path between two electrodes applied to the skin. This technique has been extensively used in animal and human research on pain, anxiety and stress levels.

According to de Waard's (1996) overview of physiological measurement techniques, the highest density of eccrine glands (sweat) is seen on the palms and soles and therefore the EDA is usually measured on either of these.

5.2.1   Metrics

Below, the definitions of some of the most common metrics are given.

5.2.1.1 Skin conductance level

Within HASTE, skin conductance level was calculated as the root mean square (rms) of the 0 - 2.0 Hz component of the skin conductance signal.

5.2.1.2 Skin conductance variation

Skin conductance variation was calculated as the root mean square (rms) of the 0.5 - 2.0 Hz component of the skin conductance signal.

5.2.2 Examples for the application of EDA or GSR

De Waard (1996) refers to a second review by Kramer (1991) who presents a wide range of experiments where EDA and GSR have been used in order to measure level of workload and stress.

In the HASTE (Östlund et al., 2004) project skin conductance level and skin conductance variation was included as part of the optional measures. Partners performed one simulator experiment where those measures were taken as well as one field experiment. In the simulator experiment an effect on both skin conductance variation and level was found for the auditory/cognitive task. The effects were very similar. For the visual task mean skin conductance and variation increased with task difficulty indicating an increased level of stress and workload. In the field experiment skin conductance level and variation were measured. However, for the auditory/cognitive task no effects were found. Effects of the visual task were found in both conductance level and variation. The significant effects were found between the baseline and task complexity levels but not in between the different difficulty levels.

5.3 Cardiac Activity

An Electrocardiogram (ECG or EKG) is a graphic diagram produced by an electrocardiograph, which records the electrical voltage in the heart in the form of a continuous strip graph. The ECG results can provide information such as whether the heart is performing normally or suffering from abnormalities (e.g. extra or skipped heartbeats - Cardiac arrhythmia), indicate coronary artery blockages (during or after a heart attack), detecting calcium, magnesium and other electrolyte disturbances and physical shape of a patient during stress tests. When measuring workload and stress in relation to the use of in-vehicle systems ECG is can used either by looking at changes in the inter-beat-intervals or how the heart rate varies.

5.3.1 Metrics

5.3.1.1 Inter-Beat-Intervals

The Inter-Beat-Intervals (IBI) measure is defined as the mean time interval between the heart beats, identified in electrocardiogram data. Heart Rate Variability is the variation of Inter-Beat-Intervals and is calculated as the mean value of the 0.07-0.14 Hz component of the IBI spectral density.

According to Deliverable 2 in HASTE (Östlund et al., 2004) the Inter-Beat-Intervals can be considered a global measure of general arousal. Since arousal and stress may be the results of mental workload, IBI may be used as an indicator of mental workload.

5.3.1.2 Heart rate variability

Heart rate variability (HRV) refers s to the beat-to-beat alterations in heart rate. According to Deliverable 1 in HASTE (Roskam, 2002), HRV is more sensitive to mental effort than heart rate per se.

5.3.2 Examples for the application of measures related to cardiac activity

The review in Deliverable 1 (Roskam, 2002) of the HASTE project refers to findings that fluctuations in the inter-beat-interval (IBI, the time between two peaks in the ECG) were reduced during more demanding mental task performance. Also, the review refers to the findings on the relation between average heart rate during driving and during rest, as well as on using the car phone while driving which resulted in an increase of average heart rate in comparison to a baseline measurement during driving.

Several studies are referred to in Deliverable 1 (Roskam, 2002) of HASTE where one has found evidence on that mental effort investment makes the heart beat more regularly. Certain frequency spans are said to be more related to mental effort (mid frequency band/ 0.10 Hz component). For example, the review refers to experiments where it was found that navigation based on a map was more loading than navigation by vocal messages, as measured by a decrease of the 0.10 Hz component. Also, de Waard (1996) found that HRV profiles provide a reliable reflection of mental effort associated with different tasks.

Heart rate and Heart rate variability were two optional measures in the pilot preparation phase in ADVISORS. However, none of the experiments in ADVISORS measured cardiac activity.

In the HASTE project heart rate and heart rate variability was measured in one simulator experiment and in two field experiments.

In the simulator experiment no effects were found.

In the first field experiment in the Netherlands inter-beat-interval and heart rate variability were measured and calculated. No task effect was found for the auditory task. However, instead an effect was evident when no-driving condition was compared to driving. Also, an effect was found between different road types where the heart rate was higher in the urban area compared to rural and motorway. For the visual task the heart rate variability was reduced when the subjects performed the task, compared to just driving. No effect could be found between the different levels of difficulty for the visual task.

In the second field experiment in Sweden, inter-beat-intervals and heart rate variability were calculated. For the auditory task the durations of the inter-beat intervals were significantly reduced, which indicated an increase of stress. Also for the visual task, effects were found on inter-beat-intervals. Again, a significant effect was found between the difficulty levels and baseline, but not between the different difficulty levels. No effects were found for heart rate variability.

One conclusion from the HASTE experiments on cardiac activity was that inter-beat-intervals and heart rate variability might not be sensitive enough to distinguish between task difficulty levels but rather just task vs. no task.

5.4 Demand for further research

In order to clearly establish if EEG as a suitable measure in traffic research of workload more research is needed. However, EEG will still be very intrusive as a measure. Also, it is difficult to analyze the data and there is a high noise-to-signal ratio which requires calibration to each individual along with rather costly instrumentation.

According to de Waard (1996) the EDA technique can be quite problematic and not very selective as the technique is sensitive to respiration, temperature, humidity, age, sex, time of day etc. In the experiments by HASTE it was evident that the measure could not fully distinguish between different difficulty levels but only between baseline drive (no secondary task) and driving with the secondary task.

Like with the EDA technique (see above), heart rate and blood pressure are sensitive to respiration. Therefore, in order to get reliable data it would be important to measure e.g. physical activity (e.g. if subject are moving, talking) and respiration at the same time as heart rate.

If high sensitivity (e.g. to distinguish between IVIS tasks) is needed, physiological methods and metrics described in this chapter might useful. However, inter-beat-intervals, heart rate variability, and EDA might be used for the assessment of workload with and without and IVIS task.

6   Secondary task methods

This chapter gives a review on the secondary task performance measures.

6.1 Definition and general description of secondary task performance measures

Generally, secondary task performance measures can be taken, when another (related or unrelated) task is added to the primary task (de Waard, 1996). The subject (driver, operator etc.) is required to perform two tasks concurrently - the primary task of interest and the secondary task. With regard to the paradigms which can be applied to dual-task performance, a further distinction can be made (de Waard, 1996):

- Loading task paradigm, i.e. the instruction given to the subject is to maintain performance on the secondary task, even if primary task performance decreases. Here, primary task performance measures can be used for the assessment of mental workload. With regard to safety, driving studies using this paradigm should only be performed in laboratory and not in real traffic tests.

- Subsidiary task paradigm, i.e. the instruction is to maintain primary task performance, even if decrements in the secondary task performance occur. In this case, subject's performance on the secondary task is used to assess mental workload (DOD, 1999). Secondary task performance varies with difficulty and indicates spare mental resources, provided that the secondary task is sufficiently demanding. As dependent variables task, completion time, reaction time and task errors can be used.

With respect to safety driving studies, they are commonly based on the subsidiary task paradigm. Within a driving environment, it is expected that both priority and performance are primarily assigned to the driving task at all times, and that attention to the secondary task is only applied when the driving task is least demanding (Roskam et al., 2002). Therefore, in road traffic research, driving is commonly called the primary task, whereas any other task additionally performed during driving is called secondary task. For a further discussion on this distinction see chapter 2 on driving performance.

In the dual-task situation "driving plus operation of IVIS" the mental workload of the driver is both affected by the demands of the driving task (lateral and frontal vehicle control, traffic situation, environmental factors etc.) and by outputs and operation of the IVIS which withdraw visual and cognitive attention from the driving task. To assess IVIS in terms of safety, the impact of IVIS on the driving task has to be investigated, i.e. how the use of IVIS affects the overall mental workload during driving. Two types of secondary task measures can be used:

- Measures can be taken from the task on the IVIS (operation; perception of messages). The use of this type of secondary task measures may be difficult. If real IVIS are to be evaluated, the necessary additional instrumentation could become complex or the use of the IVIS may be confined. Furthermore, as these tasks vary in a wide field, it could be difficult to come to a general applicable and reproducible method.

- Measures can be taken from a supplementary task which is added to the dual-task situation. In recent years this is the most common type of secondary task measures of which PDT (see below) is a well-known example.

Secondary task performance measures have some drawbacks. The most important one is their intrusion on primary task performance. Since secondary tasks generally compete for attentional demands or resources with primary task execution, this may result in poorer driving performance.

Another drawback is that secondary task performance may be affected by the driver when choosing to allocate more attention to the secondary task than to the primary task or vice versa. Therefore it is recommended that the secondary task should not compete for resources with the primary task (ISO 17287:2003; O'Donnell et al., 1986). Furthermore, the secondary task may lead to a lack of acceptance, if it is too "artificial" in comparison to the primary task (de Waard, 1996). In this case, the secondary task may be omitted, which also can occur in the case where the primary task is high demanding.

An advantage of secondary task measures compared to primary task measures is that secondary task measures ideally are able to differentiate levels of workload that do not impair driving (Jahn et al., 2003). If the secondary task is well suited to the primary task, it is assumed that secondary-task performance is inversely proportional to primary-task performance. If the driver is instructed to allocate enough resources to the primary-task to maintain primary-task performance, the secondary-task is a "subsidiary task" and secondary-task performance reflects changes in primary-task resource demand.

The following secondary task performance measures are frequently used for assessment of driver workload and distraction (de Waard, 1996; Roskam et al., 2002; Jahn et al., 2003):

- Peripheral Detection Task (PDT);
- Paced Auditory Serial Addition Test (PASAT).

Some of the problems with secondary task methods can be overcome, if embedded secondary task measures are used. Embedded task are part of the driver's role in the driving system environment, but have lower priority than the primary task. Acceptance by the driver will be higher than that of "artificial" tasks, and primary task intrusion is expected to be limited (de Waard, 1996; Roskam et al., 2002). More details on embedded secondary task measures are reviewed in chapter 2, as they are close connected to primary task

In some literature the term "secondary task method" is also used for development and use of surrogates of secondary tasks. These surrogates aim at limiting the complexity of secondary tasks and easing the implementation of experimental settings, particularly for the use in early stages of the design process (Breuer et al., 2002). Some of these methods may be in fact beyond the outline of the above mentioned definition of secondary task performance measures, but they will be touched in chapter 6.4 as they include some ideas and resources which could be useful for further optimization and development of secondary task performance measures.

6.2 Peripheral Detection Task (PDT)

Peripheral detection is the measure of the driver's ability to detect visual stimuli presented towards the edge of his field of view (ISO 17287). PDT seems to be able to indicate visual distraction and cognitive workload of different origin. The method is based on the idea that, with increasing workload, the driver's field of view decreases and the attention becomes more selective.

6.2.1 General description of method

PDT requires simple manual responses to visual stimuli usually presented on the left side of the drivers' normal line of sight. The stimuli can be presented by a LED mounted to a headset or projected to the windscreen. A horizontal angle between 11° to 23° is recommended. Stimuli are visible for 1 to 2 seconds and are presented with intervals of a few seconds, e.g., 3 to 5s. For driver feedback a push-button is attached to the index finger and connected to the data storage device. Drivers are asked to respond as quickly as possible to the signal by pressing the push-button (Van Winsum & Hoedemaker, 2000); Jahn et al., 2003). In a TRL simulator study drivers had to respond by pressing the driving

simulator vehicle horn (Brook-Carter et al., 2002). As dependent variables driver response time to the stimulus and number of accurate and missed hits are measured.

PDT method is mainly based on studies of Miura (1986, 1990) and has been further developed by Martens and van Winsum (2000). In recent years it has been used within on-road tests and simulator studies to assess changes in workload during driving, and to assess workload and distraction caused by in-vehicle systems. It is regarded as a candidate for a standard set of safety evaluation techniques.

6.2.2 Applications of the PDT and related methods

In a simulator study Martens and van Winsum (2000) demonstrated PDT's sensitivity to changes in demands of the driving task. Response times increased and hit rates decreased, with large effects being observed for critical incidents like the breaking of the vehicle ahead or an obstacle on the road. PDT was sensitive to workload induced by messages of IVIS. An impairment of PDT performance by speech warning messages was observed.

In a simulator study on collision warning systems PDT confirmed its sensitivity to variations in driver workload (Burns, Knabe, and Tevell, 2000).

To investigate drivers' visual attention Höger (2001) applied a signal detection task. Light points were shown in a video at different positions relevant for driving. Response times and error rates were less deteriorated when signals were presented, overlapping with other road users and traffic signs compared to billboards and irrelevant objects in the periphery. With increase of their virtual distance, light points were detected later. Stimuli in the periphery were detected later than stimuli in the center of the field of vision.

