**Article:**
Carsten, O, Kircher, K and Jamson, S (2013) Vehicle-based studies of driving in the real world: The hard truth? Accident Analysis and Prevention, 58. 162 - 174. ISSN 0001-4575

https://doi.org/10.1016/j.aap.2013.06.006

# Vehicle-based studies of driving in the real world: the hard truth?

Oliver Carsten[a1], Katja Kircher[b] and Samantha Jamson[a]

[a]Institute for Transport Studies, University of Leeds, Leeds LS2 9JT, UK
[b]VTI, The Swedish National Road and Transport Research Institute, SE-581 95 Linköping, Sweden

## Abstract

Real-world studies of driving behaviour and safety have face validity and have the distinct advantage of focussing on driving in its natural habitat. But their very naturalism can lead to problems with confounds and with noise in the data. This paper reviews the three major categories of on-road studies — controlled observation studies, field operational tests and naturalistic driving studies — and discusses the major applications of each study type. It also assesses some of the methodological issues that arise in one or more category of study.

Key words: On-road studies, field operational tests, naturalistic driving studies, field studies.

## 1. Introduction

Real-world studies of driving behaviour do not form a homogeneous group. They vary in scale from short-term observations using a single vehicle to investigations of driving behaviour over many months with dozens and even hundreds of vehicles. They may use simple equipment and sometimes just human observers or they may record data via elaborate instrumentation and data storage systems. In terms of study design they can range from naturalistic investigations of driver behaviour and the quality of their performance to experimental studies focussed on one or more interventions. In recent years these experimental studies have been targeted particularly at the impact of driver assistance systems on behaviour and safety.

Small-scale studies with highly instrumented vehicles go back to at least the 1960s: Michon and Koutstaal (1969) describe a general-purpose instrumented vehicle with the facility to record lane position, longitudinal and lateral acceleration, steering wheel movement, brake, accelerator and clutch activation and psychophysiological data such as heart rate. The vehicle also used video to record driver eye movements and the external scene. Earlier studies using instrumentation in vehicles include those of Hulbert (1957) on drivers' physiological response to traffic events and of Brown (1967) on the effect of time on task on driving quality. Studies of near accidents using both in-vehicle observers and instrumentation go back at least as far as the early 1950s (McFarland and Moseley, 1954). The reduced cost and growing capabilities of systems for data capture on board vehicles have stimulated field operational tests (FOTs) to evaluate how drivers behave with new driver assistance systems and to indicate whether the use of such systems improves behaviour. The same data acquisition systems have enabled large-scale naturalistic driving (ND) studies to reveal the patterns of behaviour in everyday driving and the antecedents to safety-related events and incidents.

[1] Corresponding author. Tel +44 113 34 5348. Fax +44 113 343 5334. Email o.m.j.carsten@its.leeds.ac.uk.

But the fact that studies are conducted in the real world and thus have face validity does not guarantee their usefulness or scientific rigour. Real-world studies can be difficult to manage and costly to conduct. The very feature that hallmarks real-world studies entails that they are subject to a large number of uncontrolled confounding factors so that detecting a signal, let alone the right signal, through the noise can be difficult. It is important to use studies to enhance the power of the data collection and anticipate confounds that could undermine validity.

This paper examines the broad categories of real-world studies of driver behaviour and safety. The number of such studies is huge, and there is no intention here to review them all. Rather, exemplars are used as illustrations for the purpose of a discussion of the rationale for carrying our various kinds of studies and of some of the methodological pitfalls that can lead to inconclusive results and even undermine study validity. The paper attempts to identify the major ways in which on-road studies can make a scientific contribution while acknowledging that a consensus on methodology does not generally exist and that the findings from such studies need the same scrutiny that is applied to more fully controlled laboratory and simulator studies. On-road studies fall into three major categories, which are described in more detail in the following sections.

# 2. Categories of on-road studies

The first category consists of relatively small, targeted and controlled studies conducted to study how driving behaviour and performance are affected by, for example, fatigue, alcohol or distraction or to look at an intervention in the very short term. Such studies typically collect data on minutes or hours of observed driving. The second grouping is a large-scale and more long-term evaluation of a treatment, often called a Field Operational Test. This method has been applied particularly to the evaluation of driver assistance systems such as Adaptive Cruise Control and Intelligent Speed Adaptation, but it is arguably just as appropriate for the evaluation of other types of intervention such as driver education and training programmes. Studies of this type tend to collect data on days, weeks and even months of driving. The final grouping is that of Naturalistic Driving Studies which focus not on treatment but on diagnosis — on enhancing the understanding of how safety problems arise and unfold. As with FOTs, weeks and months of data are normally collected.

Very recently a new hybrid of FOT and NDS has emerged, in which the data collection itself becomes the intervention. Here the data may be used in a fleet context for monitoring driver behaviour for use in providing positive or negative feedback to drivers (e.g. Hickman and Hanowski, 2010). It may also be used to provide post-qualification training to newly qualified drivers as in the U.S. programmes targeted at teenage drivers where video information on incidents is captured to provide feedback to the drivers and in some cases to alert parents to the behaviour of their offspring (Carney et al., 2010).

In the description of each of the approaches below, examples are provided along with the advantages and disadvantages of using the approach, supplemented by suggestions of appropriate research questions.

## 2.1 Controlled on-road studies

A controlled on-road study can provide researchers with a limited range of data that can be highly tailored to the research questions under investigation. The defining feature of a controlled on-road study is in its reliance on using a pre-set route to reveal differences in behaviour and performance, when driving under different conditions. In this respect, the onus is on the researcher to identify a route that affords them the best opportunity of being able to evaluate their hypotheses. Unlike a

naturalistic driving study, there is no question as to whether the road user does or does not encounter a particular traffic situation – this is predetermined by the characteristics of the route and the instructions provided by the experimenter.

Such studies also retain many features of a traditional experimental study, whereby extraneous (confounding) factors such as weather, time of day, and traffic conditions can be controlled for, to some extent. It is not possible, of course, to absolutely control such natural phenomena but rather to control when and where the study might take place. One of the advantages of using a predetermined route is that traffic data (counts and flows) can be used a priori to establish appropriate scheduling, for example if one wishes to study behaviour in peak versus off-peak traffic conditions, or time of day. Weather, on the other hand, is less predictable but can be inferred via meteorological data or via CAN data (e.g. activation of fog lights or windscreen wipers). Researchers may typically limit their controlled studies to weather conditions that are of relevance to their research hypotheses, and where that is not possible, use them as covariates in the data analysis.

Taking the definition of a "controlled study" to its very extreme, use of facilities such as the Virginia Smart Road (http://www.vtti.vt.edu/virginiasmartroad.php), a closed test-bed research facility which features weather-making capabilities (rain, snow, fog), can provide some insight into behavioural adaptation in such conditions. However, whilst weather can be added in at this type of facility, it is significantly more challenging to remove it; so, to a certain extent, the problem remains.

It can be the case that the use of a predetermined route cannot be avoided. This is particularly so where the study aims to evaluate infrastructure interventions such as new road design (de Waard et al. 1995) or infrastructure based ITS applications. For example, Brewer et al. (2011) conducted an evaluation of an intersection violation warning prototype with 87 drivers navigating a predetermined route on public roads. The route included 13 intersections with roadside communications equipment: three intersections controlled by traffic signals and 10 controlled by stop signs.

Controlled trials, due to their relatively short duration, have the added bonus of being able to accommodate an observer. This can be invaluable in providing context to drivers' behaviours, where even the most sophisticated camera system may fail (such as observing driver-to-driver non-verbal behaviour). Using observers as data collection tools predates the present sophistication of black-box monitoring. For example, McGlade (1963) evaluated almost 30 aspects of driver behaviour including parking skill, gear use, lane observance, attention and the use of the accelerator. Quenault (1966, 1967) focussed less on aspects of basic driving skills and more on style, by obtaining measures of speed, use of signals, overtaking and mirror usage.

Furthermore, the presence of observer(s) in the test vehicle can supplement the objective data collection by using techniques such as the Wiener Fahrprobe which is essentially a method akin to a driving test. Total counts of the number of negative interactions are made including unsafe merging/gap acceptance at junctions, incorrect lane changes, poor interaction with other road users, unsafe overtaking and headway choice. The Wiener Fahrprobe has been used in evaluation studies of a number of driver support systems (e.g. Chaloupka and Risser, 1995). Researchers have adapted the original tool, which was designed to be used overtly, to study driver behaviour covertly. For example, Brühning et al. (1989) used the technique to observe car drivers from a vehicle following behind. The Wiener Fahrprobe has also been adapted by reducing the number of observers (Almqvist and Nygård, 1997).

Controlled studies also permit the compression of exposure to the elements under investigation, creating the opportunity to study learning or adaptation effects (e.g. Jamson, 2006). Due to their

relatively short duration, controlled drives may not be able to study long-term adaptation. Even if participants return for numerous drives, it is unclear if the period between the controlled drives affects the subsequent drives, or indeed if we can truly consider experience as being cumulative in this case.

A further criticism of using controlled studies where there is an observer present is the unknown effect that the presence of that observer may have on driver behaviour. For example, whilst Höfner (1967) reported that the behaviour of moped riders did not change when they were aware of being observed and Hjälmdahl and Várhelyi (2004) report similar findings for car drivers, Rathmayer et al. (1999) found that participants' mean speed was approximately 2 km/h lower when an experimenter was present. Additionally, the authors report reduced lateral and longitudinal accelerations. So there is limited and conflicting evidence with regard to this type of bias, perhaps partly due to the ethical implications of this type of study: informed consent is an inherent part of participant handling. There is some comfort in the results obtained in on-road studies (whether data are collected subjectively or otherwise) whereby participants are captured violating traffic laws, even when they are aware (or have forgotten) they are being monitored (Dingus et al., 2006).

Controlled on-road studies are best suited to research questions that are likely to be independent of exposure (i.e. there are unlikely to be novelty effects of a system or road design element) and that utilise independent factors that are stable over shorter periods of time, such as age and personality. They are also excellent tools in the early stage of either system development (a researcher can get to see and hear how the driver really interacts with the system) or FOT design. With the latter, the observer can start to understand what the requirements of a data logger might be; for example if during the controlled drives it is apparent that drivers' headway or their interaction with vulnerable road users is affected, then additional sensors (e.g. radar) may have to be installed in a large-scale FOT. We could thus consider controlled drives to be part of a piloting phase that extends beyond software and hardware test procedures.

