Deliverable **D3.2** /

Experimental procedure

Version: 1.0 DRAFT (approval by EC pending)

Dissemination level: PU

Lead contractor: VTT

Version date: 28.02.2019

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 723051.

Document information

Authors

This deliverable originates from the collective contribution of a group of authors participating in all the L3Pilot subprojects.

Coordinator

Aria Etemad
Volkswagen Group Research
Hermann-Münch-Str. 1
38440 Wolfsburg
Germany

Phone: +49-5361-9-13654
Email: aria.etemad@volkswagen.de

Project funding

Horizon 2020
ART-02-2016 – Automation pilots for passenger cars
Contract number 723051
www.L3Pilot.eu

Legal Disclaimer

The information in this document is provided “as is”, and no guarantee or warranty is given that the information is fit for any particular purpose. The consortium members shall have no liability for damages of any kind including, without limitation, direct, special, indirect, or consequential damages that may result from the use of these materials, subject to any liability which is mandatory due to applicable law. Although efforts have been coordinated, the results do not necessarily reflect the opinion of all members of the L3Pilot consortium.

Table of contents

List of figures

List of tables

Glossary

The glossary provides a list of key terms in this deliverable and their definitions based on previous work in the field. Glossary definitions as of 21st December 2018.

Term	Meaning
Automated Driving Function	Activity or purpose of a vehicle to enable automated driving.
Automated Driving System	A combination of hardware and software required to realise an ADF.
Assist	Automated driving function operating at SAE L2.
Baseline	Set of data to which the performance and the effects of the technology under study are compared.
Chauffeur	Automated driving function operating at SAE L3.
Derived Measure	A single measure calculated from a direct measure (e.g. by applying mathematical or statistical operations) or a combination of one or more direct or derived measures (FOT-Net Data, 2016, pp. 55-56).
Direct Measure	A measure logged directly from a sensor, without further manipulations except linear transformations (e.g. m/s to kph) before saving the data to the log file (FOT-Net Data, 2016, p. 55).
Driving Scenario	The abstraction and general description of a driving situation without any specification of the parameters of the driving situation; thus, it summarises a cluster of homogeneous driving situations. Driving scenarios are typically short in time (t < 30 s) and only a few vehicles are involved. An example is lane change to the left lane (AdaptIVe).
Driving Situation	A driving situation is a specific driving manoeuvre (e.g. a lane change with defined parameters). Thus, the driving situation describes in detail a situation that can be simulated and analysed. An example of a driving situation is a lane change at 60.8 km/h with a second vehicle driving at a distance of 10 m behind the host vehicle in the adjacent lane and with a velocity of 65 km/h.
Events	Events are either single time-points or segments of time in time-series data for which one or several criteria are fulfilled. An event can be short (e.g. crash) or long, such as the start of an evasive manoeuvre, car following, overtaking or speeding.
Hypothesis	A specific statement linking a cause to an effect and based on a mechanism linking the two. It is applied to one or more functions and is tested typically with statistical means by analysing specific performance indicators in specific scenarios. A hypothesis is expected to predict the direction of the expected change (FOT-Net Data, 2016, p. 48).
Operational Design Domain	The specific conditions under which a given driving automation system or feature thereof is designed to function, including, but not limited to, driving modes (SAE, 2016).

Term	Meaning
Performance Indicator	Quantitative or qualitative indicator[s], derived from one or several measures, agreed on beforehand, expressed as a mean, percentage, index, rate or other value, which is [are] monitored at regular or irregular intervals and can be compared to one or more criteria. (Mäkinen et al., 2011, p. 45). In some cases will be the same as a derived measure; in other cases further processes are required to generate a PI.
Pilot	Automated driving function operating at SAE L4.
Pilot Test	Field test of applications and functions not as mature as in FOTs. The methodology for testing, however, may in principal be the same. The test is used to decide how and whether to launch a full-scale project.
Research Question	A general question to be answered by compiling and testing related specific hypotheses (FOT-Net Data, 2016, p. 39).
SAE L3 - Conditional Automation	Driving mode-specific performance by an Automated Driving System of all aspects of the dynamic driving task, under specific ODD, with the expectation that the human driver will respond appropriately to a request to intervene (SAE, 2018).
SAE L4 - High Automation	Driving mode-specific performance by an Automated Driving System of all aspects of the dynamic driving task, even if a human driver does not respond appropriately to a request to intervene (SAE, 2018).
Traffic Scenario	Describes a larger traffic context, which includes different (not pre-defined) driving scenarios. Typically, in a traffic scenario a large number of vehicles is analysed over a longer time period. An example could be on a three-lane motorway section with ten highway entrances and exits and a speed limit of 130 kph, for a period of one hour. (AdaptIVe)
Use Cases	A specific event in which a system is expected to behave according to a specified function (FOT-Net Data, 2016, p. 42). This includes the interaction with users, defined as “anyone who uses the road” (CARTRE, 2017, p. 3).
User	A general term referencing the human role in driving automation (SAE, 2016).

List of abbreviations and acronyms

Abbreviation	Meaning
AD	Automated Driving
ADAS	Advanced Driver Assistance Systems
ADF	Automated Driving Function
AV	Automated Vehicle
BL	Baseline
FOT	Field Operational Test
HMI	Human-Machine Interaction
NDS	Naturalistic Driving Study
ODD	Operational Design Domain
OEM	Original Equipment Manufacturer
PI	Performance Indicator
RQ	Research Question
SAE	Society of Automotive Engineers
SAE L3	SAE Level 3
SAE L4	SAE Level 4
SP	Sub-project
TR	Treatment
TRL	Technology Readiness Level
WP	Work Package

1 Introduction

1.1 Background to the L3Pilot Project

Over the years, numerous projects have paved the way for automated driving (AD). Significant progress has been made, but AD is not yet ready for market introduction. However, the technology is rapidly advancing, and today we are at a stage that justifies the pilot testing of automated driving.

Automation is not simply achieved by integrating more and better technology. The implementation of automation and deployment of automated vehicles on our roads needs a focus on understanding driver behaviour, willingness to use, and acceptance of automated driving systems (both from the perspective of the driver and the wider society). User acceptance is one key aspect of the successful deployment of ADFs, in addition to other factors such as understanding the legal challenges and restrictions, which need to be discussed and solved in this context. It is also crucial to investigate the technical feasibility of novel automated driving systems.