A driving simulator study was performed to evaluate methods for the assessment of ADAS and to identify driving performance while using ACC (Brook-Carter et al., 2002). All Subjects drove with and without ACC and in both traffic conditions (low/high demanding traffic). At a point during driving, participants had to respond to a simple reaction time task. A red rectangle appeared on the simulator screen and the driver was required to respond as quickly as possible by pressing the simulator vehicle horn. As dependent variable response time to the stimuli was measured. Response time increased in heavy traffic in comparison to light traffic and increased when driving without ACC in comparison to driving with ACC, but neither of these effects was found to be significant. It was concluded, that the stimulus detection task was too easy. The stimulus was large and appeared directly in front of the participants. It was presumed that a PDT task might have been more sensitive to measure workload and attention.

In EU project ADVISORS (Wiethoff, 2003) an inter-urban ACC system was assessed in a driving simulator. The focus was on driver's acceptance and driving performance in response to the presence of the ACC. A study was performed in order to gain further knowledge of possible influence from measurement devices (PDT; heart rate measures) on obtained effects of ACC use. A PDT device according to Martens and van Winsum (2000) was used. Simulator sessions were performed only on motorway drives and only with one setting of ACC minimum time headway (1sec.). Each participant drove with and without ACC support in balanced design. Heart rate was sensitive to ACC. For PDT measure no effects of ACC were found, i.e. PDT did not discriminate driving with ACC from driving without. It seemed that PDT test affected speed, but this question needs further research.

In a real traffic study by Olsson and Burns (2000) LED signal projections to the windscreen in an area of 11 to 23° to the left side of the drivers' normal line of sight and 2 to 4° above the horizon were used for PDT. Signal duration was 1 sec. and the interval between stimuli was 3 to 6 sec. In 30 s intervals surrounding the tasks PDT performance suffered from radio tuning and even more from backward counting and CD changing. No difference was found between driving on country roads and on a motorway.

A different detection task has been used in driving studies by Verwey (2000). The stimuli were digits from 20 to 99. They were presented on a display located at 27° horizontally to the right and 20° above the normal fixation point on the road ahead. Stimuli were presented for 750 ms and with intervals between stimuli varying from 2 to 4 s. Participants had to respond to the number stimuli verbally. During the intervals between the number stimuli the letters "GG" were presented on the display "to prevent peripheral detection of stimuli". Detection performance was sensitive to the demand of different traffic situations, e.g., driving straight ahead at intersections with priority vs. without priority or driving straight ahead vs. turning at uncontrolled intersections.

In the EU project COMUNICAR on-road tests were performed to verify an information managing system and an integrated multmedia HMI (Schindhelm et al., 2003). PDT was applied to study the interactions of the Information Manager with the criticality of the predefined scenarios. The Scenarios were varied through road demands, driving task and operation on the IVIS thus inducing different levels of mental load to the driver. Operation on the IVIS included incoming messages of the phone, mail, SMS and HMI input to the radio, phone. The expectation was that driving with the information manager "on" would decrease driver workload and thus have a positive effect on PDT performance in comparison to driving with the information manager "off". A head-mounted LED signal (1sec., interval 3 to 5 sec.) and a response button attached to the index finger were used. Response times and missed signals were measured. The PDT results confirmed the findings on the Information Manager which showed to be sensitive to the different risk levels of the scenarios. A strong effect was found for the factor "Information Manager". Larger driver responding times and a higher number of missed PDT signals indicated that, when driving without Information Manager, drivers have less spare capacity to react in time to the PDT signals. Some drivers reported that signal detection was very hard during situations where bright sunlight glared the drivers. For some subjects artifacts in PDT performance were found in the data and had to be removed (missed hits; additional hit not associated to stimulus).

The IHRA-ITS Swedish-German (Jahn, 2003) study aims at effects of driving complexity and also evaluates route guidance systems. In the present study taxi drivers performed the PDT in inner city traffic following the instructions of route guidance systems. Route sections differing in complexity and demands of the driving task were compared. Two route guidance systems, one with a small display, the other with a larger one, were used with "full" guidance (verbal and visual), to test whether the subtle difference in display size would show up in PDT performance. The PDT LED signal was projected to the left side of the windscreen at a horizontal angle of 11° to 23° and a vertical angle between 2° to 4° above horizon. The location of the signal (duration max. 2 sec; interval between signals 3 to 5 sec.) varied randomly. Participants responded by pressing a push-button attached on their index finger. PDT hit rate and response time were taken and collected on a PC. The mean hit rates for both systems (large/small display) were 93.5 %. Hit rates were significantly lower for high demanding route sections than for low demanding route sections. No significant effect was found for the factor display size. Mean response times increased on high demanding route sections (683 ms) in comparison to low demanding routes (567 ms). Similar to the findings with the hit rates the response time results proved to be significant for the factor route demand but not for the factor display size. The effect of route demands and the effect of route guidance could not be separated within the chosen test design. The difference in display size was not reflected by PDT. The findings also show the importance of clearly and conclusively defined route sections that are comparable between subjects. When varying both variables, route and the design of the IVIS, effects might get blurred. It is recommended to select low-demand sections as baseline in the case of evaluating IVIS. The results give evidence for the PDT´s sensitivity to short lasting workload peaks at different levels of workload, not only when traffic elevated visual and mental workload, but also during IVIS output.

6.2.3 Requirements for further research

A number of advantages of the PDT have been reported (Jahn et al., 2003):

- The simple responses can easily be performed during most driving scenarios and the PDT does not consume resources needed for safe driving. The PDT therefore is unobtrusive with regard to driving and both suitable for studies in real traffic and in the laboratory.
- It can detect short peaks of workload that may be missed by methods that integrate over longer intervals, thus having a favourable bandwidth.
- It proved sensitive to differences in driving demands and to effects of in-vehicle systems, in simulator studies as well as in real traffic tests.
- The equipment is simple and inexpensive and data analysis is quick and straightforward.
- Furthermore, peripheral visual stimuli are related to objects and events that have to be noticed during driving, therefore validity is claimed for the PDT.
- For safety evaluation it is useful that PDT performance is affected by both general and by selective withdrawal of attention and that PDT seems to be able to indicate visual distraction and cognitive workload of different origin.

The following drawbacks and weak points of PDT have to be considered:

- Diagnosticity is weak as the PDT can hardly differ between visual distraction and cognitive workload. Furthermore, PDT performance is affected both by attention allocated to the driving task and the attention paid to IVIS; but PDT is not able to differ between workload induced by driving on the one hand, and distraction and workload induced by the IVIS on the other hand. For future research, it is important to find out which mode of IVIS output and which source of workload affects PDT performance and in which way (Jahn et al., 2003).
- The IHRA-ITS study (Jahn et al., 2003) found that PDT is able to distinguish between IVIS which differ in the support of strategic and self-paced time sharing of visual attention. Thus the authors presume that PDT might be sensitive to systems which differ in the support of task switching. Research has to go further in this question.
- It seems that PDT performance is affected by eye and head movements in settings where the stimuli are projected to the windscreen. For example, at an intersection PDT performance might suffer because the signal is not in the driver's field of view. But this drawback can be solved by using head-mounted LED signals as applied in COMUNICAR (Schindhelm et al., 2003).
- PDT performance is presumably affected by covert shifts of attention which may be caused by an in-vehicle task or by traffic situations. For example, at an intersection PDT performance might be influenced because of gazing behaviour required for safe driving. In this case PDT´s sensitivity to different outputs of IVIS might suffer.
- The impairment of PDT performance resulting from high workload seems to be mediated by restricted patterns of visual search (e.g. see Recarte & Nunes, 2000 and Verwey, 2000). If the stimuli were not presented on the windscreen but on a head mounted LED, some of the problems mentioned in previous studies could have been avoided. The implications of the interaction of PDT performance and gazing behaviour have to be studied further.
- Visual detection performance is influenced by many variables that may be difficult to control especially in field studies. Lighting conditions and background of stimuli are reported as problems: Visibility of PDT signals is impaired in bright sunlight, and traffic lights as well as braking lights were mistaken for PDT stimuli in previous studies (Schindhelm et al., 2003; Jahn et al., 2003). Even if luminance, size, location and presentation parameters of PDT stimuli are controlled, there remains inter-individual variance in detection ability, response speed, driving competence and performance with in-vehicle systems.
- Performance of PDT may be affected by the acceptance of the subjects. Missed hits as well as additional hits not associated with a stimulus may be a result of decreasing acceptance of the subjects (Schindhelm et al. 2003). It is not always possible to identify those "artefacts" in the data file. Presumably, an enhanced PDT of a choice reaction time task type may be a solution to minimize these artefacts (see chapter 6.4).

6.3 Paced Auditory Serial Addition Test (PASAT)

The PASAT (Gronwall, 1977) is a measure of cognitive function that assesses information processing speed and flexibility, as well as calculation ability. It was developed to assess psychological effects of concussion and later adapted for clinical studies on multiple sclerosis. De Waard (1996) used the PASAT in a car-phone study during driving.

Single digits are presented every few seconds, either orally or visually, and the subject must add each new digit to the one immediately prior to it. To ensure standardization in the rate of stimulus presentation the PASAT is presented using audio cassette tape or compact disk. Normally an interval of 3 sec. has been used. Varying inter-stimulus intervals affects the difficulty of the task and enables to fine-tune the task to the subject's capabilities. Administration time is approximately 15 minutes. The score for the PASAT is the total number of correct answers out of 60 possible answers.

The PASAT is administered in person by a trained examiner, since subtle changes in the administration procedures can have a significant effect on the outcome. Performance on the PASAT is sensitive to practice effects, that is, patients often display poorer performance when first tested due to lack of familiarity with the task.

6.4 Potentially suitable secondary task methods

A summary of potentially secondary tasks methods which test perceptual and cognitive abilities is given by Roskam et al. (2002). The tests described are considered suitable tasks in assessing the interaction between driving and secondary task performance.

In a **visual choice reaction time task**, a different response has to be made to each particular stimulus which is given on the screen, e.g. the right button must be pressed for even numbers and the left one for odd numbers. In a study (Verwey, 2000) this type of secondary task was used including the detection of numbers between 20 and 99 presented at a random rate on the corner of the dashboard.

Via headphones a series of distinguishable sounds are presented in an **auditory detection task**. Subjects are required to press a specific button for each particular sound. Task difficulty can be manipulated by the number of auditory stimuli, presentation rate, and randomized presentation of stimuli.

Many tests of **spatial memory and perception** require a rather high degree of visual attention and are considered unsuitable as secondary tasks during driving.

Finger tapping tasks have been used in dual-task situations (e.g. Smyth et. al, 1994). The effect of driving on finger tapping tasks can be examined by comparing tapping accuracy during different scenarios. Accuracy is monitored by number of keys pressed, conformity with the required pattern and keeping a regular tapping pace. The advantage of tapping tasks is that workload can be manipulated by setting the task from simple to random tapping. Care must be taken to ensure that tapping does not impair driving task with regard to manual resources.

Continuous memory tasks (CMT) can be presented both auditory and visual. A button is pressed every time a target letter is presented, and pressed twice if the target letter has been presented more than once. The level of difficulty increases by increasing target letters. Load can be assessed by using either a manual or a vocal response.

In the **random number generation (RNG)** task subjects are required to randomly generate a list of numbers from a specific list (e.g. any number from 1 to 10). As a control the list has to be uttered by the subjects. The index of randomness is reduced if subjects are asked to perform a demanding task during random number generation. The administration of this task during driving is reasonably simple, although task difficulty is arbitrarily chosen by the subject.

7   Subjective assessment methods

The following chapter will present and discuss a range of methods which often are called 'subjective' or 'self reported' assessment methods in the literature. The chapter will follow the following structure:

1) Self reported (level of workload):
   a) One dimensional (section 7.1.1), global rating of workload which provide a measure more sensitive to manipulations of task demand
   b) Multidimensional (section 7.1.2): preferable when diagnosticity is of great concern
2) Self reported driving performance measures (section 7.2)
3) Expert reported driving performance measures (section 7.3)

Negative and positive aspects with the methods will be brought up. However, there exist some general positive aspects of workload which often are true for most of the methods in this chapter:
- High face validity
- Application Ease
- Low costs
- Low primary-task intrusion secured as long as the scale is administered after completion of the task

General negative aspects would be (according to O'Donnell et al. 1986):
- 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. E.g. the subjective and objective measures do not always correlate.
- Subjective Measures are limited in identifying conflicts among resource assignment, as both peak and average workload levels could be used as an index of workload by a user, who could lack the experience to correctly assess workload.

7.1   Self-Reported workload measures

Self-Reported workload measures refer to the concept that a user reports his/her vision of workload during a trial. The subjectivity of the reporting mechanism provides great insight of individual Workload, but also is prone to misunderstandings and biases. In order to obtain correct measurements, the used techniques must be validated experimentally, to give a simple way for users to understand the meaning of different dimensions of workload, thus obtaining sound data. To measure workload, experts developed several *scales*, either numerical or descriptive. The user marks a point between the minimum and maximum of a given scale, and the mark represents the amount of workload for the user. Several dimensions of workload can be assessed, or just an overall result can be taken into account. Below multi-dimensional scales in the first case and one-dimensional scales in the second case will be presented.