## 2.2 Field Operational Tests

Field Operational Tests apply real-world studies for the evaluation of an intervention. Typically the data collection process is automated, at least as regards objective data on situation and behaviour. Subjective data, for example on acceptance of the intervention, is normally collected manually, although questionnaire data and user feedback can also be collected electronically. The increasing storage capacity of modern data logging equipment enables researchers to record more and more data over longer time periods without frequent visits to physically download data from the installed data acquisition system. It has even enabled the use of multiple video cameras that continuously record both the driver and the surrounding traffic. Extended video recordings are, though storage intensive and time consuming to analyse, advantageous for several reasons. One is quality assurance; anomalies in the logged data can be double-checked, which can remove doubts regarding the correctness of the data. Further, video data provide valuable additional information on the situation and activities both in the vehicle and outside. For example driver engagement in most secondary tasks can only be reliably detected via the analysis of videos, and the behaviour of surrounding vehicles as well as the status of traffic lights is often more reliably obtained from video recordings. Fast wireless communications even permits the manipulation of experimental conditions remotely and, within limits, the automatic transmission of the recorded data over the air.

FOTs have been used extensively in the last twenty plus years to study the impacts of new driver assistance systems and other new technologies such as cooperative systems on driving and traffic with a particular focus on safety impacts, although FOTs have also been used to investigate the impacts of systems targeted at reductions to the environmental impacts of driving (green driving

aids) and even at traffic and transport information more generally (e.g. travel information delivered by nomadic devices).

The latest version of the FESTA Handbook on FOT methodology (FESTA, 2011) defines an FOT as: “A study undertaken to evaluate a function, or functions, under normal operating conditions in environments typically encountered by the participants using quasi-experimental methods.” A function is here defined as one capability delivered by a system: “an implementation of a set of rules to achieve a specified goal”. A system can deliver one or more functions.

The major justification for FOTs is their ability to provide continuous information about the performance indicators in the context of the driving situation during “normal” (i.e. relatively uncontrolled) driving over a long time period. The participants use the vehicle in their everyday driving, whether in a private or fleet context, and generally have few interruptions and little direct contact with the experimenter. The data collection systems are made as unobtrusive as possible in the hope that the participants will forget that they are being observed. Subjective data may also be collected, often at particular time intervals during the study, but such contacts tend to be infrequent.

The standard approach uses a quasi-experimental methodology in which the function or functions are inactive for a period to provide a baseline, and in which subsequently there is a period with the function(s) active to assess the impact of function availability and/or usage. Fully counterbalanced designs tend not to be used because it is assumed that there will be a substantial learning impact of function availability, but ABA designs have been applied to examine such learning effects (Lai et al., 2007) and counter-balancing has been used when investigating the impacts of a range of functions both singly and in conjunction (Regan et al., 2006). Thus the major rationale is to compare baseline behaviour to behaviour with the system activated, and to assign behavioural changes to the influence of the system. A participant may use the vehicle over an extended period of time, making it possible to study long-term effects, which includes how the participant adapts to the system over time, and possibly also whether the driver learns to misuse the system.

Typically (but not necessarily) the participants are handed special test vehicles. This reduces the cost of equipping the vehicles and reduces one aspect of variability between participants, but also implies that participants may not drive in the same manner as they would in their own vehicles.

This approach is neatly summarised in the report from an early FOT on a driver assistant system, that on Adaptive Cruise Control by Fancher et al. (1998): “In this manner, the vehicles were put into naturalistic use, without constraining where the person drives, or when, or how. Each driver was also free to choose between operating manually or with conventional cruise control [all the vehicles had cruise control] during the first week and between manual or ACC driving during the second (or subsequent) weeks” (p. 12).

FOTs were originally called field trials, large-scale trials (e.g. Biding and Lind, 2002), experiments (e.g. Rillings and Lewis, 1991) or “Operational Field Tests” (e.g. Fleischman, 1991; Gilbert et al., 1991). The first use of the term “Field Operational Test” seems to date from the mid-1990s (e.g. Horan et al., 1994). The term is used extensively in the report of a U.S. workshop report held in 1995 (ITS America Safety and Human Factors Committee and National Highway Traffic Safety Administration, 1996). More important than the label is the concept, and here the real-world trial of the Siemens-designed route guidance system LISB (Leit- und Informationssystem Berlin), carried out in Berlin in 1989-1990, has a legitimate claim to be the first FOT. There were 700 equipped vehicles and the study featured system usage data, questionnaire data and automatically recorded routeing information in order to capture compliance with system advice (Sparmann, 1989). Unfortunately, the full results were never

published, reportedly because they were not very positive, but the evaluation methodology was widely discussed (e.g. May et al., 1988).

In LISB, HMI and safety aspects were considered as side-effects — the central focus was on journey time and traffic efficiency. To some extent the same was true of the evaluation of the TravTek traffic information and navigation system trial with 100 vehicles, mainly rental cars, in Orlando, Florida, from 1992 to 1993. But in TravTek, safety-related data was observed directly rather than evaluated by questionnaire. Eighteen visitors to Orlando and twelve locals were observed while using the in-vehicle system in an instrumented vehicle equipped with video cameras for eye movement recording (Dingus et al., 1995). Other safety-related data such as longitudinal and lateral acceleration and lane tracking was also collected. An in-vehicle experimenter/observer was also present. Drivers drove from fixed origins to fixed destinations using a variety of navigation methods, including a paper map and various configurations of the TravTek system (map display or turn-by-turn display in each case with and without voice guidance). The overall conclusion was that the system led to better safety outcomes than the alternatives and that the turn-by-turn configuration performed best.

However, this study really falls into the category of a controlled on-road experiment, as opposed to a naturalistic FOT. There was an additional safety investigation in TravTek, using the data from all the general fleet vehicles, but this study mainly relied on micro-simulation of the routes chosen by drivers of vehicles equipped with the system to the drivers to calculate risks of involvement in crashes (Perez et al., 1996). The first “classic” FOT, featuring an investigation of a driver assistance system using continuous recording of system usage accompanied by vehicle-based data, was the investigation of Adaptive Cruise Control (Fancher et al., 1998). This launched a series of safety-related FOTs conducted by the University of Michigan Transportation Research Institute, most of which were focussed on prototype systems: the Advanced Collision Avoidance Study on Forward Collision Warning and Adaptive Cruise Control (University of Michigan Transportation Research Institute and General Motors, 2005); the Road Departure Crash Warning FOT on lane departure warning and curve speed warning (LeBlanc et al., 2006); and the light vehicle and heavy vehicle Integrated Vehicle-Based Safety Systems FOT investigating driving with Forward Collision Warning, Lane Departure Warning, Lane Change Warning and for cars Curve Speed Warning (Sayer et al., 2010a; Sayer et al., 2010b). While the initial study on Adaptive Cruise Control had video recording of the roadway but not of the driver, the next two studies captured continuous video of the driver’s face. In the integrated systems study, a different approach was adopted: multiple interior and exterior video was stored around events and video was also stored at fixed intervals in non-event driving.

In Europe, there was a parallel set of FOTs focussing on a single safety system, namely Intelligent Speed Adaptation (ISA), using simpler data acquisition systems with no video recording. Examples are the Dutch trial in Tilburg (Duynstee et al., 2001), the Swedish large-scale trials of ISA (Biding and Lind, 2002), the UK trials of ISA (Carsten et al., 2008), the French LAVIA project (Ehrlich et al., 2006) and two projects in Denmark (Lahrmann et al., 2001; Lahrmann et al., 2012). Not all were of the same quality: in the Swedish large-scale trials only a small minority of the vehicles were fitted with data acquisition systems and the only thorough analysis was that of the Lund trial (Hjälmdahl and Várhelyi, 2004). An FOT in Australia investigated behaviour with multiple systems — ISA, following distance warning and seatbelt reminders — both alone and in combination (Regan et al., 2006).

In the last few years, with funding from the European Commission, an FOT has examined the safety impacts of a number of driver assistance systems: the euroFOT project. Among the systems investigated were Adaptive Cruise Control, Forward Collision Warning, Lane Departure Warning, Blind Spot Monitoring and Speed Limiting. They were generally studied separately in a set of sub-

FOTs hosted in different parts of Europe. Although the project as a whole featured a large number of vehicles, the individual sub-FOTs were obviously much smaller, leading to inconclusive results on the safety impact of most of the systems investigated. Only for the combination of Adaptive Cruise Control and Forward Collision Warning was it possible to conclude that there was a positive effect on safety (Malta et al., 2012).

FOTs provide almost the only sensible methodology for assessing long-term driver behaviour with new in-vehicle systems and how that behaviour does or does not affect safety in comparison with a non-equipped baseline. They are able to capture the nuances of behavioural change and also, with the right equipment, to identify side effects such the response to false positive warnings and problems with missed warnings, if those can be identified. Usability issues can also be identified. Data is normally retained, so that it can be reused for subsequent analysis. An example here is the examination of the environmental impacts of ISA, using the speed data from the UK ISA trials (Lai et al., 2012). Indeed, the data is often so rich that it is under-exploited.

But FOTs are by no means straightforward. They are complex and costly to conduct, and as a result are frequently of small size. The largest cost can be associated with equipping the test vehicles. For example, the UMTRI Automotive Collision Avoidance System FOT featured only ten equipped vehicles, but these vehicles were used by 66 drivers who experienced the “mature” version of the system, thus counteracting the sample size problem (University of Michigan Transportation Research Institute and General Motors, 2005). As already discussed, the power of the euroFOT observations was insufficient to detect a safety benefit from several of the systems that were assessed, though of course it is also possible that this was because the systems in fact have no positive effect. If the effect size for the parameters of interest were known in advance, then it would be possible to design a study of sufficient size, but FOTs are by their nature exploratory, i.e. intended to reveal the effect size.

Another methodological issue is that there is little understanding of the time required for familiarisation with a system so that the process of assimilation is imperfectly understood. Few studies separate or reject the data from the first few hours or days with a function. The literature review on behavioural adaptation to new driver support systems conducted in the AIDE project concluded that the process of learning how to use and behave with a new system had hardly been studied at all (Saad et al., 2005). So one can recommend separating out the phase of learning how to use the new system, but it is not possible to provide a categorical recommendation on how long that phase should be. And there is an additional complex: some systems such as speed advice or support systems are virtually continuous in operation and users mat receive frequent system responses, while other systems such as Forward Collision Warning are only triggered rarely. In the U.S. Advanced Collision Avoidance System FOT, the FCW alert was triggered approximately once every 148 km (University of Michigan and General Motors, 2005). The experience accumulated with a system that triggers only rarely, such as a FCW, is likely to be quite different to one that provides continual feedback, such as a green driving support system.

One topic that is relatively little discussed in the literature is the time period that is needed to capture long-term behaviour. One might hope that an FOT of several weeks or months would be sufficient to allow driver behaviour to stabilise, but evidence from at least some FOTs would contradict this. In the Swedish trial of Intelligent Speed Adaptation (ISA) in Lund, nearly half the participants did not reach any stability in their overriding behaviour of the system after 2,500 km. In the counterpart UK trial, most drivers never stabilised their overriding behaviour (Lai et al., 2010). A study of a purely warning ISA found no real indication of stable behaviour in response to the system even when looking at three years of driving (Wallén Warner and Åberg, 2008).