L3Pilot is taking important steps towards the introduction of automated cars in daily traffic. The project will undertake large-scale testing and piloting of AD with developed SAE Level 3 (L3) functions (Figure 1.1) exposed to different users including conventional vehicle drivers and Vulnerable Road Users (VRUs), in mixed traffic environments along different road networks (SAE, 2018). Level 4 (L4) functions and connected automation will also be assessed in some cases. It should be noted that an important distinction between Level 2 and Level 3 systems is the shift in supervising responsibility from the human to the AD system (SAE, 2018). With a Level 2 function, the onus is on the human in the driver's seat to constantly supervise the driver support features, and the human is driving even when the feet are off the pedals and (s)he is not steering. With a Level 3 function, the human is not driving when the AD features are engaged but (s)he must drive when the feature requests. This difference means that there is a considerable change in the technical capabilities of a Level 3 automated driving function (ADF) compared to Level 2.

Figure 1.1: SAE Levels of Driving Automation J3016 JUN2018 (Copyright 2018 SAE International).

Extensive on-road testing is vital to ensure sufficient AD function operating performance, to allow an assessment of ongoing user interaction and acceptance of the system. A large and varied sample of test users needs to be involved in this work, to ensure effective piloting, testing and evaluation of ADFs.

L3Pilot will investigate four ADFs performing automated driving tasks in three driving environments: motorway, urban and parking. In the motorway environment, there will be functions capable of performing either high-speed driving or operating in traffic jams, or both. In this project, L3 systems of this type will be termed 'chauffeurs', for example a *Motorway Chauffeur*. L4 systems will be termed 'pilots', for example a *Motorway Pilot.* However, this distinction does not necessarily reflect the publicly-marketed names of the AD functions.

The data collected in these pilots will support the main aims of the project, which are to:

- Lay the foundation for the design of future, user-accepted, L3 (and L4) systems, to ensure their commercial success. This will be achieved by assessing user reactions, experiences and preferences relating to the AD systems' functionalities.
- Enable non-automotive stakeholders, such as authorities and certification bodies, to prepare measures that will support the uptake of AD, including updated regulations for the certification of vehicle functions with a higher degree of automation, as well as incentives for the user.

- Create unified de-facto standardised methods to ensure further development and testing of AD applications (Code of Practice).
- Perform detailed data analysis to show the performance and effects of ADFs in all relevant conditions, in terms of weather, visibility and traffic volumes within current ODD.

The consortium addresses four major technical and scientific objectives listed below:

1. Create a standardised Europe-wide piloting environment for automated driving.
2. Coordinate activities across the piloting community to acquire the required data.
3. Pilot, test and evaluate automated driving functions and connected automation.
4. Innovate and promote AD for wider awareness and market introduction.

The L3Pilot consortium brings together stakeholders from the whole value chain, including original equipment manufacturers (OEMs), suppliers, academic and research institutes, infrastructure operators, governmental agencies, the insurance sector, and user groups. More than 1000 users will test approximately 100 vehicles across Europe.

The work in L3Pilot is structured into different sub-projects that deal with different aspects. An overview is given in Figure 1.2 below:

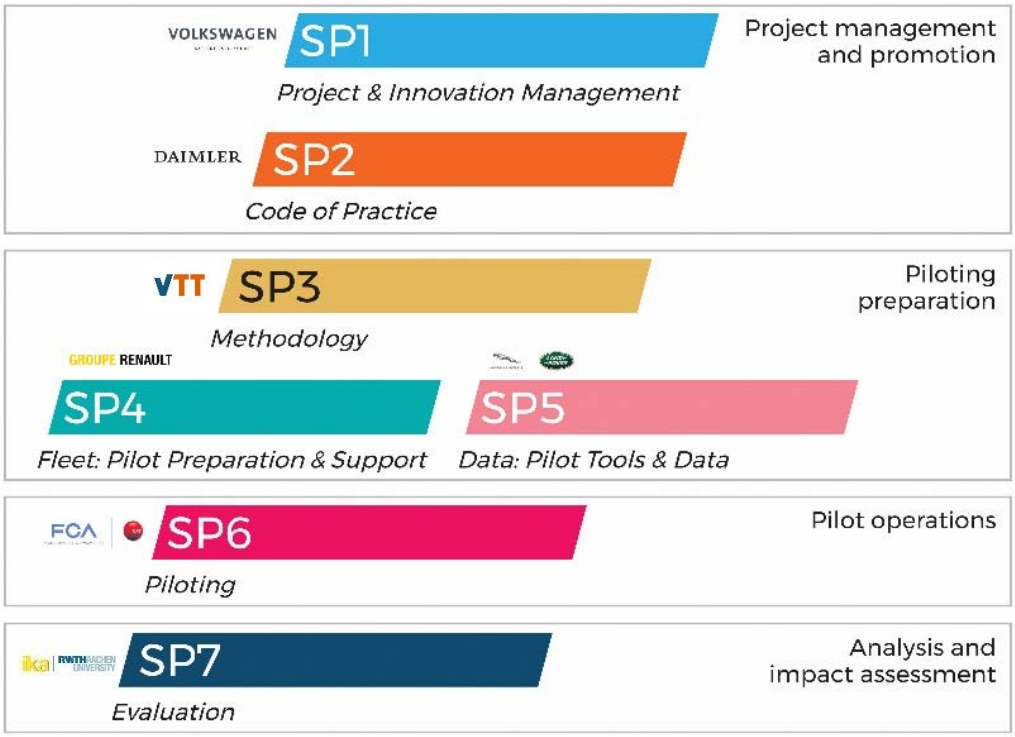

Figure 1.2: Project structure of L3Pilot.

This report focuses on the work of the methodology sub-project, which is closely linked to the work that will be done later in the evaluation sub-project.

1.2 Objectives of the Methodology sub-project in L3Pilot

The objectives of the *Methodology* sub-project (SP3) in L3Pilot are to:

- Develop a methodology for the piloting, testing and evaluation of AD systems for achieving reliable results;
- Reconsider the theoretical background and impact mechanisms required for building a multidisciplinary evaluation methodology;
- Consider not only the expected positive impacts on road and driver safety and traffic flow, but also the unintended, and possibly negative, impacts of AD;
- Facilitate good understanding of a variety of possible effects of AD on the transport system, including the effects on mobility and well-being of people, behavioural adaptation, safety and capacity, fuel consumption and emissions;
- Provide input to a Code of Practice for AD testing, interface design, and investigation of Human Machine Interaction (HMI).

In this context, SP3 provided a list of Research Questions (RQs) as one of its outputs (see D3.1 by Hibberd et al. 2018), meeting the objectives defined above. It is accompanied here by the development of innovative and appropriate experimental procedures to collect the data required to answer these questions, and the development of a structured and robust evaluation plan to ensure that reliable and valid results are achieved from the pilot testing. To accomplish this, L3Pilot follows the *FESTA V-process methodology*.