Muckler & Seven (1992) also stress that self-reported scales should be as simple as possible, immediate and comprehensible, to reduce the need for interpretation and to aid in the precision of measure definition. This is mainly true for one-dimensional scales.

7.1.1   One-Dimensional Scales

As described above, 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.

When assessing workload by using one dimensional scales, multiple aspects of workload can be assessed separately, using an one dimensional scale for each dimension. This way, multiple aspects of workload can be assessed through a one dimensional scale. Results are then aggregated in some way. Zijlstra and Meijman (1989) used the RSME (see section 7.1.1.3 below) in this way, asking people to rate different dimensions of task performance separately. In this study a RSME rating was obtained by rating the effort required to perform different sub-tasks, such as navigation, machine-use and communication. The advantage of this method is that a more differentiated picture emerges. It can be argued, however, that multiple use of a one-dimensional scale in this way is not fundamentally different from multidimensional scales. (de Waard, 1996).

Below, the following one-dimensional scales will be presented: MCH, OW and RSME.

7.1.1.1 Modified Cooper-Harper Scale (MCH)

The MCH is a one-dimensional scale consisting of ten items that add to a single score (Cooper et. al., 1969). The Cooper-Harper scales were developed for the purpose of assessing aircraft handling characteristics. The major advantage of using the Cooper-Harper Scale is that it is easy to administer; it is popular in the scientific community, and the resulting ratings highly correlate with other workload scales (Charlton, 1996; Lysaght et al. 1989). Wierwille and Casali (1983) modified the Cooper-Harper Scale, calling it the Modified Cooper-Harper (MCH) Scale. This one can be used for applications beyond that of aircraft control, as was the original one. Additionally, the MCH allows the experimenter to measure workload associated with tasks that require the participants perceptual, cognitive, and communication skills. The rating scheme is synthesized below in Table 7.1.

Table 7.1 MCH Rating Scheme

Difficulty Level	Subject Demand Level	Rating
Very Easy, Highly Desirable	Operator Mental Effort is minimal and Desired Performance is easily attainable	1
Easy, Desirable	Operator Mental Effort is low and Desired performance is attainable	2
Fair, Mild	Acceptable Subject Mental Effort is required to attain adequate system performance	3
Minor but Annoying	Moderately high Subject Mental Effort is required to attain adequate system performance	4
Moderately Objectionable	High operator mental effort is required to attain adequate system performance	5
Very Objectionable but tolerable	Maximum Subject mental effort is required to attain adequate system Performance	6
Major	Maximum Subject Mental effort is required to bring errors to moderate levels	7
Major	Maximum Subject mental Effort is required to avoid large or numerous errors	8
Major	Intense Subject mental Effort is required to accomplish task, but frequent or numerous errors persist	9
Impossible	Instructed task could not be accomplished reliably	10

Strong points of this method are:
- Sensitive to variations in task difficulty
- Very easy to administer
- Experimentally proven that results achieved with this method correlate with those obtained with other WL scales.

7.1.1.2 Overall Workload (OW)

OW (Vidulich & Tsang, 1987) is a bipolar scale, ranging from 0 (very low) to 100 (very high), with steps of 5 units. It is used to assess a global measure of Workload, and therefore it is easy to administer. Its reliability is lower than NASA-TLX, but it is proven more sensitive and acceptable than MCH and SWAT. In addition, among One-Dimension scales, it is the more reliable, along with RSME (Chin & Nathan, 2002).

Strong points of this method are:
- Easy to use
- More Sensitive than MCH and SWAT
- More Acceptable than MCH and SWAT

7.1.1.3 Rating Scale Mental Effort (RSME)

The RSME (Zijlstra & Van Doorn, 1985) is a one-dimensional scale in which ratings of *invested effort* are indicated on a line (from 0 to 150). Along the line are a number of anchor points that are labelled with a verbal descriptor of effort. It is easy to administer both after and during driving.

In comparison with other techniques for measuring workload, RSME is one of the most sensitive measures of workload (Verwey & Veltman, 1995). A higher degree of invested effort is considered as an attempt of the driver to keep performance on a certain level in response to increased task demands. It means that, when effort gets higher, it is because the driver tries to keep performance high; consequently, if effort becomes too high, performance can degrade because the resources are not sufficient to maintain a safe behaviour. The technique has a high reliability since it consistently results in higher workload ratings as a function of task load. It can be assumed that at a certain level of task demands, performance of the primary task may deteriorate if a certain amount of effort is exceeded. Because of this, high levels of invested effort are considered detrimental for driver safety. This suggests that a high validity is assumed with RSME. Then again, this needs to be demonstrated in experimental research.

On the RSME the amount of *invested effort* into the task has to be indicated and not the more abstract aspects of mental workload (e.g., mental demand, as is in the TLX). These properties make the RSME a good candidate for self-report workload measurement. Below is a representation of RSME. A graphical representation is depicted in
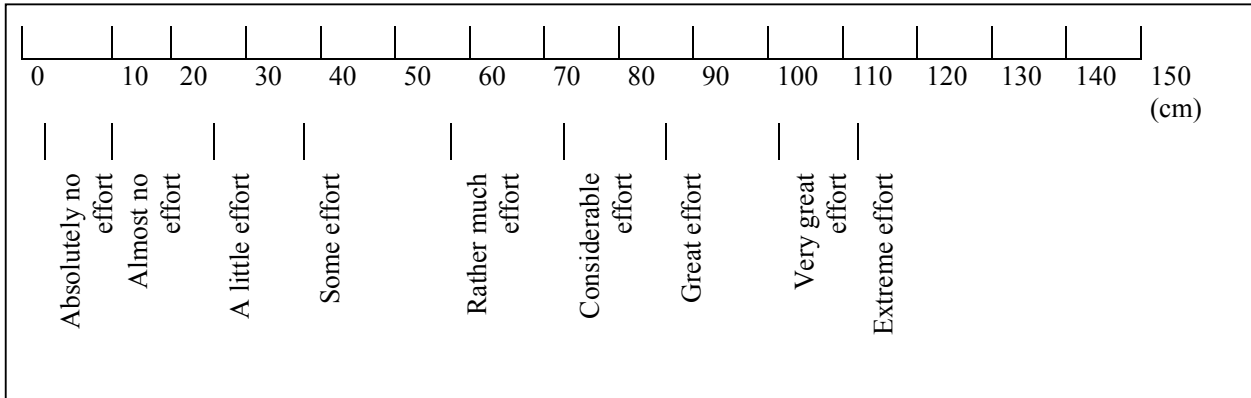Figure 7.1.

Figure 7.1 Graphical Representation of RSME.

Strong points of this technique are:
- 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 (i.e. when task load increases so does workload ratings)
- Not intrusive
- Good Acceptability

Veltman and Gaillard experimentally proved RSME more sensitive than NASA-TLX (de Waard, 1996, p33).

7.1.1.3.1 *Examples of applications of the method*

In COMUNICAR the Information Manager was assessed by RSME which was calculated twice, before the testing and after the testing.

In ADVISORS RSME was used in an evaluation of the Lateral Support System (LSS). RSME showed quite low values for the driving scenarios developed: High Traffic Situation and Low Traffic Situation. The effort ranged from 0 to 24 for High Traffic Scenario and from 0 to 21 for Low Traffic Scenario. This result might be due to a floor effect. The driving tasks might have been too easy, i.e. the route might have been not long enough and the creation of the two experimental conditions 'high traffic density' might have been not successful. The last assumption is supported by the finding that the actual traffic density for each condition is rated by the test leader on average only as 'more low' or 'more high' respectively.

7.1.1.4 One dimensional scales: Issues for further research

In the section below general issues which should be considered in future research is presented for the some of the one dimensional scales presented above.

For the Modified Cooper-Harper Scale the following should be taken into account:
- Not suited for measuring short-lasting variations in WL when driving
- Less sensitive than NASA-TLX or RSME
- Validity to be fully determined in experimental research
- Low reliability compared to other more popular methods

As described in Chin & Nathan (2002), the issues below have been identified during applications of the Overall Workload scale:

- Less valid than NASA-TLX
- Less reliable than NASA-TLX
- No Diagnosticity

7.1.2 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 which factors are more meaningful for workload in a given case, workload 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 workload, which is useful to understand which parts of the system need revision, and to have an overall value, useful to compare different systems. Many of the methods described here derive from the most used and common, NASA-TLX, as an improvement or specification. In the section below NASA-TLX, PSA-TLX, DALI, SWAT, SWORD will be presented.

7.1.2.1 NASA-TLX

The NASA Task Load Index is a multi-dimensional rating procedure that provides an overall workload score based on a weighted average of ratings on six subscales (see Table 7.2).

Table 7.2 NASA-TLX Dimensions

	Dimension	Endpoints	Description
Demand to Subject	MENTAL DEMAND	Low / High	How much mental and perceptual activity was required (e.g. thinking, deciding, calculating, remembering, looking, searching, etc.)? Was the task easy or demanding, simple or complex, exacting or forgiving?
	PHYSICAL DEMAND	Low / High	How much physical activity was required (e.g. pushing, pulling, turning, controlling, activating, etc.)? Was the task easy or demanding, slow or brisk, slack or strenuous, restful or laborious?
	TEMPORAL DEMAND	Low / High	How much time pressure did you feel due to the rate or pace at which the tasks or task elements occurred? Was the pace slow and leisurely or rapid and frantic?
Interaction with task	PERFORMANCE	Perfect / Failure	How successful do you think you were in accomplishing the goals of the task set by the experimenter (or yourself)? How satisfied were you with your performance in accomplishing these goals?
	EFFORT	Low / High	How hard did you have to work (mentally and physically) to accomplish your level of performance?
	FRUSTRATION	Low / High	How insecure, discouraged, irritated, stressed and annoyed versus secure, gratified, content, relaxed and complacent did you feel during the task?

An earlier version of the scale had nine subscales. It was designed to reduce between-rater variability bias (that is, the difference between two different dimensions) by using the *a priori* workload definitions of subjects to weight and average subscale ratings. The technique (referred to as the "NASA Bipolar Rating Scale") was quite successful in reducing between-rater variability, and it provided diagnostic information about the magnitudes of different sources of load from subscale ratings (Hart et. al., 1984; Vidulich et. al., 1986). However, its sensitivity to experimental manipulations, while better than found for other popular techniques and for a global one-dimensional

workload rating, was still not considered sufficient (HSIAC, 2004). In addition, it was felt that nine subscales were too many, making the scale impractical to use in a simulation or operational environment. Finally, several of the subscales were found to be irrelevant to workload (e.g. Fatigue) or redundant (e.g. Stress and Frustration). For these reasons, the NASA Task Load Index was developed. Some of the subscales from the original scale were revised or combined, others deleted, and two added. Three dimensions relate to the demands imposed on the subject (Mental, Physical, and Temporal Demands), and three to the interaction of the subject with the task (Effort, Frustration, and Performance).

Since subjects can give ratings quickly, it may be possible to obtain them in operational settings. However, a videotaped replay or computer regeneration of the operator's activities may be presented as a mnemonic aid to the subject that can be stopped after each segment, to obtain ratings retrospectively. It was shown in a helicopter simulation and in a supervisory control simulation that little information was lost when ratings were given retrospectively: a high correlation was found between ratings that were obtained "online" and those that were obtained retrospectively with a visual re-creation of the task (Hart et. al., 1986; Haworth et. al., 1986).

NASA-TLX has been tested in a variety of experimental tasks that range from simulated flight to supervisory control simulations and laboratory tasks (e.g. the Sternberg memory task, choice reaction time, critical instability tracking, compensatory tracking, mental arithmetic, mental rotation, target acquisition, grammatical reasoning, etc.). The results of the first validation study are summarized in Hart et. al. (1987). The derived workload scores have been found to have substantially less between-rater variability than one-dimensional workload ratings. Moreover, the subscales provide diagnostic information about the sources of load.

See Table 7.3 for a summary of the method:

Table 7.3 Summary of NASA-TLX

1	Subjects are presented with C_2^6= 15 pairs of Dimensions, that are all the possible combinations of the 6 dimensions; the Subjects are asked to choose which of the two has a major impact of Workload;
2	Subjects are presented with 6 scales, one for each Dimension, ranging from 0 to 20, and are asked to rate that dimension according to overall Workload;
3	An overall Workload Index is computed: a. Each dimension d_i , $i=1,...,6$ is assigned a *weight w_i , $i=1,...,6$*: this is the sum of the choices of dimension d_i in pairs confrontation; b. Each rating, r_i, $i=1,...,6$ is multiplied by 5, giving a scaled rating x_i, $i=1,...,6$; c. Overall Workload Index is computed by the following formula: $\text{OWI} = \frac{1}{15}\sum_{i=1}^{6} w_i x_i$, which gives an Overall Workload Index ranging from 0 (where all ratings are 0) to 100 (where all ratings are 20)

NASA-TLX is useful for measuring workload over a longer period of time, but is not suited for detecting peaks or short lasting increases in workload. It has also been proven to be sensitive to differences in workload in a number of studies of car driving. Studies have shown that it is more sensitive to workload than other multidimensional subjective workload scales, such as the MCH and SWAT discussed below (Hill et al., 1992). However, there is no evidence that it is more sensitive than one-dimensional scales such as the RSME, which are easier to administer. It is assumed that a higher mental workload is detrimental to driver safety, although the experimental evidence on the relationship between NASA-TLX performance and driver behaviour and critical incidents is lacking. The validity of this measure still needs to be established in research. Nonetheless, since NASA-TLX is consistently sensitive to workload increases in different studies it is regarded as a reliable method of measuring workload.