Experimental design can also be an issue. It is becoming increasingly difficult to disable driver assistance systems, particularly when using production vehicles as opposed to prototypes. As a consequence it is also increasingly difficult to identify the appropriate baseline. If many recently produced vehicles are fitted with a driver-set speed limiter, does that system then become the baseline for looking at the impact of Intelligent Speed Adaptation? Even in the 1990s, Fancher et al. (1998) used cruise control as the comparison for Adaptive Cruise Control.

Finally there is the fraught issue, particularly relevant to FOTs but also an issue in Naturalistic Driving Studies, of the relationship between observed safety indices and risk of involvement in a crash. For FOTs on safety-relevant systems, it is of course vital to come to conclusions about likely impacts on safety with large-scale adoption of the system or systems being investigated. It is possible just to look at the *direction* of changes and conclude, for example, that, if a system such as Forward Collision Warning produces fewer severe incidents when active than in the non-active baseline situation, it must be positive for safety. However, this finding does not provide a precise forecast of the size of the expected reduction in say injury accidents with widespread introduction and it may not consider the possible side effects of system usage. Here an interesting example is the perception by vehicle purchasers that Lane Departure warning is an aid for fatigued driving and the consequent potential for users to engage in more prolonged night-time driving because they now have a system which can assist them.

In the case of FOTs with systems that are intended to address directly certain negative behaviours with well-known relationship with accident and/or injury risk — alcohol, speed and belt-wearing for example — translating observed behavioural effects into global prediction of changes in accident numbers or severity is relatively straightforward. Models from the literature on alcohol and risk (e.g. that of Hurst at al., (1994) or on speed and risk (such as the version of the power model relating speed to crash severity developed by Elvik et al., 2004, or one of the models discussed by Aarts and van Schagen, 2006) can be applied to translate behavioural changes into predicted changes in accident numbers. But this is far more problematic for say a driver alertness monitoring system.

The FESTA definition makes clear the focus of FOTs is on technology-based systems. This focus is not inherent in the methodology — there is little reason for not considering its application to non-technology-based interventions such as the use of a training regime as an alternative to longitudinal studies focused on such outcomes as recorded or self-reported crash involvements. FOTs have the potential to provide rich data on the effectiveness of such interventions. In the area of work-related road safety Helman and Grayson (2011) recently lamented the lack of robust evidence on the efficacy of interventions. Their literature review "concluded that the task turned out to be a difficult one, largely because of the scarcity of good data. If one adopts the criterion that an evaluation study should assess whether an intervention has brought about a statistically reliable change in crash rates, then the results were meagre in the extreme" (page 6). They advocated greater use of in-vehicle data recorders for the evaluation of interventions: "IVDR data offer a proxy measure that can be used to assess the impact of interventions and changes over shorter timeframes than accident statistics, and with more objectivity than attitudinal measures" (page 17). However, the reality is they have not been used much if at all for studies beyond the realm of technical systems. While the methodology is not inherently limited to safety-related systems, most FOTs to date have focussed on such systems.

## 2.3 Naturalistic Driving Studies

ND studies usually serve several purposes, such as the collection of baseline data, reflecting "normal driving", and the investigation of associations between different variables, such as mobile phone use and crash occurrence (Klauer et al., 2006, 2010). The baseline data can be used to investigate the

prevalence of certain behaviours, driver states, or any other phenomenon of interest that can be found in the log data. Behavioural changes over time as well as reactions to external influences that happen to take place can be investigated — examples of which could be as diverse as financial crises or new traffic regulations. The association of certain behaviours with road layout, road type or other static factors in the environment might be investigated, as this information could be used for strategic countermeasures against, for example, sleepiness or distraction.

In many cases one main goal is to collect pre-crash data, often with the aim to find the true reasons for crashes (Dingus, et al., 2006; Transportation Research Board, 2012). This is nicely illustrated by a quotation from Kenneth Campbell, the SHRP2 Chief Program Officer:

> Progress in traffic safety has been limited up to now by a lack of accurate and objective information on pre-collision conditions and contributing factors. The lack of such information particularly affects our ability to assess the role of driver factors in a collision. It is commonly believed that driver behavior or actions play a significant role in nearly all collisions. Up to now, rigorous, exposure-based risk analysis of driver, and most other pre-collision factors, has not been feasible. Consequently, researchers have been limited in their ability to determine how roadway and traffic conditions interact to increase or decrease high-risk driver behaviour/actions.
>
> The same advanced technology that enables intelligent vehicle safety – not previously feasible – also enables the near-continuous collection of a vast array of data. This includes data on the driver inputs and vehicle motion and position relative to the roadway and other vehicles. This new capability allows study of the entire driving process, including pre-collision and collision events, with an accuracy that could previously only be achieved under laboratory conditions. In particular, objective measures of driver actions in normal driving are now achievable. Continuous recording capability can provide accurate and detailed exposure data as well. (Campbell et al., 2003, p.11)

This vast array of data will allow a reconstruction of the last seconds, minutes and, if need be, longer time periods leading up to a crash in as great detail as the available data allow. In this sense ND studies have more of a diagnostic character, as they are used as an instrument to find out which factors are associated with crashes and conflicts, while they, in contrast to FOT studies, do not systematically investigate a countermeasure or other treatment that ultimately might prevent crashes from occurring. As the nature of the study is rather explorative, research questions tend to be more open-ended than for other study types.

In a typical ND study a large number of vehicles is equipped with a host of sensors, somewhat similar to what is used in FOT studies (see above). The instrumentation often includes video of the road scene outside of the vehicle, complemented by a view of the driver's face, and possibly with an over-the-shoulder view of the driver's hands. The instrumentation is installed as unobtrusively as possible, both to make the driver forget about his or her being constantly observed while driving, and also to prevent other drivers from changing their behaviour upon noticing cameras in a car in the vicinity. Usually the data logging is continuous at a given frequency (e. g. Neale, 2002; Hanowski et al., 2000), but recently DriveCams installed in vehicles have been used as data acquisition system as well (Hickman and Hanowski, 2010; 2012; Carney et al., 2010). Those systems only save the sequence around a triggered event and discard the rest.

While typically in FOTs the drivers are provided with instrumented vehicles owned by the research institute, in ND studies it is common that the private cars of the participating drivers are instrumented. The drivers then use their vehicles in their daily lives, just as usual, without any special instructions. The typical duration of an ND study is longer than of an FOT, partly due to the fact that it is costly to instrument private cars, making it uneconomical to move the instrumentation from one car to another very frequently. This enables a systematic evaluation of long-term effects like seasonal variations or, especially in the case of novice drivers, of learning effects.

The first ND study was conducted on short haul trucks, where 42 drivers drove instrumented trucks for two weeks each (Hanowski et al., 2000). The 100 car study conducted by VTTI in 2003 (Dingus et al., 2006; Klauer et al., 2006) was a leap in size from there, with 100 cars collecting data over one year. In Japan a naturalistic driving study with 60 vehicles was conducted, with the vehicles running for up to two years (Uchida et al, 2010). A current study running as part of the SHRP2 programme is the biggest of its kind so far, with a planned 3000 vehicles providing data over a period of two years (Transportation Research Board, 2012). In Europe, the UDrive project commenced 2012. It is the first large-scale naturalistic driving study of the continent, and will collect data from just short of 500 vehicles, including passenger cars, trucks and powered two-wheelers (ERTICO, 2013).

While in most cases ND studies focus on passenger cars and trucks, an ND study with motorcycles was launched in autumn 2011, with 100 participating motorcycles collecting data for one year (McLaughlin, 2010). In Europe too an ND study for powered two-wheelers is under way (Spyropoulou et al., 2010).

The selection of drivers can be from a certain group of interest, such as teenage drivers (Lee et al., 2011) or long-haul truck drivers (Barr et al., 2011; Blanco et al., 2011). The participants in the 100 car study were drawn from one geographical region, while in SHRP2 drivers are being recruited from six different regions of the United States. All drivers have to fill in informed consent forms, and are therefore aware of their being observed and have agreed to it.

At the time of the 100 car study the storage capability of the data acquisition systems (DAS) was somewhat limited, such that immense logistic efforts had to be made to exchange hard disks in the vehicles on a regular basis without disturbing the driver's natural rhythm of using the car, and without reminding him or her too much of being in a study (Neale et al., 2002). The quick technical development allows larger and larger datasets with higher data rates to be collected over ever longer time periods, which is a necessity considering the scope of SHRP2.

The largest advantages with ND studies are clearly the high external validity, the possibility to study behaviour over an extended time period, and the possibility to obtain prevalence data for different types of behaviour. As large amounts of data on "normal driving" are gathered, it is possible to find out when drivers choose to use their mobile phones, how they use their navigation systems and for how long they drive without taking a rest, among many other things. These prevalence data are a necessary complement to crash databases, as they allow the assessment of whether certain variables are overrepresented in crashes or not. In ND studies a vast amount of data is gathered, which can be used for a broad number of research questions, and also for data mining to generate new questions and hypotheses. This justifies the high costs associated with data collection and data reduction.

One of the purposes of ND studies is to collect pre-crash data from actual crashes that happen in the real world. These data are unique for ND (and possibly FOT) studies, especially since for controlled on-road studies precautions are taken to avoid crashes. While event triggered logging equipment like DriveCam and other "black box" systems usually record short sequences before a crash, the

continuously collected data still allow a deeper analysis of crash-preceding factors over a longer time span. This way, not only the last movements and constellations leading up to the crash can be evaluated, but also the underlying factors that may have led to the driver's ending up in a certain situation at all.

The practical disadvantages with ND studies are that they are expensive and require a large logistic effort to conduct. Also, some variables are not or not yet possible to collect during ND studies. Physiological data that nowadays require the application of electrodes can serve as an example where developments in technology still may allow remote data in the future. Other data that require the participant to be active, as for example think-aloud reports, could possibly be collected in terms of technical maturity, but there may be ethical objections against recording sound. Also, encouraging the participant to do something that he or she otherwise would not have done, like thinking aloud while driving, goes against the very idea of observing naturalistic behaviour.

On a more theoretical note, as no variables are controlled by the experimenters, causal conclusions can, strictly speaking, not be drawn. Therefore, this type of study is not suitable for the investigation of how a certain support system or any other kind of treatment affects behaviour in a given situation. However, associations between different variables can be uncovered. For example, if it is found that talking on mobile phones is associated with fewer critical incidents (e.g. Hickman and Hanowski, 2012), it cannot be directly inferred that talking on a mobile phone in itself makes driving safer. Instead, a third variable might influence the two others; it could be possible that drivers only use their telephone in low-risk situations. There might also be an intermediate process, in the sense that drivers who use a mobile phone are aware that they need to exert extra effort, and thus, they even overcompensate by being extra attentive. Hickman and Hanowski (2012) speculate that drivers unintentionally focus their glance more to the road ahead while talking on the phone, which improves threat detection in the forward roadway. Whether any, all or none of those explanations are true cannot be concluded based on ND data, but valuable information for hypothesis generation is delivered.