FESTA (Field opErational teSt supporT Action, 2007-2008) was a project set up to produce comprehensive guidance on the evaluation and delivery of driver-assistance systems and functions using a field operational test (FOT) methodology. The aim of the FESTA project was to provide a structured methodology that would ensure that the systems are appropriately evaluated. This aligns with one of the key objectives of the L3Pilot project, hence the selection of this approach even though L3Pilot is a pilot project, not an FOT.

The FESTA Handbook (FOT-Net 2017) describes a process for evaluating driver assistance systems and functions. The four main pillars of this methodology will be followed in this project. These are: Prepare, Drive, Evaluate, and Address legal and cyber-security aspects. This process has been adapted to suit the needs of L3Pilot, taking into account the fact that the methodology was developed for driver support systems long before the need for testing Level 3 AD functions arose. Therefore, the changes to this process will be documented as recommendations for a Code of Practice for the evaluation of AD functions (see D3.4 Final Evaluation Plan).

The SP3 methodology covers the steps on the left ('PREPARE') of the modified FESTA 'V' (Figure 1.3), laying the foundations and methods for the successful execution of the 'DRIVE' and 'EVALUATE' steps. The work is carried out in close cooperation with other sub-projects.

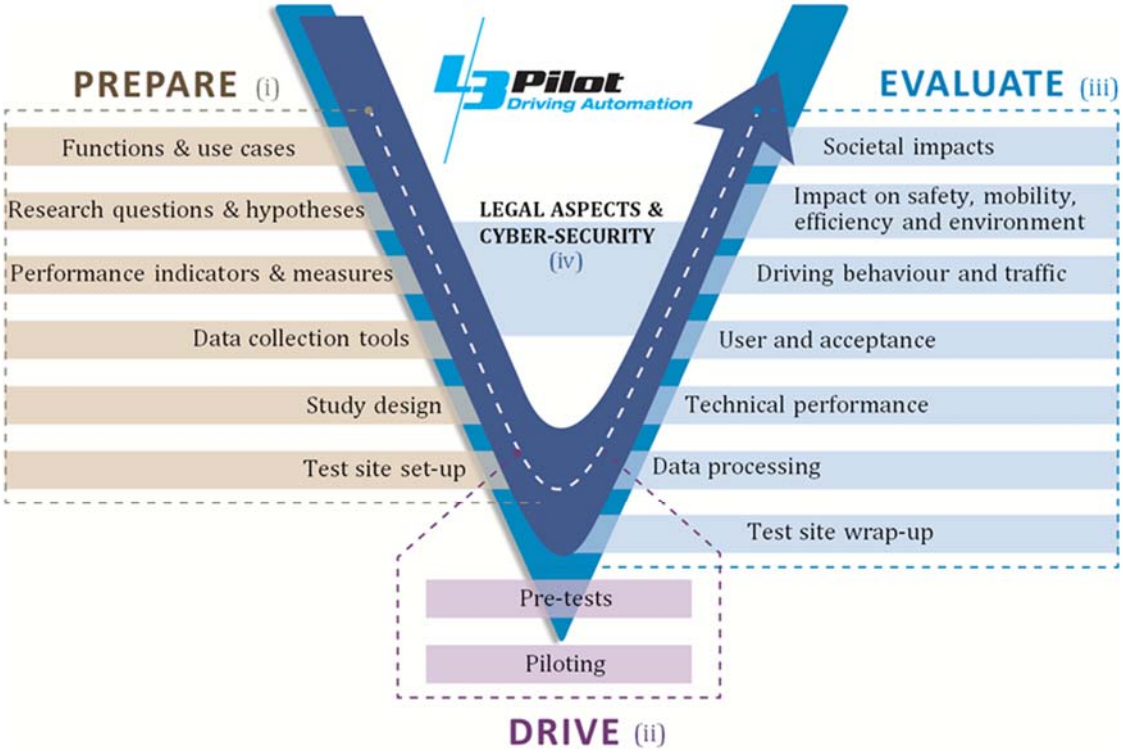

Figure 1.3: L3Pilot methodology overall structure

In L3Pilot, an evaluation of ADFs will be conducted to consider their technical and traffic , user and acceptance aspects resulting from knowledge of, or interaction with, the ADF, and driving and travel behaviour impacts of the ADF, see Figure 1.4.

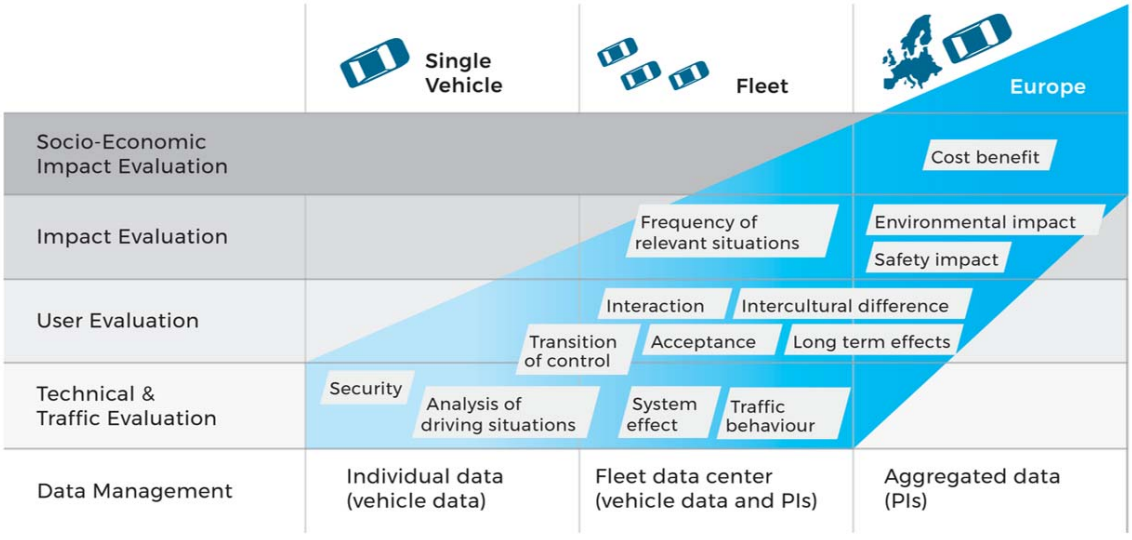

Figure 1.4: Process for evaluation in L3Pilot

The evaluation process will include subjective data collection from study participants and members of the public (e.g. questionnaires) and objective data collection from the pilot vehicles (e.g. vehicle CAN Bus). Based on these data, the evaluation will cover analysis of driving situations and user acceptance and behaviour. Based on these findings, the impact of ADFs in terms of their safety, efficiency and environmental effects will be scaled-up and evaluated. Finally, the socioeconomic impact of the ADFs will be evaluated at EU level in the form of a Cost-Benefit Analysis.