Byers et al. (1989) proposed a Raw Task Load Index (RTLX), simple average of the six TLX scales. This method does not require task-paired comparison, and in comparable to TLX according to means and standard deviations. It is, therefore, a simpler alternative to TLX. This result is further confirmed by Fairclough (1991).

Strong points of this method are:
- 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 identifying causes of workload

7.1.2.1.1 Examples of applications of the method

NASA-TLX could be said to be the starting point of different other methods to assess workload (some of which are described below). Given its power, it is a widely used tool. Below some examples are given where the method has been used.

In IN-ARTE NASA-TLX was used to assess a wide range of driver tasks: navigation, keeping distance to preceding cars (by means of adaptive cruise control), lane keeping and Collision Avoidance.

In ADVISORS e*valuation of the Lateral Support System the results showed that* under both Low Traffic and High Traffic conditions, subjects experienced only little effort with no difference between both experimental conditions (T22= 1,476; p= .154). This result might be due to a floor effect where the driving tasks might have been too easy, i.e. the route might have been not long enough and the creation of the two experimental conditions 'high traffic density' might have been not successful. The last assumption is supported by the finding that the actual traffic density for each condition is rated by the test leader on average only as 'more low' or 'more high' respectively.

Fairclough et al (1991) performed a test consisting in driving and having a conversation. Results showed an increase in overall WL when the conversation took place;

Vaughan et al (1994) performed a test where three different modes to present information were used: only auditory, auditory and continuously visible on a display, auditory and temporarily visible on a display. Overall RTLX was lowest for the second method. This test allowed to show RTLX diagnosticity, because of the higher scores of the time-pressure factor.

Alm & Nilsson (1994), performing a test with a phone task during driving, observed an incidence on all TLX scales.

In a study by Gunn (2001) the impact of cognitive distraction on driver visual behaviour and vehicle control was assessed. This study examines driver distraction impact on driving behaviour, as more than 20% of car incidents' primary cause appears to be distraction. The test is conducted by examining driver workload, taking into account modifications of braking usage in different workload levels, according to pressure intensity (>0.25g, >0.30g). Results in applying NASA-TLX to assess workload, perceived reduction of safety and distraction, indicate that workload perception increases as the complexity of the cognitive task. Mean ratings of workload were 1.94 (SD=.87) for No Task, 3.55 (SD=1.62) for Easy Task and 5.73 (SD=1.33) for Difficult Task Conditions. Significant differences in workload were found for all comparisons among the three conditions.

7.1.2.2 PSA-TLX

This method, developed by PSA Peugeot Citroen, is a mix between NASA-TLX and RSME, as described below. The purpose is to have a method with the following characteristics:

- Easy to use for experimenter and for drivers
- Cost-effective
- Not intrusive
- Able to identify components of workload
- Able to identify impact on driving task
- System-independent

After a focus group of experts (psychologist and ergonomists) and non-experts, where identification of mental demand sources has been brought, seven dimensions, ranging from 0 to 100 were defined to assess driving workload, as well as four factors increasing workload (see Table 7.4).

Table 7.4 PSA-TLX Dimensions and Factors

Dimensions	Factors
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 - aspect of WL related to a mental representation of itinerary, based upon extraction of info about the itinerary from a map, road signs, written instructions. 6) Appropriate use of controls and driving equipment 7) Reactivity to safety and status signs	1) Stress (tension, constraint, apprehension, unsafe feeling) 2) Fatigue 3) General Dissatisfaction (discontent, disappointment) 4) Discouragement (loss of motivation, interest missing)

Three axes to measure Driving Workload are used: *Effort*, *Disruption*, *Driver State*. These axes have a graduated scale from 0 to 100 (5 by 5) including six verbal reference marks**.** Marks are used to help drivers giving sense to graduations and reduce variety in interpretation of values. Their use is inspired from RSME.

The seven driving dimensions are evaluated on two axes, Effort and Disruption**.** The four factors about driver state are evaluated on one axis (State)**,** completed by two questions**,** determining if the factor is the *source* or the *consequence* of the effort spent. The Evaluation takes place immediately after the task achievement; the experimenter explains the questionnaire, to make sure of driver's understanding of questions. It is important that each driver marks each question, that is, they evaluate each dimension and factor on each axis. 2 scores for each Task Dimension (Score Effort, Score Disruption) and one score for each State Factor are given. Any comment must be noted by the experimenter who should attract them. In the end, the experimenter fills in the questionnaire. Having a situation as a reference is recommended to compare with target-situations.

Strong points of this method are:
- 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)

7.1.2.2.1 *Examples of applications of the method*

So far the PSA-TLX has only been applied in PSA's own research:
- PSA Peugeot Citroen Subjective Assessment Methods; PSA – Proposition of a new Method presents a new method (PSA-TLX) based on NASA-TLX to assess Driving Workload. It is just a proposition with an application to two internal experiments.

7.1.2.3 DALI

DALI (Driving Activity Load Index) is a modification of NASA-TLX, focusing on specific dimension to take into account driving task (Pauziè et al, 1996). The principle is the same, with a scale rating procedure for six pre-defined factors, followed by a weighing procedure in order to combine the six individual scales into a global score.

The main difference is the choice of the main factors composing the workload score:
- Physical Effort (such as turning, pulling, pushing, suitable in aviation field, but hardly in automotive with experienced drivers and electronic nowadays cars) can be ignored;
- Mental Effort consists in activities such as thinking, deciding, calculating, remembering, looking, searching, and so forth, and it might be interesting to differentiate those modalities;
- Performance can be assessed by means of objective measurements.

A team of experts led to the definition of six Workload Dimensions, as defined in the following table below (Table 7.5).

Table 7.5 DALI Dimensions

Dimension	To evaluate the
Effort of Attention	Attention required by the activity (think about, decide, choose, look for…)
Visual Demand	Visual demand required for the activity
Auditory demand	Auditory demand required for the activity
Temporal Demand	Specific constraint due to timing demand when running the activity
Interference	Possible disturbance when running the driving activity simultaneously with any other supplementary task such as phoning, using systems or radio and so forth;
Situational stress	Level of constraints / stress while conducting the activity (fatigue, insecure feeling, irritation, discouragement…)

After rating for each Factor, the paired comparison procedure is conducted in the same way as TLX scoring. The computation of the weighed rate for each of the six dimensions allows getting a score corresponding to the Subjective Evaluation of workload by factor.

Strong points of this method are:
- Same as NASA-TLX:
  - Workload measurement over a long period of time
  - More sensitive than MCH and SWAT (according to RoadSense, where DALI focuses on more specific aspects of automotive field, so DALI appears to have the same characteristics TLX has.
  - Reliable WL measuring technique due to consistency sensitiveness to increases in workload in different studies
  - Weighting increases sensitivity and allows to identify causes of workload
- Specific for Automotive application

7.1.2.3.1 Examples of applications of the method

There are a range of examples of the application of the DALI where some are described below.

DALI was applied to evaluate navigation system to subjects with presbyiophia. The system had auditory and symbolic messages displayed (Pauziè et al., 1996). The results from this experiment are summarized below:

- Stress, Attention: workload introduced by unusual situation of using a driving aid system;
- Vision: Using messages displayed on-screen not considered Heavy
- Auditory, Temporal: Not always right timing, difficult to follow Navigation instructions;
- Interference: Clear indication of interference with the guidance situation, also confirmed by verbal comments during and after the session, and on observation of subjects' behaviour.
- Young Subjects showed better results on Auditory Demand

DALI was also used in an experiment in which a hands-free car phone was evaluated. The results indicated that:

- On the Simulator, the Effort of Attention and the Visual Demand are the main increasing components;
- Under Real Driving situation, the factor Interference, and, to a lesser extend, the Auditory Demand, have been evaluated by the Subjects as the main factors that were at the origin of Mental Workload.

7.1.2.4 SWAT

SWAT (Subjective Workload Assessment Technique, (Vidulich et al., 1986) is a multidimensional scale relying on three Loading dimensions:

1. Time
2. Mental Effort
3. Psychological Stress

Such values are discrete, and range from 1 to 3 each. It was originally designed to assess the workload associated with the operators' activities in aircraft cockpits and other crew-station environments to assess the workload associated with the operators' activities.

When using the SWAT, subjects must perform a Card Sorting pre-task procedure (CS), followed by a task-scoring procedure. During CS, 27 SWAT cards are to be ordered. Such cards are the outcome of combinations of the three discrete dimensions at three discrete levels. For each dimension, the levels have descriptors that represent the lowest mental workload (level 1) to highest mental workload (level 3). Subjects rank such cards, from lower mental workload to higher mental workload.

After CS, the level of coherence among all subjects is calculated using Kendall's coefficient of Concordance (W). Based on the relative importance of each dimension (time, psychological stress, effort) six hypothetical orders exist (TSE, TES…). Subjects are asked to rate the three dimensions of a task using a scale of 1 to 3 points. According to subjects ordering, the most suitable ordering is chosen, and a conjoint workload scale, ranging from 0 to 100, is developed. This scale implies correlation among the three dimensions. Subjects are asked to rank them in increasing workload levels. Such correlation is yet to be proven valid.

Strong points of this method are:

- Reliable
- Sensitive
- More Sensitive than MCH Scale

7.1.2.4.1 Examples of applications of the method

Janssen et al (1994) reported different SWAT ratings when using or not using a GIDS system, which supported drivers by means of route guidance messages related to speed, Collision Avoidance and lane keeping. Results ranged from 3 to 9.

Verwey & Veltman (1995) found that card-sort task for SWAT did not produce more accurate WL estimation, and that SWAT ratings were as sensitive as RSME in WL increases.

7.1.2.5 SWORD

SWORD (Subjective WORkload Dominance) measures workload of different tasks as a series of relative subjective judgements compared to each other.

SWORD has three steps:

1) *Collect subjective between-tasks comparative ratings using a structured evaluation form (see example in Figure 7.2) after the subject has finished all the tasks.*

Figure 7.2 An example of a SWORD evaluation form (Vidulich et. al., 1991)

All possible paired combinations of the tasks are listed in one row, with one task on the leftmost side and the other task on the rightmost side. There are 17 comparative levels between any two tasks as shown in figure. Namely, nine levels are assigned to each task with "EQUAL" belonging to both tasks in one row. A Nine levels scale is regarded as a reasonable upper limit for the response since "seven plus or minus two items represents a pervasive limit in cognitive processing" (Vidulich et al., 1991). If the subject perceived that the two tasks induced the same level of workload, then he/she marks on the "EQUAL" slot in the corresponding row of the evaluation form. If he/she perceived either task had workload dominance over the other, he/she marks a slot closer to the dominant task according to the level of workload difference.

2) *Construct a judgement matrix based on the subjective ratings.*
The diagonal of the SWORD judgement matrix (see Figure 7.3 below), is filled with ones representing tasks which are compared to themselves. The upper right triangular area is filled with the subjective ratings (possible value are 2 to 9) of the dominance of the task in a row over the task in a column extracted from the evaluation form. The lower left triangular area is filled with the reciprocals of the numbers in the diagonally symmetric cells of the upper right area.

Figure 7.3 SWORD judgment matrix

3) Calculate the relative ratings for each task.
The judgement matrix is subject to a *Consistency test*. Consistency means that if, for example, the rating of a task A is twice as much as the rating of task B, and the rating of a task B is three times as much as the rating of a task C, then the rating of A should be six times as much as the rating of task C. A measurement of consistency (S^2) can be calculated using Williams and Crawford's method (Williams et al., 1980) and compared to a critical value from a table developed by Budescu et al. (1986). If the S^2 of a subject's judgement matrix is greater than the critical value, then the judgement matrix is too inconsistent to be included in the final workload measurement.

The final rating for each task is the normalized geometric mean for that row of the judgement matrix. The rating represents a ratio scale of a task's Workload level compared to all other tasks.

Strong points of this method are (as described in Chin & Nathan (2002)):
- Sensitive
- Reliable
- Easy to use

7.1.2.6 Multi dimensional scales: Issues for further research

For the multi dimensional scales presented above a range of issues have been identified for future research.