Typical research questions addressed by ND studies are connected to finding out what really does happen out on the roads. The research questions are often of a more open, explorative nature, looking for relationships rather than for cause and effect. In a sense, the researchers are "at the mercy" of the participants in the study, as the drivers choose whether they expose themselves to certain situations or not. For example, if one is interested in collecting naturalistic data on elderly women driving in the dark, it is of course necessary to include elderly women in the driver sample, which in itself can be challenging. To wait until they have driven enough in the dark to sample the amount of data needed can be quite time consuming. Then again, if they are asked to drive in the dark, it is not sure whether they exhibit their natural behaviour, as they may feel pressed to do something they otherwise would not have done.

In addition to dedicated ND studies, FOT baselines have also been used to extract data that are considered to be equivalent to pure ND studies (Green et al., 2007). The assumption is made that, as no new system had yet been activated for investigation, the baseline observations represent the drivers' natural behaviour. Usually the baseline periods of FOTs are relatively short in comparison to dedicated ND studies; this has to be taken into account during analysis and interpretation.

## 2.4 Impact of on-road studies

While lacking some of the control available in laboratory and simulator studies, on-road studies provide the unique possibility to study driving in real traffic, with all its complexity, and especially for FOTs and ND studies the ecological validity of the data can hardly be disputed. For stakeholders such

as governments and politicians, FOTs constitute a gold standard of evidence on safety impacts that can persuade them to promote or even require new systems. The progress on deployment of Intelligent Speed Adaptation provides a good example. The various real-world tests proved that the technology was reliable and that there was a substantial impact on speed compliance. Swedish government vehicles are equipped with ISA technology as a result of the Swedish trials (Biding and Lind, 2002), and the evidence from FOTs has persuaded Euro NCAP to offer extra rewards to ISA-equipped vehicles under its "Safety Assist" scheme for evaluating advanced safety systems (Euro NCAP, 2012). Shortly before he failed to win re-election, President Sarkozy vowed to deploy ISA because he was convinced of its safety benefits, presumably based on the results of LAVIA (Ehrlich et al., 2006).

Without naturalistic studies with highly instrumented vehicles we would for example have to rely on self-reports and observations for estimating the frequency of different types of secondary task engagements, making the studies an invaluable tool for the objective establishment of secondary task prevalence, both during incident free driving as well as in crash relevant events (Green et al., 2007; Klauer et al., 2006; Stutts et al., 2001).

The research on driver distraction benefits greatly from this, as prevalence data on distracting activities are notoriously hard to obtain. Only with reliable prevalence data obtained both for incident-free driving and for crashes and safety critical events it is possible to establish risk assessments of certain activities, such as for example mobile phone usage. Of course, it still has to be made sure that the crash relevant events as well as the incident-free driving periods evaluated are extracted from the collected data material in a scientifically sound manner.

Also, FOTs and naturalistic driving studies provide access to behavioural data leading up to crashes and near-crashes, which affords insight into how crashes can occur. The data obtained can complement that from in-depth studies and site-based studies. This knowledge can help advancing effective countermeasures, as well as generate new hypotheses that can then be tested and repeated under more controlled conditions ,either on-road, on test tracks or in simulators.

FOTs and ND studies also allow us to study the development of behaviour over time – be it the learning curve of young drivers, the adaptation to new technology, or possibly even behavioural changes caused by changes in laws and regulations. The recent debates on whether a law that prohibits hand held phone use is complied with and for how long, and whether it has any effect on traffic safety, could have been built upon much harder facts if ND data had been obtained both before and after the introduction of such a law (Elvik, 2011; Kircher, Patten and Ahlstrom, 2011; Redelmeier and Tibshirani, 1997; Young ,2011; Young, 2013).

Falling asleep at the wheel is frequently associated with severe crashes. A controlled study contributed with knowledge about the drivers' physiological state, their self-reported sleepiness and their driving behaviour leading up to the point when the drivers chose to stop driving due to excessive sleepiness (Åkerstedt et al. 2013). These data advance knowledge in how to avoid falling asleep at the wheel in the future. Controlled on-road studies are also particularly well suited to validate simulator results, as for example done by Hallvig et al. (2013), again in the area of sleepy driving, who concluded that the relative validity of simulators was acceptable for a number of different variables, but that absolute values differed.

Controlled studies can also be used to obtain user feedback on new systems relatively early in product development. An example can be found in the assessment of a speed limiter system that was carried out in the Gothenburg area in 1992 ((Almqvist and Towliat, 1993). Roadside transponders to transmit speed limit and other information were placed along a 35 km route around

Lake Aspen. The route was mainly rural, but incorporated a number of villages and speed limits varying from 30km/h to 110 km/h. The system had an information-only mode and a mode in which speed limit was set automatically. The automatic mode was a hybrid of Intelligent Speed Adaptation and Adaptive Cruise Control in that the vehicle would drive automatically at the set speed unless the driver intervened by applying the brake. The drivers commented negatively on this functionality: they felt pressured to driver to fast in the villages and on sharp curves outside the villages.

# 3. Methodological issues

## 3.1 Introduction

For all study types the data that can be logged are limited by the budget of the study, by restrictions on where to put the sensors, and by sensor range and working envelope. While some sensors are very reliable, there can be a number of concerns about others. The data output can be very noisy in general, making it hard to detect the real underlying signal. Eye trackers can serve as an example here – the data quality obtained in field studies is often rather low and tracking is lost often, especially when only one-camera systems are used. This was evident both in the SeMiFOT and the euroFOT projects (Ahlstrom, 2012). So far, only one field study of FOT type delivered largely reliable eye tracking data, with consistently less than 30 % of lost tracking. In this study a two-camera system was used (Ahlstrom, Kircher and Kircher, 2013). Some sensors may perform systematically worse when exposed to certain conditions, like a GPS, that will not find satellites in tunnels, and that may not give a sufficiently accurate signal when between high-rise buildings. Eye trackers can have trouble tracking accurately in strong sunlight, or when the participants wear mascara or glasses. Some sensors are of a somewhat diffuse character, where the logged data cannot necessarily be directly related to driver behaviour. An example is the use of a passive alcohol detection device in the car compartment in the SHRP 2 ND study, which may be as likely to pick up the breath alcohol of the passenger as that of the driver. It can still be useful, both in cases in which the driver is alone in the car, and to serve as a trigger for closer monitoring of a driver for other indications of intoxication. These examples illustrate that each data set is very likely to have limitations with biased, omitted and noisy data. The more information there exists on the limitations, the easier it is to deal with them in a methodologically sound manner.

There is a fundamental difference in data sets from controlled on-road studies and from ND and FOT-type studies. In the latter case the data collection has to be completely autonomous, usually with the sensors being activated upon the turn of the ignition key, and with the DAS being shut down when the ignition is turned off. Additionally, the sensors should be well hidden, and requirements on crashworthiness are high. For controlled on-road studies it is usually perfectly feasible to start and stop the logging equipment manually and to attach sensors like electrodes or head mounted eye trackers to the participant. The equipment does not have to be as well hidden, as it has to be presupposed anyway that the driver is aware of being monitored. Direct observations by experimenters are only possible for controlled studies.

## 3.2 Participant selection

The range of research questions under investigation will determine whether there is a need to select certain groups of participants in terms of their demographics and driving patterns. Whilst age and gender are the most commonly used demographic factors, socioeconomic factors, such as income, education and employment status can influence the exposure to different driving situations and the willingness to pay for e.g. a specific driver support system. Given the well documented influence of personality and attitudes on driving behaviour, many on-road trials incorporate a battery of

psychometric measures. For example, studies have suggested that high sensation seekers drive more recklessly (Burns and Wilde, 1995; Jonah, 1997), whilst Rudin-Brown and Parker (2004) suggest that those with an internal locus of control (Rotter, 1966) adapt differently to ACC, compared to those with an external locus of control.

Saad (2006, p.178) concludes that such findings "reveal that some individual characteristics seem to amplify the behavioural changes observed when driving with new systems and influence the drivers' subjective assessments of the impact of the systems on their driving." Thus personality is an important issue when examining behavioural adaptation in an on-road study, and researchers should consider recruiting on such criteria in order to fully test the impact of a system on driver behaviour. Pre-screening participants according to a personality trait/attitude using psychometric instruments allows the researcher to ensure that a range of drivers with the desired characteristics are included within the study. For example, if researchers are interested in recruiting drivers who express positive attitudes towards speeding, the literature would suggest targeting young males. Male drivers perceive the negative outcomes of speeding as less likely than female drivers (Parker et al., 1992a), younger male drivers perceive greater social pressure to speed (Conner et al., 2003) and younger drivers evaluate the positive outcomes of speeding more positively than older drivers (Parker et al., 1992b). Thus targeting these demographics would produce a sample of drivers with the desired attitudes.

When testing new variants of existing systems, researchers may consider recruiting on drivers' previous experience with systems. For example, Fancher et al. (1998) only recruited participants who reported themselves a priori as being frequent cruise-control users. This is a technique that can be used to avoid the novelty effect associated with first-time system users. For other FOTs it might be of interest to understand how those most likely to buy a particular product (e.g. driver support system) will use it. Thus controlling for vehicle ownership allows researchers to target drivers who are most likely to purchase the candidate system.

## 3.3 Sample size and power analysis

When too few participants are used in an on-road study, statistically proving the effects of the system is more difficult. However, a researcher is normally limited in budget (for equipping vehicles with systems/data collection tools). Sample sizes have varied enormously in the on-road studies reported in this paper. For example, in the ACAS study (Green et al., 2007) ten equipped vehicles were given to 66 drivers whilst the Swedish ISA FOT involved several thousands of cars equipped with Intelligent Speed Adaptation (ISA).

Ideally, a power analysis should be undertaken to determine the appropriate sample size, although various assumptions have to be made regarding the effect size, particularly when a FOT is examining a new system. For example, in the power analysis conducted in the euroFOT project (Jamson et al. 2009) the simulations suggested that as effect sizes become more modest the number of required cars increases substantially. When at least 120 participants, who drive 15,000 kilometres per year, are included, sufficient power would be attained. It was also recommended that including more vehicles or more unique participants should take precedence over measuring for longer periods. For example, collecting data for a year from 60 participants is not as powerful as collecting for six months with using 120 participants. Finally, it was suggested that reducing the variance between participants would improve power. This can be achieved by choosing a homogenous group of drivers, for example male drivers between 30-40 years of age with similar mileage. However, this would be at the cost of the generalisability (external validity) of the results.