1.3 Introduction to experimental procedures

When designing the experimental procedures for a pilot study, one must understand the difference between FOTs of close-to-market products and pilots of systems on earlier technology-readiness levels. In an AD pilot, satisfactory levels of field tests are controlled tests with a safety driver and OEMs' employees. This procedure is very different to FOTs, where ordinary drivers use the system as part of their daily lives. Thus, in a pilot study, field tests produce indicative estimates of impacts, while further assumptions need to be made on market-ready versions, and their use utilising other sources of information to complement the field measures. In a FOT, one can expect more direct proof of impacts from the field measurements.

The goal of the L3Pilot project is to demonstrate and assess the Level 3 ADFs in real or close-to-real use contexts and environments in the pilots. However, the pilot nature of the field tests will bring some practical limitations to the possibilities of how to conduct them as described above. To receive meaningful and valid results on impacts of the ADFs, it is important to carefully consider the principles underlying the approach to collect the evaluation data. The project's experimental procedure was developed to provide a solid evaluation methodology and to ensure that the results from tests across all pilot sites can lead to an L3Pilot-wide evaluation, taking into account the practical limitations mentioned above. Furthermore, the aim is to harmonise the evaluation criteria by providing detailed requirements for the pilots with the intention to create holistic evaluation results of the L3Pilot project.

The experimental procedure should be based on established scientific methods presented in the literature. Consequently, the general rules and principles found in the literature need to be applied to the specific L3Pilot ADFs, their technology readiness level, and their operational design domains (ODDs) paying specific attention to the safety of experiments made on open roads. Furthermore, the experimental procedure defines the role of each pilot site and facilitates the synchronisation and harmonisation of evaluation across sites. Boundary conditions that set limits to the tests at each pilot site are discussed, and an optimal adaption of the common methodology to practical requirements at the different pilots will be ensured.

This document has two target audiences: (1) research scientists responsible for planning the other parts of methodology and for the evaluation activities, and (2) the persons designing

and running the pilots, both in L3Pilot ----but also in other evaluation projects in the domain of automated driving.

Recommendations related to experimental procedures will also be provided for the 'Code of Practice for Automated Driving' created during L3Pilot for the development of ADFs.

1.4 Purpose of this deliverable

1.4.1 The role of methodology deliverables in L3Pilot

The main purpose of this deliverable D3.2 is to continue the methodology work by describing experimental procedures to be carried out at the pilot sites, and to introduce how to organise the data collection in such a way that an L3Pilot-level evaluation can be conducted across the pilot sites in SP7. This deliverable follows from deliverable D3.1 "From research questions to logging requirements" (Hibberd et. al., 2018), which covered the theoretical basis for the L3Pilot evaluation framework, overall description of the included ADFs, research questions generation process with actual research questions, and logging needs associated with the research questions.

Following this deliverable on the experimental procedure, deliverable D3.3 will present the evaluation methods for all impact areas and each research question in more detail. Finally, D3.4 will provide the overall evaluation plan as a concluding deliverable of the work conducted in the Methodology sub-project.

1.4.2 Content of this deliverable

Chapter 1 sets the scene introducing the L3Pilot project and the methodology work within it. Chapters 2 and 3 determine the general principles of experimental procedure, and how these should be interpreted in the context of automated driving. Specifically, they address what approach (Chapter 2.1) will be applied for data collection in the pilots, i.e. whether the approach will be an experimental study or some type of simulation study. Definition of participants (Chapter 2.2) is closely related to the approach – who the test participants will be, how to generalise the findings for the general public, and the optimal sample size required to ensure sufficient statistical power. The definition of the experimental design (Chapter 2.3) encompasses several topics (within- versus between-participants design; before-after measurements; definition of baselines; variable types) and, in practice, determines the framework for both data collection and analysis. By carefully planning and following the experimental design in evaluation, we can mitigate the effects of random fluctuation or seasonal trends on results. Experimental environments (Chapter 3.1) will be dependent on circumstances at pilot sites. Chapter 4 of this document discusses practical guidance given to the pilot sites and remarks related to experimental procedure plans at pilot sites for the evaluation team. The intention is to support the pilot sites in identifying the critical items of the test setup, and determine how to ensure that the requirements set by the evaluation are met. Finally, this chapter discusses the aspects of the pilot plans that the evaluation teams should consider when conducting analysis.

2 Selection of the approach, test persons and experimental design

2.1 Approaches for data collection

2.1.1 Objective and overview of approaches

The main purpose for selecting the approaches to the study is to adapt the common research methodology to practical requirements and limitations at the different pilot sites. It is important not only to list the approaches and fit them into each pilot site, but also to ensure that the overall combination of various approaches in different pilot sites provides the L3Pilot evaluation with a representative example of impacts of various level 3 automated driving functions of passenger cars. In order to evaluate comprehensively the impacts of level 3 ADFs, methods for both objective and subjective data collection are needed. Methods for objective data collection range from driving simulator studies to naturalistic driving studies (NDS). As shown in Figure 2.1, as the external validity (i.e. transfer of results to the real world) of the results increases, the controllability of the conditions/experiment decreases. Among other factors, this interdependence needs to be considered when designing the procedure of an experiment.

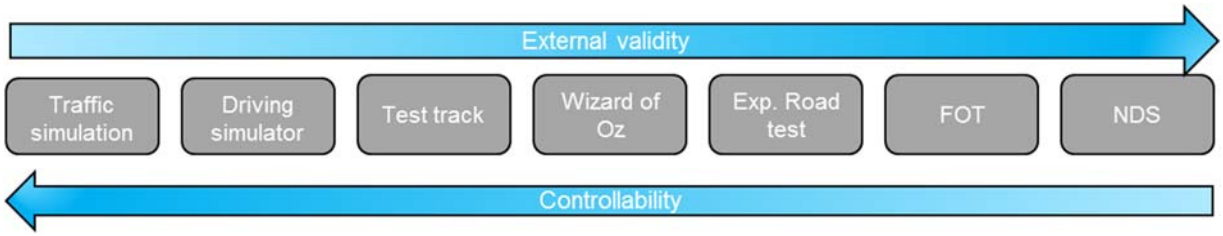

Figure 2.1: Overview of methods for objective data collection

Methods for getting subjective data range from qualitative to quantitative data collection (see Figure 2.2). Qualitative methods are often used for exploratory research to better understand human behaviour in order to answer *why* and *how* questions: reasoning, opinions and motivation. Quantitative methods, on the other hand, are characterised by a systematic empirical investigation of phenomena. Generally, numerical data is generated to enable the quantification of human behaviour.