For NASA-TLX those issues are:
- 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. (de Waard, 1996)
- 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

For the DALI scale the following should be considered:
- 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

For the SWAT method the following was identified:
- High time demand for card sorting
- More relevant for studies focused on 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

And finally for SWORD as pointed out in Chin and Nathan (2002):
- Depends on the number of tasks to compare
- Scale format problems: Definition of workload may be different for subjects, factors evaluated may not be the same

7.2 Self-reported driving performance measures

The purpose of this kind of measures is to understand driving performance by means of self-reporting means. Data about errors and violations as well as an overall judgement about the driving session is given by drivers themselves. This kind of measurement can be used as a predictor of accident involvement. Users are tested with Questionnaires and Scales.

7.2.1 Driver Behaviour Questionnaire

The Manchester Driver Behaviour Questionnaire (DBQ) operates a division of driver behaviour in:
- Errors, (errors of omission or commission resulting from a lack of knowledge or information)
- Lapses (inadvertent or inappropriate occurrences of highly practiced behaviours)
- Violations

Such definition can be specifically found in Norman (1981), Rasmussen (1982), Reason (1990) DBQ was developed to explore the role of human error in crashes and has been found to be a good predictor of crash involvement (Reason, et. al., 1990; Parker et. al., 1995; Stradling et. al., 2000) DBQ is formed by questions whose purpose is to test driver's propensity to commit any of the issues defined above. There was correlation between answers to the questionnaires and the crash history of users. (Parker et al, 1995, Rothengatter, 1997)

7.2.1.1 Examples of applications of the method

The method has been used in order to examine Road User Interactions: Patterns of Road Use and Perceptions of Driving Risk. The questionnaire was used in order to test New Zealand Drivers' Behaviour, according to Driving Violations and Risk Taking.

7.2.2 Driving Quality Scale (DQS)

This scale is a self-reporting measure of how well the subject thinks he/she 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" (Brookhuis et. al., 1999).

7.2.2.1 Examples of applications of the method

The DQS has been used in different projects where one example is given below.

Effects of MDMA (ecstasy), and multiple drugs use on (simulated) driving performance and traffic safety was assessed (Brookhuis et al., 2004). The aim of this project was to understand how much the use of MDMA, common dance drug, modified driving behaviour.

Self-reported effort increased on a scale from 0 to 150 from 40.2 in the non-drug condition to 47.6 in the MDMA condition to 50.7 in the multi-drug condition [drug versus non-drug is significant; $F(1,29)=5.02$, P=0.037]. The control group indicated an average of 49.1 [$F(1,31)=1.83$, NS]. On the driving quality scale, ranging from +100 (extremely well) to –100 (extremely badly), the control group on average rated normal driving (which was 0.2). The experimental group in the non-drug condition indicated driving as better than normal (+16.5), under the influence of MDMA about normal (+3.8), and under the influence of multiple drugs lower than normal (–4.6). The difference between non-drug state and the two drugs conditions was significant [F (1, 19) =5.13, P=0.035].

7.3 Expert-reported measures of driving Performance

These kinds of measures are gathered by filling in ready-made prospects, representative of important aspects of driving performance. Such prospects are filled by experts while users perform a task. Performance is achieved by judging user's behaviour according to normal levels of driving. Salient aspects of good performance are identified and during trials they are certified to be respected. We present two methods to assess Expert-reported measures: TRIP and Wiener Fahrprobe.

7.3.1 TRIP

TRIP (Test Ride for Investigating Practical Fitness to Drive) is a tool which identifies what aspects of driving is a concern to the driver. Different *sections* (i.e. different aspects of the driving behaviour by users; examples of them are presented below in

Table 7.6) are present and a scoring system determines the Practical Fitness to Drive. The *scoring modes* (i.e. the way a performance by the driver is ranked by experts) 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.

Table 7.6 Sections for TRIP

Lateral Position on the road	a) Average positioning (in scale [too left] 1 2 3 2 1 [too right]) b)Steadiness of steering (I D S G)
Lane Position Change (I D S G)	
Distance from car in front ([too short distance] 1 2 3 2 1 [too long distance])	a) Goodness of distance adaptation in town /off town areas (I D S G)
Speed	a) Driver classification according to style of speed choice ([too fast] 1 2 3 2 1 [too slow]) b) Goodness of speed adaptation in town /off town areas (I D S G)
Visual Behaviour and Communication (I D S G)	
Traffic Signals well perceived and responded to (I D S G)	
Mechanical Operations operating brakes, accelerator, steering (I D S G)	
Anticipation, with regard to changing road situations / changing traffic situations (I D S G)	
Understanding, perception and quality of traffic participation: perception and traffic inside (I D S G)	
Turning Left	a) When approaching crossing or junctions b) At the crossing or junction
Joining the traffic stream (I D S G)	
Conclusions	
Expert Judgement	

Each section is provided with a blank space, for in which to make annotations when the scoring is Insufficient. 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 description, see HASTE *Development of Experimental Protocol* (de Waard, Brookhuis et al, p 109). However, in the actual HASTE experiments Wiener Fahrprobe (see below) was selected instead of TRIP.

7.3.2 Wiener Fahrprobe

The method was originally designed by Risser (1985), 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. See

Table 7.7 below for an overview of the observation plan.

Table 7.7 Tasks for observers during Wiener Fahrprobe

Observer on the rear seat records	Observer on the front seat
Standardised Variables (such as speed adaptation at junctions/obstacles, lane change, interaction with other road users, driving above the limit etc), that is types of behaviour that is likely to appear and can be predicted. The Highway code can be used to identify erroneous behaviours.	Free observations about: • Conflicts (collision and evading manoeuvres), • Communication, • Interaction with pedestrians and • Special events

The technique is best used on a predefined experimental route, thereby minimising variance in road and traffic conditions between drivers (although of course this can never be entirely reduced). The technique requires a significant amount of preparation in advance, and when assessing IVIS/ADAS systems, it would be of use to perform two tests, one with the system under test enabled and one with the system disabled.

Analysis of collected data is rather simple, and requires a total count of the number of negative behaviours including:
• Unsafe merging/gap acceptance at junctions.
• Incorrect lane changes.
• Ignores other road users e.g. by not adapting their speed.
• Unsafe overtaking manoeuvres.
• Adoption of short headways.

In addition, the total number of conflicts can be calculated. Below (in Table 7.8) WF is presented as taken from HASTE and re-organized.

Table 7.8 Scheme for Wiener Fahrprobe, graphically modified HASTE layout.

Approaching a place of interaction	
	Checks the situation
	Drives with anticipation
	Does not drive with anticipation
	Inappropriate speed
	Inaccurate lane choice

Overtakes or Changes lane	
	Cuts up
	Too small lateral distance
	Aborted

Conflict: Subject…	
	provokes conflict
	Does not provoke conflict

Communication		Comments	
	Positive		Positive
	Negative		Negative

Interaction			
	Insists on right of way		Does not insist on right of way
	Does not allow to continue/merge		Allows to continue/merge
	Does not reduce speed		Reduces speed
	Presses other cars		Obstructs others when turning right
	Obstruction others (e.g. at crossing, etc.)		Obstructs others when turning left
	Others move into the safey distance of the subject		Makes other road users decelerate
	Turns right near oncoming traffic		Makes others accelerate
	Impedes cyclists / pedestrians		Endangers cyclists / pedestrians

Standardized Observation			
Overtaking or lane change		Speed	
	Correctly		Inappropriate for road geometry
	Not correctly		Too fast near VRUs
	In spite of oncoming traffic		In the platoon
	Without sufficient vision		Without platoon
	While forbidden		Above the speed limit
	Because of a stationary obstacle		At / below the limit
	Lane change in time		Considerably slower than the limit
	Uses right lane mainly		Brakes abruptly
	Uses left lane mainly		Unsteady speed

Use of the indicator		Distance to the road user ahead		
	Indicates in time		Correct		Too short
	Does not indicate	Behaviour when merging		
	Does not indicate in time		Safe		With Traffic
	Indicates ambiguously		Unsafe		Without Traffic

Lane use		Behaviour at traffic lights	
	Inaccurate, weaving		Drives against red
	Extremely on the right side of the lane		Drives against amber
	Extremely on the left side of the lane		Does not start when it is green
	Cuts the curve		Starts too early

Lane choice for proceeding			Checks situation with respect to other road users		
	Correct		At the last moment		Yes		No
	In time		Incorrect	Number of cars overtaking		

7.3.2.1 Examples of applications of the method

Below, two experiments are presented where the Wiener Fahrprobe has been used.

In an assessment of ISA (Intelligent Speed Adaptation), the so called *speed camera hypothecation pilot - the South Wales experience*, subjects were monitored with WF with and without the system support. A group of 20-25 drivers was studied to differences in workload when driving with and without ISA. All test drivers were interviewed four times; before their vehicle was equipped with the ISA, after driving with the ISA for one month, at the half time of the project and finally at the end of the project.

In Hjälmdahl (2002) presents results from in-vehicle observations of subjects who drove with active gas pedal (ISA). The Wiener Fahrprobe questionnaire was used here to assess the ISA system with regard to e.g. speed adaptation and car following distance.

8   Situation Awareness measures

Situation Awareness measures could be seen as a mix of many of the earlier described techniques and metrics. 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* (Endsley & Garland, 2000). 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 (Endsley & Garland, 2000):

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. This phase allows experts to gather the most important data about SA.

According to an object's position in space and their kinetic properties, people tend to project the object's position at a given time, in order to understand what the implications for them and their task will be in the future. Such aspects are strictly tied to Comprehension and Projection Phases, and are more accurate when people's expertise with environment is higher.

Several factors affect SA, such as working and long-term memory, other cognitive processes, mental models, goal precedence, expectations, and automaticity. (For a detailed description of such aspect, see Endsley et. al., 2000, pp12-22)

8.1 Situation Awareness, Workload, and Performance

Situation Awareness can be measured along with other measurements, such as Workload and Performance, when evaluating systems, to understand how the new system (or some components of it) modifies SA.

SA and workload sometimes are seen as related, e.g. high levels of workload can lead to low SA, or a high SA might lower workload. As stated by de Waard (1996), when complex systems take over the driver, and a "deactivation" phase takes place (Driver Vigilance), SA can decrease, with consequent increase of accident probability. On the other hand, systems may take care of secondary aspects of car driving, leaving the driver free to accomplish primary task safely, thus allowing him increase his SA. But Endsley (1993) showed during several experiments that only when workload exceeds people's capacity will SA be necessarily modified. In conclusion, it can be stated that workload and SA are quite independent in the most cases, having interrelations in certain cases only.

It is important to have workload and SA measured independently when assessing a system, since, as stated above, workload measurements not necessarily reflect SA ones. For instance, Endsley (2000) states that a system could be harder to use than a previous one, thus raising workload, but delivering the same level of SA; (this is an index of poor design, because the new interface, or introduced modification, did not bring any good increase for SA, but also rendered the job heavier, ndr);

conversely, a system with new features might deliver less SA while not varying workload (another index of poor design: no better results were achieved according to SA side, and no better results according to workload, ndr).

Moreover, performance is not a function of SA: having a higher SA does not mean the driver *will* perform better; several factors, such as slips or system failures might bring the user to poor performance even if he had the best available SA; on the other hand, poor SA does not mean the driver *will* have an accident, though it is likely to happen. The best statement is that the more complete SA, the higher the Probability to have high Performance levels.

SA can be assessed in terms of self-reported measures and in terms of performance measures.

8.2 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 a specific aspect. 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 that users sometimes 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 is 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 lead to poor Performance (Endsley et. al., 2000).

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 data, 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.

SART and SAGAT will be discussed below as examples of Self-Report Measures.

8.2.1 SART

Situation Awareness Rating Technique (SART) is one of the best known and widely used scales to assess SA (Endsley & Garland, 2000). Characteristics of understanding by the users are:
1) Understanding of situations in making decisions
2) Understanding is available to consciousness
3) Understanding can readily be made explicit and quantifiable.

In order not to take into account only the aspects outlined in the definition of SA, such definition was not used by the crew, who started from another point of view. The elements the crew considered suitable for testing SA were as described in

Table 8.1.

Table 8.1 SART salient aspects

	Domain	Dimension		Description
Rating on a -mm scale	Attentional Demand	Instability of Situation	Rating discretely 1 to 7	Likeliness of situation to change suddenly
		Variability of Situation		Number of variables which require one's attention
		Complexity of Situation		Degree of complication (number of closely connected parts) of situation
	Attentional Supply	Arousal		Degree to which one is ready for activity (sensory excitability)
		Spare Mental Capacity		Amount of mental ability available to apply to new variables
		Concentration		Degree to which one's thoughts are brought to bear on situation
		Division of Attention		Amount of division of attention in the situation
	Under-standing	Information Quantity		Amount of knowledge received and understood
		Information Quality		Degree of goodness or value of knowledge communicated
		Familiarity		Degree of acquaintance with situation experience

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.