## 3.4 When to record

In carrying out FOTs and ND studies, one major consideration is whether to have continuous or only event –related data recording. Obviously this decision depends in part on the focus of the study: it would be perverse to collect data in an FOT examining the effects of Intelligent Speed Adaptation, an ACC or a headway warning system only on events, particularly if there is a concern about side effects such as changes in route choice or lane choice. But the situation is not always so clear-cut, and there may be arguments for restricting data collection to events, as was done in the UMTRI Integrated Vehicle Based Safety Systems FOT, where audio data was only stored for the pre-defined events (Sayer et al.,2010a). Alternatively event data may be coupled with a sample of continuous data. This was the approach used in the UMTRI Automotive Collision Avoidance System FOT (University of Michigan Transportation Research Institute and General Motors, 2005). There pre-determined events caused video to be permanently stored for a period of four seconds before and four seconds after the event. Exposure video data consisting of a single video frame was stored at 1 Hz intervals, giving in effect low quality continuous video.

One major reason for not collecting continuous data is cost savings. DAS storage space can be reduced, data can perhaps be transmitted over the air thus potentially reducing the risk of data loss such as from a disk failure, and analysis effort can be cut substantially. For an FOT, particularly one that is designed for the evaluation of a system targeted at reducing the frequency of certain type of event, such as small times to collision or small times to line crossing, it may make sense to focus the data acquisition on those events. But for a more generally focussed investigation such as a naturalistic driving study, this make much less sense. Granted, data collection that focusses only on triggered events is likely to be substantially cheaper and that increased efficiency can perhaps be translated into a larger sample size. However that reduction in effort comes at considerable cost in reducing the subsequent flexibility of analysis. Pre-defined triggers can mean that it is not possible to examine the impact of setting the triggers at less severe levels. Events that do not have any obvious in-vehicle triggers are unlikely to be captured — thus , for example, potential side collisions which are often denoted by very small Post-Encroachment Time (PET) values may not be captured. Since side collisions form a substantial proportion of overall crashes and since occupant injuries tend to be more severe in side impacts than in frontal impacts, such omission is highly undesirable.

There are also issues concerning how to sample non-incident episodes, especially in naturalistic studies. If the sampling procedure is based in elapsed time or travel distance, more dangerous locations in the road network, such as intersections or sharp curves, may end up with inadequate representation. Therefore it may not be feasible, when using such strategies, to address some specific research questions on driver attention such whether drivers are more attentive in more risky situations and therefore less prone to engaging in non-driving-related tasks.

It is perhaps not totally surprising that the findings of the naturalistic driving studies carried out so far have tended to concentrate on the human element in event causation. They have applied a methodology which is retrospective, i.e. identify an event and then try to determine its causation. In this sense, the methodology can be likened to that applied in in-depth accident studies: identify a crash and determine the contributory factors. Application of the latter methodology has tended to produce findings that emphasise the role of human error as the immediate precursor to a crash as opposed to identifying traffic-system-based problems that make the occurrence of human error more risky (Carsten, 2002). A methodology that allows the analyst to identify, for example, when and where human error is more problematic or whether particular groups of drivers systematically engage in rule violations would permit a focus on the traffic system as a whole, including deficiencies in infrastructure design.

### 3.4.1 Hypothesis-driven data analysis

For any type of study, the collected data will only be a subset of what is happening, and also of what is technically possible to obtain. It is important to keep that in mind when analysing and interpreting the findings from the data available.

More controlled and experimental studies usually investigate more specific hypotheses. Consequently, the data acquisition is geared towards answering the hypotheses in question, which increases the likelihood that all relevant data are collected, but which also makes the data very specific, such that they cannot easily be used for other purposes. For example, in a study that compared sleepy driving in a simulator and on a real road, the collected data contained driving variables, physiological variables like EOG and EEG, and self-reported sleepiness scores (Hallvig et al., 2013). Also, the trials were run distributed over the day, to make sure that the data represented the whole circadian rhythm of the drivers. While these data provide useful and detailed information on sleepiness, as was intended, they are quite specific and therefore not very suitable to investigate other research questions.

In FOTs the effect of a treatment are evaluated against baseline driving, without the treatment in place. The data acquisition system obviously has to be able to record data that are relevant to answering the hypothesis in question. In earlier FOTs, when storage capacity was much lower than nowadays, a common approach was to increase the logging frequency around triggers, which indicated an activation of the treatment under investigation (e. g. LeBlanc et al., 2006), or to record video clips based on triggers and set time intervals, instead of continuously (e. g. University of Michigan Transportation Research Institute and General Motors, 2005).

The data acquisition systems in ND studies are very similar to those used in FOTs. While in FOTs the triggering can be based on the treatment evaluation, for ND studies it is usually not as clear beforehand which data to capture and which to discard. Therefore, one approach is to record a host of data continuously, leaving the data selection process for later (e. g. Klauer et al., 2006). However, the triggered approach has also been used in ND research (e. g. Uchida et al., 2010). Furthermore, triggered data logs that were not originally recorded with research as the first purpose have also been used for the evaluation of naturalistic behaviour (e. g. Hickman and Hanowski, 2012).

The advantage of data logged for a less specific purpose is that they are usable for a broader range of questions. However, especially when using data for another purpose than what was intended originally, it is necessary to be observant, and to acknowledge both the limitations in the data that are collected, and to be aware of the data that were not collected. In order to determine sleepiness it is usually not sufficient to look at short video clips, as shown by Anund et al. (submitted). Therefore, caution has to be exerted when drawing conclusions about sleepiness related behaviour when no other data sources are available to judge a driver's sleepiness level. Similarly, cognitive distraction cannot be observed from facial expressions (e. g. Peng, Boyle and Hallmark, 2013). Therefore, the studies that report on driver distraction (e. g. Klauer et al., 2006; Olson et al., 2009; Hickman et al., 2010; Klauer et al., 2010) are limited to observable instances of "eyes off road" or activities like phone use or eating, that are usually classified as distracting.

The detection of crashes, near-crashes and incidents is a goal of a number of ND studies (e. g. Klauer et al., 2006, Uchida et al., 2010). While crashes typically can be detected with a ND data acquisition system, those near-crashes and incidents that do not set off a trigger in the logged data will be missed systematically, however. While some could be detectable with the help of different trigger criteria, some will not leave any traces and may only be detectable in future studies with a more complete sensory coverage. For example, when there is no side radar, near lateral impacts will be missed completely if the driver does not take any evasive action. Such an event can only be identified if it happens to be observed during a random video scan. A systematic viewing by human

analysts of the complete video material from thousands of hours or even days of driving is and will probably remain practically impossible, but improved image recognition might help in identifying events.

### 3.4.2 Extraction of data from the complete data set

In controlled on-road studies typically the driver's behaviour in a baseline and one or several treatment conditions is compared, which partly determines how the data subsets should be extracted from the data stream. In an evaluation of the effect of a navigation system on driver behaviour, data from the same junction driven once with and once without a system could be compared, and in a study that examines the influence of time-on-task on sleepiness the relevant performance indicators might be monitored at regular time intervals for all wakefulness conditions, for example every fifth minute.

For FOTs and ND studies the picture looks different. As drivers can choose for themselves where and when to drive, and as, in case of FOT studies, not only the operational and tactical, but also the strategic driving behaviour (Michon, 1985) can be influenced by the treatment under investigation, the data extraction procedure is not quite as straightforward. Most FOT studies so far have evaluated driver assistant systems with various functions, from forward collision warning systems, which are intended to intervene in the last second, meaning that warnings can expected to be rare, over intermittent systems like lane keep assist and speed limiters to convenience systems like adaptive cruise control, which can be expected to be in use for a large percentage of the driving time.

The research question might be whether collision avoidance systems usually are triggered when the driver has been distracted. To answer this question, at first all events in which the system triggered would be identified, and the prevalence of distraction during those events would have to be established. Then it is necessary to identify comparable baselines, such that the prevalence of distraction can be studied for those cases and be compared to the prevalence of distraction in the events. The choice of baseline selection is very important for the result, and it requires thorough consideration for which variables the baselines have to be matched to the events. Should they stem from the same driver (case-crossover approach), or from different, but similar drivers (case control)? Should the baselines be collected from the time before the warning was triggered, as such a rather drastic warning might lead to long-term behavioural changes, or rather from all data, as not to bias the selection with respect to time? Should the road type, the weather, the traffic density, the time of day, or any other factor be controlled for? The choice of baselines for systems which operate more frequently may have a different starting point, but in principle the reasoning is the same. Each research question has its own particular set of more or less suitable baselines, such that no general advice can be given except to be aware that the choices made can affect the obtained results.

Depending on the type of system, different behavioural changes can be expected, but it is common for all systems that they may influence behaviour both immediately at the time of the warning or feedback given, and on a larger time scale. A lane departure warning system, for example, will lead to the driver's steering back onto the road as response to a warning, but it may also alter the driver's behaviour such that he or she will use the indicators more to avoid false alarms. An adaptive cruise control system will likely change a driver's behaviour while the system is switched on, but may also lead the driver to choose routes that are suitable for the system, thereby altering the strategic behaviour. As FOTs usually are intended to evaluate safety, acceptance and possibly other aspects of the systems under investigation, it is necessary and important to consider behavioural changes on different structural levels, and to evaluate them with respect to appropriate comparison cases. These can either stem from the baseline phase or from the treatment phase, depending on the hypothesis.

For ND studies the typical event of interest is not the activation of a certain system that usually leaves convenient traces of its activity in the log data. Rather, it is the occurrence of so-called safety critical events (SCE), which usually are split into crashes, near-crashes and incidents. The first challenge is therefore to define such an event in a way that it can be detected automatically in the available data. The sensitivity of the triggers for event candidates should be as high as possible, without sacrificing too much specificity, as the amount of event candidates has to be reasonable to allow visual control for confirmation or rejection of events. In the 100 car study triggers were based on longitudinal and lateral acceleration, forward and rear time to collision, yaw rate and the activation of an event button by the driver (Dingus et al., 2006). When tested on a pre-categorised data set each of those triggers detected between 3.5 % and 57 % of the events, while at the same time arriving at a rate of 60 % to 91 % of invalid triggers. Crashes with a measurable transfer of kinetic energy were identified more reliably than near crashes or incidents. The authors argued that for a larger study, which is likely to produce many more SCE, triggers need to be more restrictive. For feasibility reasons it is more important to arrive at a high specificity, limiting the number of false positives, even though this will lead to a larger number of missed actual events. Moreover, by extrapolating from a smaller, fully categorised set it is possible to estimate the number of SCE that will be missed.

Once the events are identified satisfactorily, baselines have to be found for comparison, and here the issues are quite similar as described for FOTs. Furthermore, there are cases, like in the DriveCam study reported by Hickman and Hanowski (2010), in which only triggered scenes are recorded. The trigger was based on accelerometer data and led to the saving of a video clip of duration of 12 s when a value of |0.5 g| was reached or exceeded. These clips could consist of actual “events”, that is, hard braking, swerving or the like, but they could also be triggered by rough roads etc. A trained analyst separated the recordings into “safety critical events” (SCEs) and “baseline” based on visual inspection. The authors acknowledge themselves that the baseline may not be representative for “true baseline” driving, even though they report similar odds ratios as found by Olson et al. (2009). For a continuous naturalistic data set clips that would have been triggered by the DriveCam equipment could be extracted from the data and compared to “true baseline” clips, in order to investigate this issue further.