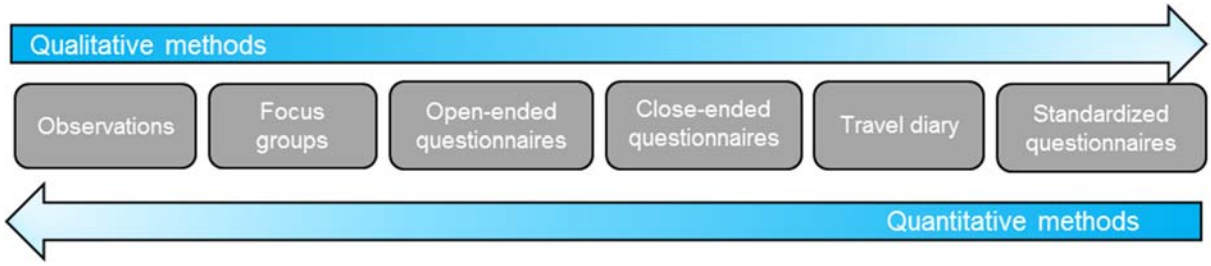

Figure 2.2: Overview of methods for subjective data collection.

2.1.2 Methods for objective data collection

Naturalistic driving study (NDS)

In naturalistic driving studies, participants usually drive an instrumented car (often their own) for a longer period of time on their usual routes without any limiting instructions. NDSs do not follow any experimental control in terms of group assignment or control conditions (i.e. variables are not actively manipulated) and no instructor is present while driving. The data is recorded continuously (Lietz et al., 2011).

Advantages

Participants are not asked to alter their behaviour nor to encounter situations/conditions they would normally not encounter. Therefore, NDS data is very realistic and conclusions on general driving can be drawn. Naturalistic driving studies are characterised by a high external validity. In addition, mobility effects can be studied in NDSs. In NDs it is possible to get deeper information of the behaviour of the subject that in FOTs by monitoring the behaviour in detail in real use context and motives to the behaviour.

Disadvantages

As NDSs are characterised by no experimental control, many factors may influence drivers' behaviour. This means that the internal validity of NDSs is rather low. Replication studies only produce the same results in very few cases. As a high variance of behaviour is observed, a large number of participants and/or high number of kilometres driven are required. The analysis of NDS data requires time-consuming data processing for e.g. specific event detections. Finding and matching the baseline and treatment conditions in the recorded NDS data may also pose a challenge.

Field operational test (FOT)

Field operational tests (FOTs) are defined as "a study undertaken to evaluate a function, or functions, under normal operating conditions in road traffic environments typically encountered by the participants using study design so as to identify real-world effects and benefits" (FESTA Handbook v07). FOTs aim at investigating the effect of one or more independent variables (e.g. introduction of assistant systems, different groups and different conditions) on driving behaviour. In FOTs, baseline and treatment data can be collected. The experimental design allows for limited hypothesis testing (i.e. driving is not as free as in naturalistic driving studies) and manipulation of conditions. Data is collected continuously (Lietz et al., 2011).

Advantages

FOTs offer more experimental control by making causal interferences possible compared to naturalistic driving studies (e.g. driving with system: experimental; driving without system: baseline). They can be designed as both within- and between-participant studies. The external validity is higher than in simulator studies. Conclusions can be drawn on the effects of ADFs in the field.

Disadvantages

On the other hand, FOTs are very time- and cost consuming. Due to the lack of experimental control (e.g. various confounding/intervening variables cannot be fully controlled) and the naturalistic driving, the internal validity (i.e. successfully eliminating confounding variables) is lower compared to lab studies. Testing ADFs in real-world settings most often requires permission from public authorities.

Experimental road tests

Experimental road tests are carried out with instrumented vehicles in real traffic conditions on a predefined test route(s). In order to cover different experimental conditions, participants often have to drive the same test route several times. Generally, a researcher accompanies participants giving instructions and observing behaviours. In case of a prototype vehicle, the presence of a safety driver may also be required.

Advantages

Experimental road tests are suitable for systematically studying various aspects of ADF functionality and ADF behaviour in a realistic environment and in realistic traffic conditions. In experimental road tests, ADFs are tested in predefined sections, appropriate for the research goal at hand, of the road network systems; therefore, compared to FOTs and naturalistic driving studies, the amount of data relevant for the analysis is maximised. In addition, implemented test protocols offer experimental control.

Disadvantages

On the other hand, as is true for a test track study, the effect of ADFs on certain driver aspects cannot be assessed in experimental tests. These aspects include, for instance, what kind of trips and in what conditions the system will be used in real life, changes in travel patterns, etc. Even though the testing area is defined and can be controlled, as well as e.g. testing hours, other factors such as behaviour of other road users cannot be controlled and may introduce some bias. The experiment leader present in the test vehicle may also introduce bias by unintentionally making the subject behave in the way she/he assumes “good” subject behaviour. In addition, permission from the road authority is needed for testing prototypes and ADFs on public roads.

Wizard of Oz

Wizard of Oz is a technique used to give the appearance that an application/system/function is automated, when in fact it is not. One way of simulating the behaviour of an automated vehicle is by using a hidden driver in the backseat or in the front passenger seat. This method allows for evaluating the effects of an imitated ADF on driver behaviour.

Advantages

The Wizard of Oz method is more realistic than other simulation methods in the laboratory due to having real road-users around the test vehicle. This method allows for safely testing drivers’ (adverse) reactions to imitated ADFs in the field. This way, naïve participants can be

assessed in real traffic conditions. Various aspects of ADFs and driver experience can be systematically varied and studied.

Disadvantages

The duration of the experiment is limited due to strains on the hidden driver. This driver needs to be extensively trained to be able to control the vehicle from the backseat, as the first driver input (driver seat) needs to correspond to the automation reaction (hidden driver). As this method is used in real-world driving, the situations can hardly be replicated. Additionally, the hidden driver is still a human with human capabilities when it comes to e.g. reaction times.

Test track studies

In test track studies, cars are driven on specifically designed tracks and not on public roads. Compared to road tests, test track studies take place in a controlled setting and it is possible to evaluate the effects of ADF on drivers' behaviour and perceptions.