Strong points of this method are:
- Measures directly derived from users (ecological validity);
- General Construct allows 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 workload constructs), it should provide some measure of how changes in workload affect SA

8.2.1.1 Issues and demand for further research

There are certain issues to consider in future research:
- 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 workload, as SA is different from workload
- It is not clear whether the 10-dimensions division into three domains is sufficient and necessary (Taylor et. al., 1991)

8.2.2 SAGAT

The Situation Awareness General Assessment Technique (SAGAT) is perhaps the most used technique to assess SA. The technique is described below.

During a simulation with a system being assessed simulation is frozen at randomly times, displays are blanked during SA assessment and 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.

This measure is highly subjective, because it probes in depth into subjects' own thoughts, and as such it has all potential and drawbacks of usual subjective methods. This method allows subjects to be queried during test, which provides coherent, in-time information, while the subject's memory is fresh and SA can be assessed trustfully. Because of stopping the tasks during SA testing, intrusiveness is reduced, and subjects will resume the task from the point where they stopped with full mental potential available. Widely spreading questions about all aspects of system help the testing to be more reliable, as it reduces the possibility of biases due to expectancy and the preparation for answering. The randomization of moments queries take place and this is another way to reduce biases.

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 (Endsley et. al., 2000, p148):
- Goal
    - Sub Goal
    - Decision
    - Projection (Level 3 SA)
    - Comprehension (Level 2 SA)
    - Data (Level 1 SA)

After a Goal has been chosen to be tested, all sub goals to achieve the parent are enumerated. For each sub goal, there is a list of decisions which has to be accomplished for the identified sub goal and for each decision the three levels of SA information required is identified. Such information is the source of the queries. When a subject answers a specific query, it can be checked whether the subject's level of SA is appropriate for goal completion. The definition of queries can take years to complete, and are therefore quite expensive, but are quite general and can be applied to a wide variety of situations and people.

It is important that subjects are familiar with the method, in order to make answers as accurate as possible; to this extent, pre-test phases should be acted repeatedly (3 to 5 times), until subjects feel accustomed with the technique. During the test phase, collected data will be more spontaneous and precise. During SAGAT gathering, other measurements can take place, as SAGAT was not proven to be invasive or contaminating other results. Usually, in the first 3 to 5 minutes of testing no freezes take place, and freezes have a distance in time of at least 1 minute. This allows subjects to create their mental image, and ensures unpredictability of freezes. According to complexity and variability of tasks, more freezes can be done. In a within-subjects test, 30 to 60 samples for query are advised as a mean value.

Experimental results in aviation led to conclude that SAGAT has good sensitivity and reliability and intrusiveness as freezing never influenced performance (Endsley et al., 2000).

For safety reasons, SAGAT may not be applicable to all domains, especially real on-road applications. Two subjects should be used, in order to maintain safety: while the first is tested, the second takes control of the 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 workload levels, could be used 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 is followed (Endsley et. al., 2000).

Strong points of this method are:
- 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 with SAGAT.

8.2.2.1 Examples for the application of the method

One project who intended to use the assessment of situation awareness was HASTE. However, the methods were not singled out in the experimental phases as they were for example said to be too time consuming. HASTE however presented a sample of questions that could be used in automotive context. (Roskam et al., 2000), dividing them into the three levels of SA: Perception, Comprehension, and Projection. Some samples are presented in Table 8.2 where the questions could be presented e.g. during a simulation when a system was being assessed.

Table 8.2 Levels of SA and questions

Perception	1. Did you receive a message in the last 30 seconds?
	2. What colour was the light at the last intersection?
	3. How many pedestrians at the last intersection?
Comprehension	1. What action did the most recent message call for?
	2. What is the speed of the car ahead relative to you?
	3. What was the last warning received by the system?
Prediction of future events	1. Should you be driving less than 50kmph in the next 30 seconds?
	2. Should you need to be in the left lane in the next 30 seconds?

8.3 Performance measures

Performance Measures, opposite to Self-reported measures, rely only on subject's interaction with the system. These consist of any measurement that infers subjects' SA from their observable actions or the effects these actions ultimately have on system performance (Endsley & Garland, 2000). This allows experimenters to compare the desired and achieved performance of a System, and its weak points, which are 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 identifying:
- 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. (Endsley et. al., 2000)

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. Embedded 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, as long as they assess different aspects of SA, as they are not conflicting one to each other. Performance measures are often taken along with verbal reporting, because they never showed intrusive for task accomplishment or for Performance measurements. However, when assessing topics concerning knowledge-based aspects of a system, by asking subjects to name pieces of information, it may change the subject's strategies for situation assessment and decision making to include that information more than normal (Endsley, 1995; Vidulich, 1995). If any of the measures in the experiment are intrusive, the changes in subject behaviour may affect the other measures.

When measuring SA, a set of scenarios, aiming to test precise aspects, is developed. It is interesting to create abnormal behaviour situations among actors other than the subjects under test, in order to calculate how much resources are to be implied to take care of the new, unforeseen situation. It can be a car breaking abruptly, to test how much time does the subject take to brake, a pedestrian crossing where not allowed, to see how much time does the subject to adjust speed in order not to catch him, and so forth. Such scenarios, which render dangerous situations, are likely to be used in simulators for safety reasons, or certainly not in real traffic situations.

9   Summary and conclusions

This review has presented some of the most common methods, tools and metrics used to assess In Vehicle Information Systems and Advanced Driver Assistant Systems. This document has been structured according to the actual techniques and methods and not on concepts such as for example physical and mental workload. The main chapter has been on:

1) Driving performance measures
2) Visual performance measures
3) The Occlusion technique
4) Physiological measures
5) Secondary task methods
6) Subjective assessment methods
7) Situation Awareness

Since each chapter provide the reader with further issues to consider in future development the intention in AIDE is that the task members in WP 2.2 will look into these matters as well as explore other underlying factors further.

In the deliverable most of the metrics explored are used to capture the mostly negative effect IVIS have on the driver. In the future work within AIDE it is important to distinguish between methods, tools and metrics suitable for ADAS vs. IVIS. However, since the work in SP2 has a focus on IVIS it is expected that SP1 can provide the project with more of the effects related to ADAS. In the work to refine the methods, tools and metrics described in this deliverable it is important to also look at what happens in more dynamic environments (e.g. interaction with IVIS in a complex driving environment) where e.g. the time sharing behaviour between primary and secondary tasks might be different and effect the measurement of workload and distraction.

Finally it will be important in the further development of the methods, tools and metrics to identify the lowest number of methods, tools and metrics in order to have a test battery valid but still cost-efficient enough to capture the complex task of driving and the different expected effects.

10 References

Alm, H., Nilsson, L. (1995). The effects of a mobile telephone task on driver behaviour in a car following situation, Accident Analysis and Prevention, Vol. 27 (5), pp. 707-715.

Antin, J.F., Dingus, T. A., Hulse, M.C. and Wierwille, W.W. (1990). An evaluation of the effectiveness of an automobile moving-map navigation display. International Journal of Man-Machine Studies. 33, 581-594.

Applied Science Laboratories's web page: http://www.a-s-l.com/. Accessed on the 1st of September, 2004.

Bauer, A. (BASt), Tango F. (CRF) ADVISORS: An Evaluation of Lateral Support System (LSS),

Becker, S., Brandenburg, K., Feldges, J., Fowkes, M., Johanning, T., Kopf, M. (2000). Deliverable D2.1 – System, user and legal aspects: The integrated approach for the assessment of Driver Assistance Systems. Project TR4022.

Bengler, K., Praxenthaler, M., Theofanou, D., Eckstein, L. (In press). Investigation of Visual Demand in Different Driving Simulators within the ADAM Project. Driving Simulation Conference Europe 2004. Paris.

Bhise, V. D., Forbes, L.M., Farber, E.I. (1986). Driver Behavioral Data and Considerations in Evaluating In-vehicle Controls and Displays. Presented at the Annual Meeting of the TRB, Washington, D. C., January 1986.

Boer, E. (2000). Behavioural Entropy as an Index of Workload. Proceedings of the IEA 2000/HFES 2000 Congress.

Breuer, J.; Bengler, K.; Heinrich, Ch.; Reichelt, W. (2002). Development of advanced driver attention metrics (ADAM). Proceedings of the GfA-Conference, Munich.

Brook-Carter, N.; Parkes, A.; Burns, P. and Kersloot T. (2002). An experimental assessment of an urban adaptive cruise control (ACC) system. TRL Annual Research Review.

Brookhuis, K., de Ward, D. and Mulder, B. (1994). Measuring performance by car-following in traffic. Ergonomics, 1994, Vo. 37, No 3, 427-434.

Brookhuis, K., Uneken, E. & Nilsoon, L. (1999): ADVISORS common measures. Document RUG ID5_1_1.DOC. ADVISORS

Brookhuis, Karel A., de Waard, D. Samyn, N. (2004) The Effects of MDMA (ecstasy) and multiple drugs use on (simulated) driving performance and traffic Safety), http://www.maps.org/w3pb/new/2004/2004_brookhuis_6325_1.pdf

Budescu, D.V., Zwick, R., & Rapoport, A (1986). A Comparison of the Eigenvalue Method and the Geometric Means procedure of Ratio Scaling. Applied Psychological Measurement, 10, 69-78.

Burns, P.C. (2001). Behavioural Adaptation to an Advanced Driver Support System. Volvo Technology Corporation, Internal Report.

Burns, P. C.; Knabe, E. and Tevell, M. (2000). Driver behavioral adaptation to collision warning and avoidance information. Paper presented at the IEA/HFES, International Ergonomics Association, San Diego.

Byers, J. C., Bittner, A.C., Hill, S.G. (1989). Traditional and raw task load index (TLX) correlations: are paired comparisons necessary? In A. Mital (Ed.) Advances in industrial ergonomics and safety, I. London: Taylor & Francis.

Carsten, O. and Comte, S. (1997). UK work on automatic speed control. Proceedings of the ICTCT 97 conference, 5-7 November, Lund, Sweden.

Charlton, S. (1996). Mental Workload Test and Evaluation. Handbook of Human Factors Testing and Evaluation, 181-199.

Chapman, P., Underwood, G. (1998). Visual search of Dynamic Scenes: Event Types and the Role of Experience in Viewing Driving Situations. Chapter 17 in Eye guidance in reading, driving and scene perception. Ed. Underwood, G. Elsevier, 98/22009 0080433618.

Chin, E., Nathan F. (2002). Roadsense D 2.1 part 1 – state of the art on HMI metrics and target values.

COMUNICAR *from Deliverable*: Human Factor Tests on car Demonstrator – The Methodology

Cooper, G.E., Harper, R.P. Jr. (1969). The use of pilot ratings in the evaluation of aircraft handling qualities. NASA TN-D-5153.

Curry, G.A., Hieatt, D.J. and Wilde, G.J.S. (1975). Task load in the motor vehicle operator: A comparative study of assessment procedures. Ottawa, Ontario: Ministry of Transport, Road and Motor Vehicle Traffic Safety Branch.

de Waard, D. (1996). The Measurement of Drivers' Mental Workload. ISBN 90-6807-308-7. Traffic Research Centre. University of Groningen.

de Waard, D., Brookhuis, K. et al; (2000). HASTE: Human Machine Interface And the Safety of Traffic in Europe – Development of Experimental Protocol

Deering, R. K (2002). Annual Report of the Crash Avoidance Metrics Partnership,
April 2001 - March 2002 NHTSA DOT HS 809 531

Delphi (2004). Delphi public proposal release: SAfety VEhicle(s) using adaptive Interface Technology (SAVE-IT) Program DTRS57-02-R-20003. U.S. Department of Transportation. www.volpe.dot.gov/opsad/saveit/docs.html. Accessed on the 1st of September, 2004.

Dingus, T., Antin, J, Hulse, M, and Wierville, W. (1989). Attentional demand requirements of a moving-map navigation system. In Transportation Research, A 23(4), 301-315.

Dingus, T. A. (1995). Moving From Measures of Performance to Measures of Effectiveness in the Safety Evaluation of ITS products or Demonstrations. University of Iowa. Safety Evaluation Workshop.

Donges, E. (1978). A Two-Level Model of Driver Steering Behaviour. Human Factors, 20(6), 691-707

DOD, Department of Defense (1999). Department of Defense handbook, MIL-HDBK-46855A: Human engineering program process and procedures. Washington DC: DOD

Endsley, M.R., Garland, D.J. (2000) Situation Awareness Analysis and Measurement, Lawrence Erlbaum Associates, Publishers, London

Endsley, M.R. (1993) Situation awareness and workload: Flip sides of the same coin. In R.S.Jensen & D.Neumeister (Eds.), Proceedings of the Seventh International Symposium on Aviation Psychology (Vol. 2, pp.906-911). Columbus: Department of Aviation, The Ohio State University

Fairclough, S.H. (1991). Adapting the TLX to measure driver mental workload (Report V1017/BERTIE/No. 71). Loughborough, Leics,UK: HUSAT Research Institute.