While so far a substantial amount of effort has been placed on extracting SCE from naturalistic data sets, those data also hold an enormous potential for many other types of analyses. Examples are the analysis of naturalistic windscreen wiper usage (Wetzel, Sayer and Funkhouser, 2004), the assessment of the variation in fuel consumption between drivers (LeBlanc, Sivak and Bogard, 2010), and the analysis of lane keeping behaviour with and without eyes off road (Peng et al., 2013).

The different study types tend to use different experimental designs. The more fully controlled studies often employ counterbalanced repeated measures designs or case-control designs in which there is both a without treatment baseline and also a non-treatment control group who are investigated at every successive time period. FOTs tend to use a simple within-participant AB (or ABA) design, but arguably more complex case-control approaches should also be applied to them in order to take care of seasonal effects and the impact of unforeseen external events — such as changes in fuel prices or changes in the levels of police enforcement.

ND studies do not have an experimental design *per se*, though of course the location(s) of the study and participant selection is of prime importance. One major analytical approach in some ND studies, particularly when focussing on incidents, near crashes and actually recorded (though generally low severity) crashes, has been to compare events with “baseline” epochs drawn from the recorded driving of the participants in a case-control analysis:

> For these analyses, two reduced databases were used: the 100-Car Study *event database* that consists of the reduced crashes, near-crashes, and incidents; and the *baseline database*. The *baseline database* was created specifically for this analysis by stratifying the entire dataset based upon the number of crashes, near-crashes, and incidents each vehicle was involved in and then randomly selecting 20,000 6-second segments from the 6.3 terabytes of driving data. For example, a vehicle involved in over 3 percent of all the total crashes, near-crashes, and incidents would also represent 3 percent of the baselines. Vehicles that were not involved in any crashes, near-crashes, or incidents were not represented in the baseline database. This stratification of the baseline epochs was performed to create a *case-control* data set where there are multiple baseline epochs per each crash or near-crash event to allow for more accurate calculation of odds ratios. (Klauer et al., 2006, p. viii)

Here the participants provide their own controls thus accounting for long-term *individual effects* such as age, gender, driving experience, personality, etc. However more short-term driver factors such as fatigue or impairment are omitted as are the impacts of roadway type, traffic density and time of day. Therefore the contribution of those omitted factors is ignored. As in any multivariate analysis, this is not a problem provided that the omitted terms do not have a systematic contribution to the phenomenon being studied. It is, however, not difficult to formulate hypotheses about how the omitted factors might affect the results obtained — for example, drivers may be less prone to engage in distraction when traffic densities are high.

The effect of omitted variable bias in ND data analysis has recently been investigated by Jovanis et al. (2011), focussing on events involving road departure. They found large effects of environmental factors such as road surface condition and lighting and smaller but still relevant effects of driver-related variables such as measures of driver risk propensity. They conclude: “It is critically important that omitted-variable bias be identified with naturalistic data. The primary advantage of naturalistic data is that factors not previously observed or estimated through use of judgment are now observable with, it is hoped, a high degree of accuracy and reliability. It would be a shame to give up that accuracy to a poor model specification. Tests with additional data sets are needed to provide verification, but the need to include context variables in event-based analysis seems strong.” (p. 56). They argue for a multi-level analytical approach which examines how drivers with certain characteristics find themselves in a situation in which they execute a specific manoeuvre which in turn leads to specific outcomes.

A recent VTTI report on driver inattention used a case-crossover analysis of the 100 car study data (Klauer et al., 2010). Here the baseline epochs were much more carefully matched to the events, using a within participant approach and matching by such factors as time of day and day of week and to some extent location type. The resulting odds ratios were substantially lower than what was found with the case-control methodology.

## 3.5 Interpretation of results

For a typical controlled on-road study with random assignment of participants to conditions and with balanced experimental manipulations of conditions the typical interpretation is that the manipulated factor constitutes the cause for observed behavioural changes. While this usually is a rather non-controversial interpretation, caution should be exerted with a generalisation to spontaneously occurring behaviour. It is not at all guaranteed that drivers choose to engage in certain behaviours voluntarily, even though they did so when prompted by an experimenter. When the same behaviour was observed during a naturalistic driving study, however, it is a proof that this behaviour actually does occur in normal driving. On the other hand, for studies of this type it is much

more difficult to ascertain cause and effect in observed associations. One such example was a finding by Hickman et al. (2010), which received wide attention both amongst researchers and the general public. An association of the use of mobile telephones with a decreased odds ratio for crashes for commercial truck drivers was observed. This cannot necessarily lead to the assumption that telephone use makes driving safer, and this was not either suggested by the authors. The interpretations that have been offered are widespread and include that drivers may use their telephones in situations that are less crash prone to begin with, that drivers use the telephone as "activation" and therefore fall asleep less frequently, or that drivers tend to look ahead to a greater extent while on the phone, thereby improving their chances of detecting and avoiding forward crashes. The example shows that there can, but does not have to be, a direct causal link between the variables; it is also possible that a third variable, in this case the drivers' self-adaptive behaviour, influences the results.

For FOTs and ND studies it has become common practice to extract very short data clips, based on the notion to look for SCE or system activations. By only looking at those short time intervals, however, it is pre-defined that longer-term effects will not be evaluated. Changing the window size may have direct effects on the results. To give an example, it has been reported that for 55 percent of all 6 s clips in the 100 car study some secondary task engagement could be observed (Klauer et al., 2006). Extending the time window should clearly lead to an increased frequency of clips including secondary task engagement. Changing the window size may also influence the possibility to detect sleepiness, stress, confusion or other driver states. It may also provide insight into how situations build up, and whether there were any priming or prompting signals that led the driver to behave in a certain way. Of course, extending the window size also entails higher costs and efforts in data reduction.

# 4. Conclusions and future developments

The very vigour of on-road studies in recent years provides testimony to their perceived usefulness. As interest in the safety impacts of a variety of driver assistance systems has grown, so has the impetus to evaluate those systems in the real world and observe both the direct and the indirect effects of system usage. Naturalistic studies have gained in popularity in spite of their cost. They offer the prospect of new understanding of how drivers behave and how they respond to changing circumstances.

Data from FOTs and ND studies are less confounded than data from short, controlled on-road studies by factors directly associated with the awareness of taking part in a study. On the other hand, in the case of FOTs and ND studies, there are fewer possibilities to expand the data sets to obtain physiological data including, given the current state of the art, automated eye movement data. Further developments in technology and computational power are likely to lead to more detailed data sets and additional data sources in the near future.

Even though the increase in information density is promising, it is necessary to put effort into developing suitable methods, both for data extraction and data analysis. As shown above, it is not trivial to choose the correct material for comparison in studies where participants are not assigned to different experimental conditions, but where the participants choose themselves when to do what and where. Also, it would be desirable to develop methods of analysis that encompass larger time frames than a few seconds around a SCE, and to increase the exploitation of data on everyday driving, which can provide insight into what characterises crash-free driving.

Arguably we have only scraped the surface of what is achievable with ND studies, which may be more suited to looking at the prevalence of safety-related behaviours such as speed choice and

distraction than at the relationship between some rather arbitrarily defined “events” and the precursors to those events. Perhaps we would also benefit from a more focussed set of studies, examining for example how newly qualified drivers adapt over the first year of driving post-test, in an effort to better understand what constitutes experience and why performance improves so sharply in the first few months of driving.

FOT methods have so far been used mainly to investigate the impact of driver assistance systems. They could and should be applied to investigations of the impact of training regimes of or work-related road safety interventions. Whatever the application, it is vital that robust methodology be applied in study design and analysis.

To conclude, on-road studies make significant contributions to traffic research, particularly in the area of prevalence assessments, long-term studies, pre-crash behaviour, but also for the studies of naturalistic incident-free driving. However, we agree with the National Safety Council that they are not a gold standard (U.S. Department of Transportation, 2013, p.58). The National Safety Council states: “There simply is no perfect study design for an issue as complex as traffic safety.” It is necessary to consider carefully in each case which method is best suited to answer the research question at hand. It may very well be the case, especially for more complex issues like driver distraction, that there is not one single method that will provide all the answers, but that several methods have to be used in a concerted fashion to approach a problem from different angles.

# 5. References


Aarts, L., van Schagen, I., 2006. Driving speed and the risk of road crashes: a review. Accident Analysis and Prevention 38 (2) 215-224.

Ahlstrom, C., Kircher, K., Kircher, A., 2013. A gaze-based driver distraction warning system and its effect on visual behaviour. Transactions on Intelligent Transportation Systems. doi: dx.doi.org/10.1109/TITS.2013.2247759.

Ahlstrom, C., Victor, T.W., Wege, C., Steinmetz, E., 2012. Processing of eye/head-tracking data in large-scale naturalistic driving data sets. IEEE Transactions on Intelligent Transportation Systems 13(2) 553-564.

Akerstedt, T., Hallvig, D., Anund, A., Fors, C., Schwarz, J., Kecklund, G., 2013. Having to stop driving at night because of dangerous sleepiness - awareness, physiology and behaviour. Journal of Sleep Research. doi: 10.1111/jsr.12042

Almqvist, S., Nygård, M. 1997. Dynamisk hastighetsanpassning, demonstrationsförsök med automatisk hastighetsreglering i tätort. (Dynamic speed adaptation — experiment with automatic speed control in built-up areas). Bulletin 154, Department of Traffic Planning and Engineering, Lund University, Sweden.

Almqvist, S., Towliat, M., 1993. Road side information linked to the vehicle for active safety: “Aspen Track”. ARENA, Test Site West Sweden. Swedish National Road Administration, Gothenburg.

Anund, A., Fors , C., Hallvig, D., Åkerstedt, T., Kecklund ,G., 2013. Observer rated sleepiness and real road driving: an explorative study. PLOS One. doi:dx.doi.org/10.1371/journal.pone.0064782

Barr, L.C., Yang, C.Y.D., Hanowski, R.J., Olson, R.L., 2011. An assessment of driver drowsiness, distraction, and performance in a naturalistic setting. Report No. FMCSA-RRR-11-010, Federal Motor Carrier Safety Administration, U.S. Department of Transportation, Washington, D.C.

Biding, T., Lind, G., 2002. Intelligent Speed Adaptation (ISA): results of large-scale trials in Borlänge, Lidköping, Lund and Umeå during the period 1999-2002. Publication 2002:89, Vägverket, Swedish National Road Administration, Borlänge, Sweden.

Blanco, M., Hanowski, R.J., Olson, R.L., Morgan, J.F., Soccolich, S.A., Wu, S.-C., Guo, F., 2011. The impact of driving, non-driving work, and rest breaks on driving performance in commercial motor vehicle operations. Report No. FMCSA-RRR-11-017, Federal Motor Carrier Safety Administration, U.S. Department of Transportation, Washington, D.C.