Advantages

Test track studies are suitable for systematically studying various aspects of the ADF functionality and the ADF behaviour. Tests are normally done on a closed circuit on private property; therefore, no specific permissions from road or city authorities are needed to test prototype functions. Relevant aspects of the driving environment can be systematically varied (within certain limits). Test track studies offer experimental control through test protocols. In addition, situations inheriting challenges for the driver or the ADF can be tested.

Disadvantages

Compared to public roads, the variation of the driving environment is limited. Not all relevant conditions can be staged in a test track scenario (e.g. traffic jam, variety of road users, etc.). The environment is artificial and may influence the behaviour of (naïve participants) the driver. The effect of the ADF on certain driver aspects, such as travel behaviour, frequency of reduced driver attention or driver state, cannot be assessed in test track studies. In addition, the interaction with other road users may be different and limited.

Driving simulator studies

Driving simulators range from low- and medium- to high-fidelity simulators. They can either be stationary (fixed base) or dynamic simulators. In driving simulator studies, standardised driving tests and scenarios can be implemented. Therefore, comparable and reproducible results are generated. In addition, in driving simulator studies, hazardous/dangerous situations can be tested without harming the participant (Caird & Horrey, 2011).

Advantages

One of the main advantages of simulator studies is the high controllability of the setting, making hypotheses-driven assessment possible, and introducing test subjects with critical conditions and situations not possible in road tests. Influencing factors can be controlled or systematically manipulated. Data acquisition is cheap and easy. The reduced risk of

participants and other road users allows for testing premature systems, critical situations, influencing risk factors (e.g. secondary task engagement). Since the layout of ADF can be varied in a defined way, driving simulators are suitable for systematically studying various aspects of the ADF and driver experience.

Disadvantages

On the other hand, the implemented ADF is not a real system but only a simulated one, meaning that ADF behaviour can only be evaluated to certain extent. The behaviour of the user their acceptance of, trust in, and interaction with the ADF can be evaluated, but keeping in mind that the conditions are not fully natural. In addition, all experienced system boundaries are experimentally implemented; therefore, one needs to be careful when making conclusions on ADF behaviour in real context. In addition, user behaviour in a simulator may differ from real world driving behaviour. Additionally, a minor issue is a selection bias of participants due to simulator sickness.

Traffic simulation

Traffic simulation is a tool that is applied in the impact assessment of various measures. Micro-simulation models simulate the behaviour of individual agents (combination of a driver and a vehicle) in a traffic environment and allow analysis of the consequences resulting from changes to the traffic environment or to drivers' or vehicles' behaviour. Analysis of the effect of the latter aspect typically requires the use of detailed driving behaviour models to determine the actions of each relevant traffic participant.

Traffic simulations can vary from analysing single road stretches or intersections (micro simulation) to simulating traffic in entire towns (macro level simulations). Commercial and open-source software tools are available. Models are usually very flexible, allowing for the assessment of a wide range of different circumstances and conditions. Driving behaviour parameters can be adjusted according to values of the ADF (if known). Simulation can be used, for example, for assessing the effects of ADF on transport system efficiency (e.g. travel times, delays) and the environment (e.g. emissions, fuel consumption). Besides, certain type of simulations are also used to determine the safety impact of certain technology.

Advantages

Traffic simulations offer the opportunity to run analyses for different traffic scenarios (i.e. varying driving behaviour and penetration rates) in an inexpensive way. Since the analysis is done entirely virtually, there is no risk of harming someone physically. In a simulated environment, it is also possible to control the driving conditions and vary them systematically for sensitivity analysis, e.g. to see how much the selected time-gap has an impact on traffic throughput with various penetration rates of the ADFs. In real traffic this is not possible.

Disadvantages

The validity of the outcome depends on the accuracy of the selected input variables and applied models. It would be preferable to run the simulations with several tools and to utilize

versatile supporting data to assess the validity of the results. This, of course, is resource consuming and requires both several types of software and skilled persons to build the models and run the simulations.

2.1.3 Methods for subjective data collection

Observation

Observational approaches collect data by directly observing the behaviour or action of interest, in either naturalistic or controlled settings. This can be achieved using field notes, narrative descriptions, behaviour chronicles, or video recording. Data collection can occur during test drives or during day-to-day mobility.

Advantages

A particular advantage of observational research techniques is that they provide direct evidence on driver behaviour in actual situations. They can also provide explanations for behaviours or outcomes of innovations that cannot be explained by summative measures from conventional research.

Disadvantages

Observational studies inherently produce a large amount of data, which means that a clear analytical strategy needs to be incorporated into the overall research design. Reducing such a large amount of data into a meaningful and useable format is a disadvantage over other data collection techniques. Furthermore, because participants know their behaviour is being observed, they may not act in a natural manner. However, this is not specific to observational approaches.

Focus group

Focus groups are described as “organized group discussions which are focused around a single theme” (Krueger, 1986). Usually guided by a moderator and quite organised and formal, the aim of a focus group is to create an atmosphere where a range of opinions stimulate discussions that will provide a more complete and revealing picture of the theme or issue in focus. Therefore, unlike small group interviews, the goal of focus groups is not to reach consensus or solve a problem, but rather to elicit different opinions, though not to determine their strength or validity.

Advantages

A focus group is a versatile tool that is effective across a range of approaches and research purposes. Focus groups can provide new information on a specific topic in a relatively short period of time, which is enhanced by their emphasis on dynamic group interactions. Moreover, they allow researchers to probe the motivation behind answers. Focus groups are a good approach to finding new development ideas or innovations related to the “product” in question.

Disadvantages

Focus groups have a number of potential limitations. First, the quality of data collected depends first and foremost on how well the discussion is facilitated, which comes down to the skill of the moderator. Second, the large amount of video or audio data collected needs to be transcribed and analysed, which can be time consuming and open to inter-analyst variability. Third, outspoken individuals can dominate the discussion, which may discourage less outspoken individuals from expressing their views. Here, it is the moderator's responsibility to allow for a range of views to be heard. Finally, given that focus group participants are typically self-selected, it may be difficult to extrapolate the results to a wider population.

Open-ended interview questions

An open-ended interview question is designed to encourage a full, meaningful answer using the individual's own knowledge, attitudes, expectations, and/or requirements. This data can be collected either face-to-face or via telephone.

Advantages

Open-ended interview questions allow respondents to include more information, including feelings, attitudes and understanding of the subject. This allows researchers to better access the respondents' true feelings on an issue.

Disadvantages

Open-ended questions can be time intensive in terms of formulating the questions, conducting the interview and analysing the results. This, therefore, often limits the number of interviewees possible. In addition, it is difficult to control for the level of detail or scope of the respondents' answers.