Farber, E., Blanco, M. Foley, R., Curry, J., Greenberg, J., and Serafin, C. (2000). Surrogate Measures of Driver Visual Demand While Driving. Proceedings of the Human Factors Engineering Society, Vol. 44, Santa Monica, California

Finch, D.J., Kompfner, P., Lockwood, C.R., Maycock, G. (1994). Speed, speed limiters and accidents. Project Report 58. Transport Research Library, Crowthorne, UK.

Fosser, S., Saetermo, I. F. and Sagberg, F. (1997). An investigation of behavioural effects to airbags and antilock brakes among taxi drivers. Accident Analysis and Prevention, 29 (3).

Gelau, C. & Krems, J.F. (2004). The occlusion technique: a procedure to assess the HMI of in-vehicle information and communication systems. Applied Ergonomics, 35, 185-187.

Gelau, C. (2004). Fahrerablenkung durch Informations- und Kommunikationssysteme im Fahrzeug: Auswirkungen auf das Fahrerverhalten und die Verkehrssicherheit (S.297-316). In B. Schlag (Hrsg.), Verkehrspsychologie. Mobilität – Verkehrssicherheit – Fahrerassistenz. Pabst Science Publishers.

Gelau, C., Keinath, A., Baumann, M., Bengler, K., Krems, J.F. (1999). Die Okklusionsmethode als Verfahren zur Bewertung von visuellen Displaydarstellungen in Kraftfahrzeugen. Zeitschrift für Arbeitswissenschaft, 25(1), 51-57.

Gibson, J.J. (1979). The Ecological Approach to Visual Perception. Boston: Houghton Mifflin.

Gibson, J.J. and Crooks, L. (1938). A theoretical field-analysis of automobile driving. American Journal of Psychlogy, 51, 453-471.

Godthelp, J, and Konings, H. (1981). Levels of steering control; some notes on the time-to-line crossing concept as related to driver strategy. In Proceedings of the First European Annual Conference on Human Decision and Manual Control (pp. 343-357). Delft, The Netherlands: Technical University.

Godthelp, H., Milgram, P. and Blaauw, J. (1984). The Development of a Time-related Measure to Describe Driving Strategy. Human Factors, 26(3), 257-268.

Green, P. (1993, revised March 1994, August 1994, final version in 1995). Measures and Methods Used to Assess the Safety and Usability of Driver Information Systems. Technical Report UMTRI-93-12. FHWA-RD-94-088

Green, P. (1999). Estimating compliance with the 15-second rule for driver interface usability and safety. Proceedings of Human Factors and Ergonomics Society 43rd Annual Meeting.

Green, M. (2000). "How long does it take to stop?" Methodological analysis of driver perception-brake times. Transportation Human Factors, 2(3), 195-216.

Greenberg, J., Tijerina, L., Curry, R., Artz, B., Cathey, L., Grant, P., Kochlar, D., Kozak, K., and Blommer, M. (2003). Evaluation of driver distraction using an event detection paradigm. Journal of the Transportation Research Board, No 1843. Washington D.C.: TRB.

Greenshields, B.D. (1963). Driver behaviour and related problems. Highway research record. 25, 14-32

Gronwall, D. (1977). Paced auditory serial-addition task: A measure of recovery from concussion. Perceptual and Motor Skills, 44, 363-373.

Gunn, R., Mason, G., Dangerfield, K, (2001) The In-Depth Interview, http://www.pra.ca/resources/indepth.pdf

Harbluk, J.L., Noy, Y.I. (2002), The Impact of Cognitive Distraction on Driver Behaviour and Vehicle Control, Ergonomics Division, Road Safety Directorate and Motor Vehicle Regulation Directorate, Canada

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

Hart, S. G., Battiste, V., and Lester, P. T. (1984). POPCORN: A supervisory control simulation for workload and performance research (NASA-CP-2341). In Proceedings of the 20th Annual Conference on Manual Control (pp. 431-453). Washington, DC: NASA.

Haworth, L. A., Bivens, C. C., and Shively, R. J. (1986). An investigation of single-piloted advanced cockpit and control configuration for nap-of-the-earth helicopter mission tasks. In Proceedings of the 42nd Annual Forum of the American Helicopter Society, 657-671.

Heijer, T. , Oei, H. L, Wiethoff, M., Penttinen, M., Shirokoff, A., Kulmala, R., Heinrich, J., Ernst, A. C., Sneek, N., Heeren, H., Stevens, A., Bekiaris, A., Damiani, S. (2003). Problem identification and Actor Classification - D1/2.1 V12. Competitive and Sustainable Growth Programme ADVISORS GRD 1 2000 10047

Hill, S.G., Iavecchia, H.P., Byers, J.C., Bittner, A.C., Zaklad,A.L. and Christ, R.E.(1992). Comparison of four subjective workload rating scales. Human Factors,34(4), 429-439.

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].

Hjälmdahl, M., and Varhelyi, A. (2004). Speed Regulation by in-car active accelerator pedal – effects on driver behaviour. Transportation Research Part F, Vol. 7(2), pp. 77-94.

Hoedemaeker, M., Dangelmaier, M., Gelau, C., Mattes, S., Montanari, R. (2003). Deliverable 6.5: Test results evaluation. Information Society Technologies, Comunicar PROGRAMME "IST KA1".

Hoedemaeker, M., Dangelmaier, Gelau, C., Mattes, S., and Montanari, R. (2004). Test Results Evaluation. COMUNICAR Deliverable 6.5. IST 11595.

HSIAC website (2004). http://iac.dtic.mil/hsiac/docs/TLX-UserManual.pdf. Accessed 1st of September.

HUMANIST web site: www.noehumanist.org. Accessed on the 1st of September, 2004.

Höger, R. (2001). The signal location task as a method quantifying the distribution of attention. In D. Harris (Ed.), Engineering Psychology and Cognitive Ergonomics, Volume five (pp. 373-380). Aldershot: Ashgate.

IN-ARTE web page: http://www.cordis.lu/telematics/tap_transport/research/projects/in-arte.html. Accessed on the 1st of September, 2004.

ISO 17287:2003. Road vehicles – Ergonomic aspects of transport information and control systems - Procedure for assessing suitability for use while driving. International standard, International Organization for Standardization.

ISO 15007-1 Road vehicles – Measurement of driver visual behaviour with respect to transport information and control systems Part 1: Definitions and parameters.

ISO 15007-2 Road vehicles – Measurement of driver visual behaviour with respect to transport information and control systems Part 2: Equipment and procedures.

ISO. (2004). Road Vehicles – Ergonomic Aspects of Transport Information and Control Systems – Simulated Lane Change Test to Assess Driver Distraction: Preliminary Work Item, First Draft. ISO/TC22/SC 13/WG8 N416.

Jahn, G.; Oehme, A.; Rösler, D.; Krems, J. (2003). IHRA-ITS-Swedish-German Study on Method Development. Federal Highway Research Institute (BASt) Project Nr. FE 82.175/2000. Chemnitz, Germany: Chemnitz University of Technology.

Janssen, W. H., Kuiken, M.J., Verwey, W.B. (1994). Evaluation studies of a prototype intelligent vehicle. In ERTICO (Ed.) Towards an intelligent transport system, Proceedings of the first world congress on applications of transport telematics and intelligent vehicle-highway systems. Boston: Artech House.

Jex, H. R., & McDonnell, J. D. (1966). A "Critical" Tracking Task for Manual Control Research. Transactions on human factors in electronics, 7(4), 138-145.

Keinath, A., Baumann, M., Gelau, C., Bengler, K. & Krems, J.F. (2001). Occlusion as a technique for evaluating in-car displays. In Harris, D. (Ed.), Engineering Psychology and Cognitive Ergonomics, Vol. 5 (pp. 391-397). Aldershot, U.K.: Ashgate Publishing Ltd.

Kelley, C.R. (1969). The measurement of tracking proficiency. Human Factors, 11(1), 43-64.

Krems, J. F.; Keinath, A.; Baumann, M.; Jahn, G.; Bengler, K. (2004). Die Okklusionsmethode: Ein einfaches und valides Verfahren zur Bewertung der visuellen Beanspruchung von Zweitaufgaben? In: Bernhard Schlag (Hrsg.): Verkehrspsychologie. Mobilität – Sicherheit – Fahrerassistenz. Lengerich: Pabst Science Publishers.

Lamble, D. Laakso, M. Summala, H. (1998). Detection thresholds in car following situations and peripheral vision: Implications for positioning for visually demanding in-car displays. Ergonomics, 41.

Larsson, P. (2002). Automatic Visual Behavior Analysis. Dissertation for a Master of Science Degree Applied Physics and Electrical Engineering Control and Communication Department of electrical engineering Linköping University, Sweden. LiTH-ISY-EX-3259.

LC Technologies, Inc web site: http://www.lctinc.com. Accessed on the 1st of September, 2004.

Lee, D.N. (1976). A theory of visual control of braking based on information about time-to-collision. Perception, 5, 437-459

Lee, J.D., Caven, B., Haake, S. and Brown, T.L. (2001). Proceedings of the Human Factors and Ergonomics Society, Vol. 43, Santa Monica, California.

Lee, J. D., Ries, M. L., McGehee, D. E., Brown, T. L., Perel, M., (2000) Proc Internet Forum on The Safety Impact of Driver Distraction When Using In-Vehicle Technologies, (NHTSA, DOT, Washington DC), http://www-nrd.nhtsa.dot.gov/departments/nrd-13/driver-distraction/Papers.htm (July 5- August 11, 2000).

Liu, Schreiner and Dingus. (1999). Development of human factors guidelines for Advacnced Traveler Information Systems (ATIS) and Commercial Vehicle Operation (CVO): Human Factors Evaluation of the Effectiveness of Multimodality Displays in ATIS. NHTSA FHWA-RD-96-150

Lysaght, R. J., Hill, S. G., Dick, A. O., Plamondon, B. D., Linton, P. M., Wierwille, W.W., Zaklad, A. L., Bittner, A. C., Wherry, R. J. (1989). Operator workload: Comprehensive review and evaluation of operator workload methodologies (Tech. Rep. 851). Alexandria, VA: U.S. Army Research Institute.

Malaterre, G. (1994). Méthode de mesure de la charge de travail en situation de conduite simulée et réelle. Rapport INRETS n 191.

Martens, M. H.; van Winsum, W. (2000). Measuring distraction: The Peripheral Detection Task. Soesterberg, Netherlands: TNO Human Factors.

Mattes, S. (2003). The Lane Change Task as a Tool for driver Distraction Evaluation. In H. Strasser, H. Rausch & H. Bubb (Eds.), Quality of Work and Products in Enterprises of the Future. Stuttgart: Ergonomia Verlag.

McDonald, W.A., and Hoffman, E.R. (1980). Review of relationships between steering wheel reversal rate and driving task demand. Human Factors 22(6), 733-739.

McKnight, A. & McKnight, A. 1993. The effect of cellular phone use upon driver attention. Accident Analysis and Prevention, 25, 259-265.

McLean, J.R. & Hoffman, E.R. (1971). Analysis of Drivers' Control Movements. Human Factors, 1971, 13(5), 407-418.

McLean, J.R. & Hoffman, E.R. (1972). The effect of lane width on driver steering control and performance. Proceedings of the Australian Road Research Board Sixth Conference, 6(3), 418-440.

McLean, J.R. & Hoffman, E.R. (1975). Steering wheel reversals as a measure of driver performance and steering task difficulty. Human Factors, 17, 248-256.

Merat, N. (2003). Loading Drivers to Their Limit: The Effect of Increasing Secondary Task on Driving. Proceedings of the Second International Driving Symposium on Human Factors in Driver Asssessment , Training and Vehicle Design, Park City, Utah.

Michon, J.A. (1985). A critical review of driver behaviour models: What do we know? what should we do? In L.A Evans and R.C. Schwing (Eds.) Human Behaviour AND Traffic Safety. (pp. 487-525). New York: Plenum Press.

Michon, J. A. (Ed.) (1993). Generic Intelligent Driver Support: A Comprehensive Report on GIDS. London: Taylor and Francis.

Milgram, P., & van der Horst (1986). Alternating-field stereoscopic displays using light-scattering liquid crystal spectacles. Displays: Technology & Applications, 7, 67-72.

Miura, T. (1986). Coping with situational demands: A study of eye movements and peripherial vision performance. In A. G. Gale & I. D. Brown & C. M. Haselgrave & P. Smith & S. H. Taylor (Eds.), Vision in Vehicles-II. Amsterdam: Elsevier.

Miura, T. (1990). Active function of eye movement and useful field of view in a realistic setting. In R. Groner & G. d'Ydewalle & R. Parham (Eds.), From eye to mind: Information acquisition in perception, search and reading (pp. 119-127). Amsterdam: Elsevier.

Muckler, F.A. & Seven, S.A. (1992). Selecting performance measures: 'objective' versus 'subjective' measurement. Human Factors, 34, 441-455.

Nakayama, O., Futami, T., Nakamura, T. and Boer, E.R. (1999). Development of a Steering Entropy Method for Evaluating Driver Workload. SAE Technical Paper Series: #1999-01-0892

Nathan, F., Ojeda, L. (2004). Synthesis of Deliverable 2.1 part 1 – State of the Art on HMI metrics and target values. RoadSense. Internal document sent to AIDE project.