Brewer, J., Koopmann, J., Najm, W.G., 2011. System capability assessment of cooperative intersection collision avoidance system for violations (CICAS-V). Report DOT-VNTSC-NHTSA-11-08, U.S. Department of Transportation, Research and Innovative Technology Administration, Cambridge, Massachusetts.

Brown, I. D., 1967. Measurement of control skills, vigilance, and performance on a subsidiary task during 12 hours of car driving. Ergonomics 10 (6), 665-673.

Brühning, E., Chaloupka, C., Höfner, K., Lukaschek, H., Michalik, C, Pfafferott, I., Risser, R., Zuzan, W.D., 1989. Sicherheit im Fernreiseverkehr: Ausländische Kraftfahrer - insbesondere Deutsche - in Österreich. Unfall- und Sicherheitsforschung Straßenverkehr, 75.

Burns, P.C., Wilde, G J.S., 1995. Risk taking in male taxi drivers: relationships among personality, observational data and driver records. Personality and Individual Differences 18 (2), 267-278.

Campbell, K.L., Lepofsky, M., Bittner, A., 2003. Detailed planning for research on making a significant improvement in highway safety: study 2 — safety. NCHRP Project 20-58[2]): Contractor's Final Report. Transportation Research Board, Washington, D.C.

Carney C., McGehee D.V., Lee J.D., Reyes, M., Raby, M., 2010. Using an event-triggered video intervention system to expand the supervised learning of newly licensed adolescent drivers. American Journal of Public Health 100 (6), 1101–1106.

Carsten, O., 2002. Multiple perspectives. In: R. Fuller and A.G. Santos (eds.), Human Factors for Highway Engineers. Pergamon, Oxford, UK.

Carsten, O., Fowkes, M., Lai, F., Chorlton, K., Jamson, S., Tate, F., Simpkin, B., 2008. Final report: Intelligent Speed Adaptation project. Institute for Transport Studies, University of Leeds, UK.

Chaloupka , C., Risser R., 1995. Don't wait for accidents — possibilities to assess risk in traffic by applying the Wiener Fahrprobe. Safety Science 19 (2-3), 137-147.

Conner, M., Smith, N., McMillan, B., 2003. Examining normative pressure in the theory of planned behaviour: Impact of gender and passengers on intentions to break the speed limit. Current Psychology 22 (3), 252-263.

De Waard, D., Jessurun, M., Steyvers, R., Raggatt, P., Brookhuis, K., 1995. Effect of road layout and road environment on driving performance, drivers' physiology and road appreciation. Ergonomics 38 (7), 1395-1407.

Dingus, T.A., Klauer, S.G., Neale, V.L., Petersen, A., Lee, S.E., Sudweeks, J., Perez, M.A., Hankey, J., Ramsey, D., Gupta, S., Bucher, C., Doerzaph, Z.R., Jermeland, J., Knipling, R.R., 2006. The 100-car naturalistic driving study: phase ii -results of the 100-car field experiment. Report No. DOT HS 810 593, National Highway Traffic Safety Administration, U.S. Department of Transportation, Washington, D.C.

Dingus, T., McGehee, D., Hulse, M., Jahns, S., Manakkal, N., Mollenbauer, M., Fleischman, R., 1995. TravTek evaluation task C3: camera car study. Report FHWA-RD-94-076, Federal Highway Administration, U.S. Department of Transportation, Washington, D.C.

Duynstee, L., Katteler, H. Martens, G., 2001. Intelligent speed adaptation: selected results of the Dutch practical trial. In: Proceedings of the 8th World Congress on Intelligent Transport Systems, Sydney, Australia, 30 September – 4 October.

Ehrlich, J., Saad, F., Lassarre, S., Romon, S. 2006. Assessment of "LAVIA" speed adaptation systems: experimental design and initial results on system use and speed behaviour. In: Proceedings of 13th ITS World Congress, 8-12 October, London, UK.

Elvik, R., 2011. Effects on accident risk of using mobile phones: Problems of meta-analysis when studies are few and bad. Paper presented at the TRB Annual Meeting, Washington DC.

Elvik, R., Christensen, P., Amundsen, A., 2004. Speed and road accidents: an evaluation of the power model. TOI Research Report 740/2004, Institute of Transport Economics, Oslo, Norway.

ERTICO 2013. http://www.ertico.com/udrive-kick-off-meeting/

Euro NCAP, 2012. Assessment protocol – safety assist. Version 5.6. European New Car Assessment Programme (Euro NCAP), Brussels, Belgium.

Fancher, P., Ervin, R., Sayer, J., Hagan, M., Bogard, S., Bareket, Z., Mefford, M., Haugen, J., 1998. intelligent cruise control field operational test, Final Report, Volume I: Technical Report. University of Michigan Transportation Research Institute, Ann Arbor, Michigan.

FESTA, 2011. FESTA Handbook. Version 4. Revised by FOT-Net. Retrieved from http://www.fot-net.eu/download/festa_handbook_rev4.pdf.

Fleischman, R., 1991. Research and evaluation plans for the TravTek IVHS operational field test. In: VNIS '91: Vehicle Navigation & Information Systems Conference Proceedings, 827-837.

Gilbert, R.K., Underwood, S.E., DeFrain, L.E., 1991. DIRECT: a comparison of alternative driver information systems. In: VNIS '91: Vehicle Navigation & Information Systems Conference Proceedings, 397-406.

Green, P. E., Wada, T., Oberholtzer, J., Green, P. A., Schweitzer, J., Eoh, H., 2007. How do distracted and normal driving differ: an analysis of the ACAS naturalistic driving data. Report No. UMTRI-2006-35, University of Michigan Transportation Research Institute, Ann Arbor, Michigan.

Gstalter, H., 1991. A behaviour and interaction study for assessing safety impacts of a new electronic car equipment. In: Proceedings of the Third ICTCT Workshop, Cracow, Poland.

Hallvig, D., Anund, A., Fors, C., Kecklund, G., G. Karlsson, J.G., Wahde, M., Åkerstedt, T., 2013. Sleepy driving on the real road and in the simulator—a comparison. Accident Analysis and Prevention 50, 44-50.

Hanowski, R. J., Wierwille, W. W., Garness, S. A., and Dingus, T. A. , 2000. Impact of local/short haul operations on driver fatigue: final report. Report No. DOT-MC-00-203, Federal Motor Carrier Safety Administration, U.S. Department of Transportation, Washington, D.C.

Helman, S., Grayson, G.B., 2011. Work-related road safety: a systematic review of the literature on the effectiveness of interventions — Project Report. Institution of Occupational Safety and Health, Leicester, UK.

Hickman, J.S., Hanowski, R.J. , 2010. Evaluating the safety benefits of a low cost driving behavior management system in commercial vehicle operations. Report No. FMCSA-RRR-10-033, Federal Motor Carrier Safety Administration, U.S. Department of Transportation, Washington, D.C.

Hickman, J. S., Hanowski, R.J., 2012. An assessment of commercial motor vehicle driver distraction using naturalistic driving data. Traffic Injury Prevention 13 (6), 612-619.

Hickman, J. S., Hanowski, R. J., Bocanegra, J., 2010. Distraction in commercial trucks and buses: assessing prevalence and risk in conjunction with crashes and near-crashes. Report No. FMCSA-RRR-10-049, Federal Motor Carrier Safety Administration, U.S. Department of Transportation, Washington, D.C.

Hjälmdahl, M., Várhelyi , A., 2004. Speed regulation by in-car active accelerator pedal: effects on driver behaviour. Transportation Research Part F: Traffic Psychology and Behaviour, 7(2) 77-94.

Höfner, K., 1967. Fahrverhalten und Persönlichkeitsbeurteilung junger Mopedfahrer: Eine Gegenüberstellung von wissentlicher und unwissentlicher Beobachtung des Fahrverhaltens. University of Vienna, Vienna, Austria.

Horan, T.A., Hempel, L.C., Bowers, M., 1994. Institutional challenges to the development and deployment of ITS/ATS systems in California. Institute for Applied Social and Policy Research, The Claremont Graduate School, Claremont, California.

Hulbert, S. F., 1957. Drivers' GSRs in traffic. Perceptual and Motor Skills 7, 305-315.

Hurst, P.M., Harte, D., Frith, W.J., 1994. The Grand Rapids dip revisited. Accident Analysis and Prevention 26 (5), 647-654.

ITS America Safety and Human Factors Committee and National Highway Traffic Safety Administration, 1996. Safety evaluation of intelligent transport systems: workshop proceedings. ITS America, Washington, D.C.

Jamson, S.L., 2006. Would those who need ISA use it: investigating the relationship between drivers' speed choice and their use of a voluntary ISA system. Transportation Research Part F: Traffic Psychology and Behaviour 9 (3), 195-206.

Jamson, S.L., Chorlton, K., Gelau, C., Schindhelm, R., Johansson, E., Karlsson, A., Metz, B., Tadei, R., Benmimoun, M., Val, C., Regan, M., Wilschut, E., 2009. Experimental procedures. Deliverable 4.2 of euroFOT, Ford Research & Advanced Engineering Europe, Aachen, Germany.

Jonah, B. A., 1997. Sensation seeking and risky driving: a review and synthesis of the literature. Accident Analysis and Prevention 29 (5), 651-665.

Jovanis, P.P., Aguero-Valverde, J., Wu, K-F., Shankar, V., 2011. Analysis of naturalistic driving event data omitted-variable bias and multilevel modeling approaches. Transportation Research Record 2236: 49–57.

Kircher, K., Patten, C., Ahlström, C., 2011. Mobile telephones and other communication devices and their impact on traffi c safety: a review of the literature. VTI rapport 729A; VTI, Linköping, Sweden.

Klauer, S.G., Dingus, T.A., Neale, V.L., Sudweeks, J.D., Ramsey, D.J., 2006. The impact of driver inattention on near-crash/crash risk: an analysis using the 100-car naturalistic driving study data. Report No. DOT HS 810 594, National Highway Traffic Safety Administration, U.S. Department of Transportation, Washington, D.C.

Klauer, S.G., Guo, F., Sudweeks, J., Dingus, T.A., 2010. An analysis of driver inattention using a case-crossover approach on 100-car data: final report. Report No. DOT HS 811334, National Highway Traffic Safety Administration, U.S. Department of Transportation, Washington, D.C.

Klauer, S. G., Sudweeks, J., Hickman, J. S., Neale, V. L., 2006. How risky is it? An assessment of the relative risk of engaging in potentially unsafe driving behaviors. In A. Foundation (Ed.). Blacksburg, Virginia: Virginia Tech Transportation Institute.

Lahrmann, H., Agerholm, N., Tradisauskas, N., Berthelsen, K.K., Harms, L., 2012. Pay as you speed, ISA with incentives for not speeding: results and interpretation of speed data. Accident Analysis and Prevention 48, 17-28.