Closed-ended interview questions

Closed-ended interview questions are used to understand respondents' attitudes, expectations and requirements relating to a specific subject or theme. In closed-ended questions, respondents are asked to either give responses that are along a continuum containing an ordered and predefined set of answers, or give a single response to a statement such as "Which system did you prefer?" As with open-ended questions, closed-ended questions can be conducted either face-to-face or via telephone. The selection of response scales needs to be considered carefully, since different scales enable different statistical analysis. This also applies to questionnaires.

Advantages

Closed-ended interview questions can be faster for respondents to answer, which means that more interviews can be conducted. The format of the data collected allows for easy analysis as well as a comparison of answers between respondents.

Disadvantages

Because of the simplicity and limit of the answers, closed-ended questions may not offer the respondents choices that actually reflect their feelings. Closed-ended questions also do not allow the respondent to explain that they do not understand the question or do not have an opinion on the issue and, therefore, it is difficult to gain insight into the motivation behind the answers. Generally, it is preferable to have an alternative “no opinion” or “I don’t know” to avoid the respondent having to reply if (s)he does not have an opinion.

Questionnaire

A questionnaire consists of a standardised series of questions (or other types of prompts) for gathering information from respondents across different pilot sites and tested systems. The questionnaire data can be collected either via mail or, as currently more typical, by internet survey.

Advantages

Compared to other forms of subjective data collection, questionnaires are relatively quick and cheap to administer, which means that more data can be collected in a shorter period of time. The format of the data collected allows for easier and faster analysis.

Disadvantages

Because of the simplicity and limit of the answers, questionnaires may not offer the respondents choices that actually reflect their feelings. Questionnaires also do not allow the respondent to explain that they do not understand the question or that none of the alternatives fully match their opinion. Insights into the motivation behind the answers can be gained only via open answers, but they are labour intensive to analyse. Furthermore, different individuals may adopt slightly different versions of a ‘standardised’ questionnaire, such that comparison of the responses is challenging. A particular case where this might be problematic is in an international project, where translated versions of the same questionnaire are used across different sites. Therefore, attention must be paid to careful alignment of all translations. In addition, the response rate in questionnaires (both mail and internet based) can be quite low, and it is not possibly to know how the large group of non-responding persons differ from those who participate the surveys. This limits the generalization value of the results to the population at large.

Travel Diary

Travel diaries require respondents to write down their daily travel experiences, including trip time and distance, duration, purpose and number of travellers. This approach requires not only regular use of their vehicles or travelling by other modes, but a daily reflection on their travel experiences. The duration for which the travel diary is kept can be varied depending on the purpose of the study. This approach can also be used to provide repeated snapshots of travel behaviour at each stage of a longitudinal research design. The diary can be completed

in the vehicle or at home at the close of each date. Both paper-based and computerised versions, such as mobile data collection applications, can be employed.

Advantages

In terms of resource costs, this method is easier to realise than an observation, as the participant is responsible for the data collection him/herself and only needs to be provided with the materials to complete the diary and guidance to do so by an experimenter. There is also the potential for the participant to add points of note or identify 'outliers' in their dataset, which can aid interpretation by the experimenter. This method has the potential to create highly detailed subjective travel data based on revealed travel behaviour (also called RP, or Revealed Preference, study), with more insight than a Stated Preference SP-type closed-response questionnaire or interview.

Disadvantages

The participant is responsible for remembering to record their data, and furthermore for following the instructions on how to complete the travel diary. This can lead to considerable variation in the frequency and quality of data recording that can influence the analysis, like precision of reported times and travelled distances, absence of data, absence of a trip or absence of recording. Especially the shortest trips are often forgotten if the data is recorded afterwards. In addition, there can be a considerable transcription load for travel diaries completed on paper.

2.1.4 Main approach alternatives for pilot sites per ADF type

Table 2.1 shows the main approaches for testing the prototype functions at the pilot sites and lists the relevant advantages and disadvantages of selected approaches for the different function types. Simulator studies and Wizard of Oz are not included in this table, since the main focus of L3Pilot is on prototype vehicles that are tested in the real world. Nevertheless, these alternative approaches will be used to collect some data as supplementing studies for some research questions. Since the testing of parking functions will be done on private grounds only, there is no fundamental difference between test tracks and controlled testing in real traffic. Therefore, advantages and disadvantages are listed for both approaches.

Table 2.1: Summary of main advantages (+) and disadvantages (–) of approaches relevant for pilot site testing for the ADF-groups.

ADF	Test track	Experimental road tests	FOT / NDS
Traffic jam / Motorway	+ Highly controlled environment + Critical situations can be tested + No specific permissions from road authorities needed - Complex driving scenarios cannot be staged - No realistic interaction with other vehicles and road users	+ Efficient data collection within ODD + Realistic interaction with other vehicles (and road users) - Critical driving situations are too dangerous to be included systematically (not arranged on purpose) - Permissions (may be) needed Important to find comparable situations/circumstances in AD and manual driving	+ Frequency of ODD can be assessed more reliably + Usage of ADF can be assessed more reliably + Realistic interaction with other road users + Mobility impacts and deeper understanding of the motives - Data logging is less efficient (e.g. vehicle is either parked throughout most of the day or used outside the ODD) - Challenging to find comparable situations/circumstances in AD and manual driving
Urban	+ Highly controlled environment + Critical situations can be tested + No specific permissions from city authorities needed - Complex driving scenarios cannot be staged - No realistic interaction with other road users	+ Efficient data collection in ODD + Realistic interaction with other road users - Critical driving situations might be too dangerous to be included systematically (not arranged on purpose) - Permissions (may be) needed Important to find comparable situations/circumstances in AD and manual driving	+ Frequency of ODD can be assessed more reliably + Usage of ADF can be assessed more reliably + Realistic interaction with other road users - Challenging to find comparable situations/circumstances in AD and manual driving - Data logging is less efficient (e.g. vehicle is either parked throughout most of the day or used outside the ODD)
Parking	[Test track and controlled tests are the same if testing on private grounds only.] + Highly controlled environment + Efficient data collection within ODD		+ Frequency of ODD can be assessed more reliably + Usage of ADF can be assessed more reliably - Very inefficient data logging

2.2 Participants

In this chapter, an overview of the participant selection criteria is given with a description and rating of their importance. Recommendations about sample size are provided. In addition, different driver types are compared, and conclusions are drawn to support the test-participant recruiting at pilot sites.