Nilsson, L. (1995). Safety Effects of Adaptive Cruise Controls in Critical Traffic Sitauations. Proceedings of Steps Forward, Volume III, the Second World Congress on Intelligtent Transportation Systems, Yokohama, Japan.

Nilsson, L., Törnros, J., Parkes, A., Brook-Carter, N., Dangelmeier, M., Brookhuis, K,m Roskam, A-J., de Ward, D., Bauer, A., Gelau, C., Tango, F., Damiani, S., Jaspers, I., Ernst, A. and Wiethoff, M. (2000). An Integrated Methodology and Pilot Evaluation Results. ADVISORS deliverable 5.2, Part III.

Nilsson, G. (2004). Traffic safety dimensions and the power model to describe the effect of speed on safety. Bulletin 221, Lund University, Lund, Sweden

Norman, D.A. (1981). Categorization of action slips. Psychological Review, 88, 1-15.

Noy, Y.I., Lemoine, T.L., Klachan, C. & Burns, P. (2004). Task interruptability and duration as measures of visual distraction. Applied Ergonomics, 35, 207-213.

Nygård, M. (1999). A method for analysing traffic safety with help of speed profiles. MSc Thesis, Tampere University of Technology: Department of Civil Engineering.

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.

Olsson, S., Burns, P. C. (2000). Measuring driver visual distraction with a Peripheral Detection Task. Linköping, Sweden: Linköping University.

Parker, D., Reason, J., Manstead A.S.R., Stradling, S.G., (1995) Driving Errors, Driving violations and accidents involvement. Ergonomics, 38, 1036-1048

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

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

PSA Peugeot Citroen Subjective Assessment Methods; PSA – Proposition of a new Method (slides).

Rasmussen, J. (1982). Human errors: A taxonomy for describing human malfunction in industrial installations. Journal of Occupational Accidents, 4, 311-333.

Rasmussen, J. (1986). Information Processing and human-machine interaction: An approach to cognitive engineering. New York: North Holland

Reason, J. (1990) Human Error, Cambridge University Press, Cambridge

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.

Recarte, M. A., Nunes, L. M. (2000) Effects of verbal and spatial-imagery tasks on eye fixatons while driving. Journal of Experimental Psychology 2000, Vol 6 No.1.

Recarte, M. A., Nunes, L. M. (2003) Mental Workload While Driving: Effects on Visual Search, Discrimination, and Decision Making. Journal of Experimental Psychology 2003, Vol. 9, No. 2.

Risser, R. (1985) Behaviour in traffic conflict situations. Accident Analysis and Prevention Vol. 17, No. 2, pp 179-197.

RoadSense web page: http://www.eu-projects.com/roadsense/index.html. Accessed on the 1st of September, 2004.

Rockwell, T. H. (1972). Eye movement analysis of visual information acquisition in driving: An Overview. Proceedings of the 6th Biennial Conference of the Australian Road Research Board. August 1972, Canberra.

Rockwell, T. H. (1988). Spare Visual Capacity in Driving – Revisited: New Empirical Results for an Old Idea. In A. G. Gale, et al (Eds), Vision in Vehicles II, North Holland Press: Amsterdam.

Roskam, A.J., Brookhuis, K.A., de Waard, D. Carsten, O.M.J., Read, L., Jamson, S., Östlund, J., Bolling, A., Nilsson, L., Antilla, V., Hoedemaeker, M., Janssen, W.H., Harbluk, J., Johansson, E., Tevell, M. Fowkes, M. Victor, T., Engström. J. (2002). Deliverable 1 - Development of Experimental Protocol. Human Machine Interface And the Safety of Traffic in Europe. Project GRD1/2000/25361 S12.319626

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.

Rumar, K. (1988). In-vehicle information systems. International Journal of Vehicle Design, 9.

Saad, F. and Villame, T. (1996). Assessing new driving support systems: Contribution of an analysis of driver's activity in real situations. In Proceedings of the third annual world congress Intelligent Transportation Systems.

Saad, F., Hjälmdahl, M., Canas, J., Alonso, M., Garayo, P., Macchi, F., Nathan, F., Ojeda, L., Papakostopoulos, M., Panou, M. and Bekiaris, E. (2004). Literature Review of Behavioural Effects. Deliverable 1.2.1, AIDE Integrated Project, Sub-project 1. IST-1-507674-IP

SAVE-IT web page: http://www.volpe.dot.gov/opsad/saveit/docs.html. Accessed on the 1st of September, 2004.

Seeing Machine's web page: www.seeingmachines.com. Accessed on the 1st of September, 2004.

Senders, J. W., Kristofferson, A. B., Levison, W. H., Dietrich, D. W., Ward, J. L., 1967. The attentional demand of automative driving. Highway Res. Rec. 195, 15-33.

Serafin, C. (1993). Preliminary Examination of Driver Eye Fixations on Rural Road: Insight into Safe Driving Behavior (Technical Report UMTRI-93-29), Ann Arbor, MI: The University of Michigan Transportation Research Institute.

Schindhelm, R.; Gelau, C.; Montanari, R.; Moreale, D.; Deregibus, E.; Hoedemaeker, M.; De Ridder, S. & Piamonte, P. (2003). Human factor tests on car demonstrator. EU project COMUNICAR, project IST 11595, Deliverable 6.4.

Schindhelm, R., Gelau, C., Keinath, A., Bengler, K., Kussman, H., Kompfner, P, Cacciabue, P.C., Martinetto, M. (2004). Deliverable 4.3.1 Report on the review of the available guidelines and standards. Adaptive Integrated Driver-vehicle Interface. Project IST-1-507674-IP.

Shulman, M., Deering, R. K. (2004). Second Annual Report of the Crash Avoidance Metrics Partnership, April 2002 - March 2003 NHTSA DOT HS 809 663

Simons, D. J. (2000). Change Blindness and Visual Memory. Psychology Press: East Sussex.

Smart Eye's web page: www.smarteye.se. Accessed on the 1st of September, 2004.

Smyth, M. & Scholey, K. (1994). Interference in immediate spatial memory. Memory and Cognition, 22, 1-13.

Stevens, A., Board, P. A., Quimby, A. (1999). A Safety Checklist for the Assessment of in-Vehicle Information Systems: Scoring Proforma (Project Report PA3536-A/99), Crowthorne, UK: Transport Research Laboratory.

Stevens, A., Bygrave, S., Brook-Carter, N. & Luke, T. (2004). Occlusion as a technique for measuring In-Vehicle-Information-System visual distraction: a reserach literature review. TRL report, in press.

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.

Strayer, D. L., Drews, F. A., Johnston, W. A. (In Press). Cell Phone Induced Failures of Visual Attention During Simulated Driving. Journal of Experimental Psychology: Applied.

Summala, H. (1981).Driver/vehicle steering response latencies. Human Factors, 23, 6, 683-692.

Summala, H., Lamble, D., & Laakso, M. (1998). Driving experience and perception of the lead car's braking when looking at in-car targets. Accident Analysis and Prevention, 30, 401-407.

Taylor, R.M. (1990) Situational awareness rating technique (SART): The development of a tool for aircrew systems design. In Situational Awareness in Aerospace Operations (AGARD-CP-478, pp 3/1-3/17). Neuilly Sur Seine, France: NATO-AGARD)

Van der Horst, R. and Godthelp, H. (1989). Measuring road user behaviour with an instrumented car and an outside-the-vehicle video observation technique. Transportation Research Record # 1213, Washington DC: Transportation Research Board, 72-81.

Van der Horst, R. (2004). Occlusion as a measure of visual workload: an overview of TNO occlusion research in car driving. Applied Ergonomics, 35, 189-196.

van Winsum, W. and Godthelp, H. (1996). Speed choice and steering behaviour in curve driving. Human Factors, 38(3), 434-441.

van Winsum, W. & Hoedemaker, M. (2000). A road test of a protoype satellite system for in-vehicle menu control. TNO-report, TM-00-C003. Soesterberg, The Netherlands.

Vaughan, G., May, A., Ross, T., Fenton, P. (1994). A human-factors investigation of an RDS-TMC system. In ERTICO (Ed.) Towards an intelligent transport system, Proceedings of the first world congress on applications of transport telematics and intelligent vehicle-highway systems. Boston: Artech House.

Veltman, J.A., Gaillard, A.W.K. (1993). Measurement of pilot workload with subjective and physiological techniques. Paper presented at the annual meeting of the Europe Chapter of the Human Factors and Ergonomics Society, November 1993, Soesterberg, The Netherlands.

Verwey, W.B. (1991). Towards Guidelines for in-car Information Management: Driver Workload in Specific Situations. Technical Report IZF 1991 C-13, Soesterberg, The Netherlands: TNO Institute of Perception.

Verwey, W.B., Veltman, J.A. (1995). Measuring workload peaks while driving. A comparison of nine common workload assessment techniques (Report TNO-TM 1995 B-4). Soesterberg, The Netherlands: TNO Human Factors Research Institute.

Verwey, W. B. (2000). On-line driver workload estimation. Effects of road situation and age on secondary task measures. Ergonomics, 43(2), 187-209.

Victor, T, Blomberg, O., Zelinsky, A. (2001). Automating Driver Visual Behaviour Measurement, 9th Vision in Vehicles Conference, Brisbane, Australia.

Victor, T., Johansson, E. (in press). Driving information loss while performing secondary tasks.

Vidulich, Ward and Schueren (1991), Using the Subjective Workload Dominance (SWORD) technique for projective workload assessment in Human Factors, 33(6), 677-692

Vidulich, M.A. & Tsang, P.S. (1986). Techniques of subjective workload assessment: a comparison of SWAT and the NASA bipolar methods. Ergonomics, 29, 1385-1398.

Vidulich, M.A. & Tsang, P.S. (1987)Absolute magnitude estimation and relative judgement approaches to subjective workload assessment. In Proceedings of the Human Factors Society 31st Annual Meeting (vol. 2, pp1057-1061). Santa Monica, CA,: Human Factors Society.

Weir, D. H. and McRuer, T.M. (1968). A theory of driver steering control of motor vehicles. Highway Research Record, 247, 7-39.

Weir, D., Chiang, D.P. & Brooks, A.M. (2003). A Study of the effect of Varying Visual Occlusion on Driver Behavior and Performance While Using a Secondary-Task-Human-Machine-Interface.. SAE 2003-01-0128, Warrendale, PA.

Wickens, C.D. (1992). Engineering Psychology and Human Performance (Second Ed). NY: HarperCollins.

Wickens, C.D. and Gopher, D. (1977). Control theory measures of tracking as indices of attention allocation strategies. Human Factors, 1977, 19, 349-365.

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.

Wierwille, W., Tijerina, L., Kiger, S., Rockwell, T., Lauber, E. And Bittner, A Jr. (1996). Heavy Vehicle Driver Workload Assessment. Task 4: Review of Workload and Related Research. US Department of Transportation, NHTSA. DOT HS 808 467 (4).

Wiethoff, M. (2003). ADVISORS Final Report Annexes. EU project ADVISORS, project no. GRD1/2000/10047.

Williams, C., and Crawford, G. (1980, May). Analysis of Subjective Judgment Matrices (Tech. Report R-2572-AF). Santa Monica, CA: RAND

Witt, G. J., Zhang, H., Smith, M. R. H. (2004). SAfety VEhicle(s) using adaptive Interface Technology (SAVE-IT): Phase 1 Progress. National Highway Traffic Safety Administration International Workshop on Progress and Future Directions of Adaptive Driver Assistance Research May 13-14, 2004. U.S. DOT Headquarters. www.volpe.dot.gov/opsad/saveit/docs/may04/zhang.pdf. Accessed on the 1st of September, 2004.

Zang, H. (2003). Task 7: Visual Distraction. Phase I SAVE-IT Briefing, August 12, 2003, Washington DC. www.volpe.dot.gov/opsad/saveit/docs.html. Accessed on the 1st of September, 2004

Zijstra & Van Doorn, (1985), The construction of a scale to measure perceived effort. Department of Philosophy and Social Sciences, Delft University of Technology.

Zijlstra, F. & Meijman, T. (1989). Het meten van mentale inspanning met behulp van een subjectieve methode (measurement of mental effort with a subjective method). In T. Meijman (Ed.), Mentale belasting en werkstress. Een arbeidspsychologische benadering. (pp. 42-61). Assen, The Netherlands: Van Gorcum.

Östlund, J., Nilsson, L., Carsten, O., Merat, N., Jamson, H, Jamson, S., Mouta, S., Carvalhais, J., Santos, J., Anttila, V., Sandberg, H., Luoma, J., de Waard, D., Brookhuis, K., Johansson, E., Engström, J., Victor, T., Harbluk, J., Janssen, W., Brouwer, R. (2004) Deliverable 2 - HMI and Safety-Related Driver Performance. Human Machine Interface And the Safety of Traffic in Europe. Project GRD1/2000/25361 S12.319626