Lahrmann, H., Madsen, J.R., Boroch, T., 2001. Intelligent speed adaptation: development of a GPS-based ISA-system and field trial of the system with 24 drivers. In: Proceedings of the 8th World Congress on Intelligent Transport Systems, Sydney, Australia, 30 September – 4 October.

Lai, F., Carsten, O., Tate, F., 2012. How much benefit does Intelligent Speed Adaptation deliver: an analysis of its potential contribution to safety and environment. Accident Analysis and Prevention, 48, 63-72.

Lai, F., Chorlton, K., Carsten, O., 2007. Overall field trial results. Intelligent Speed Adaptation Project. Institute for Transport Studies, University of Leeds, UK.

Lai, F., Hjälmdahl, M., Chorlton, K., Wiklund, M., 2010. The long-term effect of intelligent speed adaptation on driver behaviour. Applied Ergonomics 41, (2), 179–186.

LeBlanc, D., Sayer, J., Winkler, C., Ervin, R., Bogard, S., Devonshire, J. Mefford, M., Hagan, M., Bareket, Z., Goodsell, R., Gordon, T., 2006. Road departure crash warning system field operational test: methodology and results. Report UMTRI-2006-9-1, University of Michigan Transportation Research Institute, Ann Arbor, Michigan.

LeBlanc, D.J., Sivak, M., Bogard, S., 2010. Using naturalistic driving data to assess variations in fuel efficiency among individual drivers. Report No. UMTRI-2010-34, University of Michigan Transportation Research Institute, Ann Arbor, Michigan.

Lee, S.E., Simons-Morton, B.G., Klauer, S.E., Ouimet, M.C., Dingus, T.A., 2011. Naturalistic assessment of novice teenage crash experience. Accident Analysis and Prevention 43 (4), 1472-1479.

McFarland, R.A., Moseley , A.L., 1954. Human factors in highway transport safety. Harvard School of Public Health, Boston, Mass.

McGlade, F., 1963. Testing driver performance-development and validation techniques. Highway Research News, 13-22.

McLaughlin, S., 2010. Ongoing naturalistic driving studies and field operational tests and utility in ADAS development. Presentation at the 17th ITS World Congress, Busan, Korea. Retrieved from http://www.fot-net.eu/download/international_workshops/Busan2010/ss61/vtti_naturalistic_and_fots.pdf

Malta, L., Ljung Aust, M., Faber, F., Metz, B., Saint Pierre, G., Benmimoun, M., Schäfer, R., 2012. Final results: impacts on traffic safety. Deliverable 6.4 of euroFOT, Ford Research & Advanced Engineering Europe, Aachen, Germany.

May, A.D., Bonsall, P.W., Carsten, O.M.J., van Vuren, T., 1988. The evaluation of route guidance systems. Working Paper 266, Institute for Transport Studies, University of Leeds, UK.

Michon, J.A., 1985. A critical view of driver behavior models: what do we know, what should we do? In: L. Evans and R. Schwing (eds.), Human Behavior and Traffic Safety. Plenum Press, New York.

Michon, J.A. , Koutstaal. G.A., 1969. An instrumented car for the study of driver behavior. American Psychologist 24 (3), 297-300.

Neale, V.L., Klauer, S.G., Knipling, R.R., Dingus, T.A., Holbrook, G.T., Petersen, A., 2002. The 100 car naturalistic driving study: phase 1 – experimental design. Report No. DOT HS 809 536, National Highway Traffic Safety Administration, U.S. Department of Transportation, Washington, D.C.

Olson, R.L., Hanowski, R.J., Hickman, J.S., Bocanegra, J., 2009. Driver distraction in commercial vehicle operations. Report No. FMCSA-RRR-09-042, Federal Motor Carrier Safety Administration, U.S. Department of Transportation, Washington, D.C.

Papakostopoulos, V., Panou, M., Bekiaris, E., 2005. Literature review of behavioural effects. Deliverable 1.2.1 of AIDE (Adaptive Integrated Driver Vehicle Interface), INRETS, Paris, France.

Parker, D., Manstead, A. S., Stradling, S. G., Reason, J. T., 1992a. Determinants of intention to commit driving violations. Accident Analysis and Prevention 24 (2), 117-131.

Parker, D., Manstead, A. S., Stradling, S. G., Reason, J. T., Baxter, J. S., 1992b. Intention to commit driving violations: an application of the theory of planned behavior. Journal of Applied Psychology 77 (1), 94-101.

Peng, Y., Boyle, L.N., Hallmark, S.L., 2013. Driver's lane keeping ability with eyes off road: insights from a naturalistic study. Accident Analysis and Prevention 50, 628-634.

Perez, W.A., VanAerde, M., Rakha, H., Robinson, M., 1996. TravTek evaluation safety study. Report FHWA-RD-95-188, Federal Highway Administration, U.S. Department of Transportation, Washington, D.C.

Pettersson, H. E., Aurell, J., Nordmark, S., 2006. Truck behaviour in critical situations and the impact of surprise: a pilot study of a sudden blow-out on the front axle of a heavy truck. In: Proceedings of Driving Simulator Conference Europe.

Quenault , S.W., 1966. Some methods of obtaining information on driver behaviour. RRL Report No. 25, Ministry of Transport, Harmondsworth, UK.

Quenault , S.W., 1967. Driver Behaviour: Safe and Unsafe Drivers. RRL Report LR 70, Ministry of Transport, Crowthorne, UK.

Rathmayer, R. Beilinson, L. Kallio, M. Raitio, J., 1999. The observers' and the visual instruments' effect on driving behaviour when driving in an instrumented vehicle. VTT, Esbo, Finland.

Redelmeier, D. A.,Tibshirani, R. J., 1997. Association between cellular-telephone calls and motor vehicle collisions. New England Journal of Medicine, 336(7), 453-458.

Regan, M.A., Triggs, T., Young, K.L., Tomasevic, N., Mitsopoulos, E., Stephan, K., Tingvall, C., 2006. On-road evaluation of intelligent speed adaptation, following distance warning and seatbelt reminder systems: final results of the TAC SafeCar Project. Report No. 253, Monash University Accident Research Centre, Victoria, Australia.

Rillings, J.H., Lewis, J.W., 1991. TravTek. In: VNIS '91: Vehicle Navigation & Information Systems Conference Proceedings, 729-733.

Rotter, J.B., 1966. Generalized expectancies of internal versus external control of reinforcements. Psychological Monographs 80 (1), 1-28.

Rudin-Brown, C.M., Parker, H.A., 2004. Behavioural adaptation to adaptive cruise control (ACC): Implications for preventive strategies. Transportation Research Part F: Traffic Psychology and Behaviour 7, 59-76.

Rudin-Brown, C.M., Young, K. L., Patten, C., Lenné, M. G., Ceci, R., 2011. Driver distraction in an unusual environment: effects of text messaging in tunnels. In: Proceedings of 2nd International Conference on Driver Distraction and Inattention, Gothenburg, Sweden.

Saad, F. 2006. Some critical issues when studying behavioural adaptations to new driver support systems. Cognition, Technology and Work 8, 175-181.

Saad, F., Hjälmdahl, M., Cañas, J., Alonso, M., Garayo, P., Macchi, , L., Nathan, F., Ojeda, L., Papakostopoulos, V., Panou, M., Bekiaris, E., 2005. Literature review of behavioural effects. Deliverable 1.2.1 of AIDE (Adaptive Integrated Driver Vehicle Interface), INRETS; Paris, France.

Sayer, J.R., Buonarosa, M.L., Bao, S., Bogard, S.E., LeBlanc, D.J., Blankespoor, A.D., Funkhouser , D.S., Winkler, C.B., 2010a. Integrated vehicle-based safety systems light-vehicle field operational test, methodology and results report. Report UMTRI-2010-30, University of Michigan Transportation Research Institute, Ann Arbor, Michigan.

Sayer, J.R., Funkhouser , D.S., Bao, S., Bogard, S.E., LeBlanc, D.J., Blankespoor, A.D., Buonarosa, M.L., Winkler, C.B., 2010b. Integrated vehicle-based safety systems heavy truck field operational test, methodology and results report. Report UMTRI-2010-27, University of Michigan Transportation Research Institute, Ann Arbor, Michigan.

Stutts, J. C., Reinfurt, D. W., Staplin, L., Rodgman, E. A.,2001. The role of driver distraction in traffic crashes. Washington, DC.

Sparmann, J.M., 1989. LISB route guidance and information system: first results of the field trial. In: VNIS '89: Vehicle Navigation & Information Systems Conference Proceedings, 463-466.

Spyropoulou, I., Yannis, , G., John Golias, J., Basacik, D., Chattington, M., Weare, A., Eliou, N., Lemonakis, P., Galanis, T., Karamberopoulos, D., Huth, V., Baldanzini, N.., Val, C., Krishnakumar, R., Regan, M., 2010. Design of a naturalistic riding study — implementation plan. Deliverable 5 of 2BESAFE. National Technical University of Athens, Greece.

Transportation Research Board, 2012. Revised safety research plan: making a significant improvement in highway safety. Strategic Highway Research Program (SHRP) 2, Transportation Research Board, Washington, D.C.

Uchida, N., Kawakoshi, M., Tagawa, T., Mochida, T., 2010. An investigation of factors contributing to major crash types in Japan based on naturalistic driving data. IATSS Research 34, 22-30.

University of Michigan Transportation Research Institute and General Motors, 2005. Automotive collision avoidance field operational test: methodology and results. Report No. DOT HS 809 900, National Highway Traffic Safety Administration, U.S. Department of Transportation, Washington, D.C.

U.S. Department of Transportation, National Highway Traffic Safety Administration, 2013. Visual-manual NHTSA driver distraction guidelines for in-vehicle electronic devices. Docket No. NHTSA-2010-0053.

Wallén Warner, H., Åberg, L., 2008. The long-term effects of an ISA speed-warning device on drivers' speeding behaviour. Transportation Research Part F: Traffic Psychology and Behaviour 11 (2), 96–107.

Wetzel, J.M., Sayer, J.R., Funkhouser, D., 2004. An examination of naturalistic windshield wiper usage. Report No. UMTRI-2004-35, University of Michigan Transport Research Institute, Ann Arbor, Michigan.

Young, R. A., 2011. Driving consistency errors overestimate crash risk from cellular conversation in two case-crossover studies. Paper presented at Driving Assessment 2011: 6th International Driving Symposium on Human Factors in Driver Assessment, Training, and Vehicle Design, Olympic Valley - Lake Tahoe, California.

Young, R. A., 2013. Naturalistic studies of driver distraction: Effects of analysis methods on odds ratios and population attributable risk. Paper presented at Driving Assessment 2013: 7th International Driving Symposium on Human Factors in Driver Assessment, Training and Vehicle Design, Bolton Landing, New York.