2.2.1 Selection criteria

The selection of participants for an experiment or FOT follows certain criteria depending on the research question and corresponding hypothesis. When a factor, such as age or gender, is not of particular interest for a specific hypothesis, a balanced spread is recommended to avoid effects caused by the participants being biased toward a specific factor. A well-balanced sample allows to generalise the results with more confidence and arguing for effects that are also present in the basic population. When a factor is particularly interesting for a specific hypothesis, then it may be used as a selection criterion. In other words, drivers may be sampled according to this factor. For example, when the usage of automation is of interest for older drivers compared to young drivers, the sample should be recruited accordingly to fit either of the two groups.

Furthermore, the sample size is crucial to find statistical significance in the data with a certain effect size. This is a quantitative measure in statistics for the magnitude of a phenomenon (Kelly & Preacher 2012). The appropriate sample size depends on several experimental design attributes such as the choice of between-participants (two separate groups of drivers one with and one without automation) or within-participants design (each participant drives with and without the ADF) (FOT-Net 2016, PReVENT, 2009).

2.2.2 Demographic variables

Demographic variables are often a basic part of surveys and easily retrievable through questionnaires. Examples of demographic variables include age, gender, occupation, social economic variables and driver impairments. Since there are age- and gender-related factors e.g. in technology affinity (Edison & Geissler, 2003), a balanced sample with regards to these variables is recommended. In practice, this means that the number of young drivers (e.g. 18–25 years) and older drivers (e.g. >60 years) or males and females (in all age groups) should be balanced to represent a larger population.

Socioeconomic variables, such as occupation and average yearly income, are important particularly for the socioeconomic analysis included in the project. This information can be retrieved from questionnaires and eventually leads to a possible future prospect of the spread of automated vehicles when available. This spread is evidently also connected to their potential impact.

Asking for driving impairments is a sensitive topic and might be seen as a violation of privacy. The focus is to test level 3 automation, which requires a driver as fall-back, and the participants need to be fully capable to operate a vehicle. Therefore, it is self-evident that only participants who do not have driving impairments should be recruited. Although higher-level automation has a great potential to provide increased mobility to people with driving impairments, it is not within the scope of this project.

2.2.3 Expectation and experience

Many people have become familiar with automated vehicles from the media. Moreover, participants can have different expectations of the functionality, and even driving experience with level 1 or level 2 automated functions. These expectations play a role when participants interact with a new L3 ADF. Here, an important aspect is that people involved in the development of ADFs have expectations of the functionality that can greatly divert from the expectation of the general public. One could assume that the expectations of the general public are less realistic, i.e. close to L5 automation. A common introduction can help set a common mental model for all participants testing an ADF. However, special training, although potentially necessary for safety reasons, should be well defined and needs to be considered when evaluating the results. Trained drivers are likely to interact differently with an automated vehicle, especially in the transition between automated driving and human control. Training can also influence a driver's opinion of an automated system.

Driving experience is also a component that has shown an influence on driving, particularly in challenging situations. This is commonly represented by drivers' self-reported estimates of their annual mileage driven. In more detail, driving in various conditions and environments should be checked to identify the circumstances in which the experience was gained. Potentially, the accident history could be asked. However, this might violate privacy regulation depending on the participant recruitment pool.

2.2.4 Driver types

Drivers can be categorised based on certain criteria, such as driving experience or age. As described above, different driver attributes should be considered when testing ADFs depending on the research question. Several driver types are introduced below. The optimal choice of driver type depends on the hypothesis and experimental setup. Other factors, such as safety or company-specific requirements, can limit this choice; e.g. prototype vehicles typically require a specialised trained and highly experienced driver, or at least a company employee.

Dimensions of driver categorisation

Essentially, there are two highly relevant driver attributes when investigating the effect of new vehicle systems such as automation: the driving experience and familiarisation with the tested or similar system. The driving experience is an important factor in traffic safety. Safety increases over time with more exposure to the driving task. The driving experience will play a role when a driver assesses an ADF based on his/her performance capabilities.

Familiarisation with a system is gained by instructions or training and exposure as well as usage. Driver categories basically range from naïve participant to expert/test driver or even fully professional driver. Ordinary drivers may be naïve participants when not familiar with the system or even novice drivers with only minimal driving experience. The group of professional drivers are not typically either naïve participants or novice drivers. The two dimensions of the attributes are visualised in Figure 2.3.

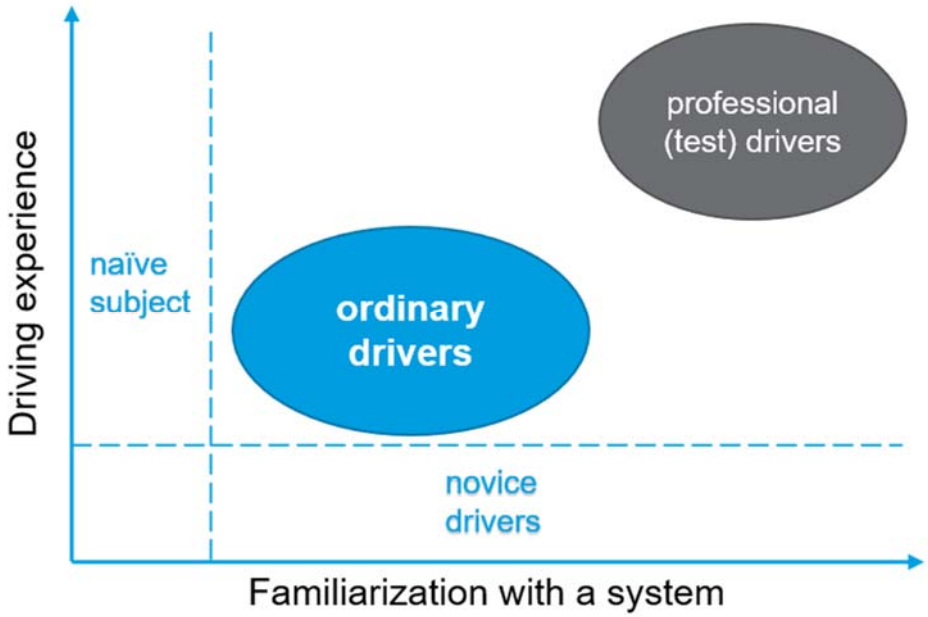

Figure 2.3: Categorisation of driver types in driving experience and system knowledge

Professional (test) drivers / safety drivers

A professional driver is mainly characterised as an individual whose main work activity involves driving. This leads implicitly to a high driving experience over time. An overview of different professional driver types is shown in Figure 2.4. The displayed proportions are based on German data and are roughly taken from publicly available statistics (Bundesagentur für Arbeit, 2018). It serves as an outline of the term generally understood as professional driver.

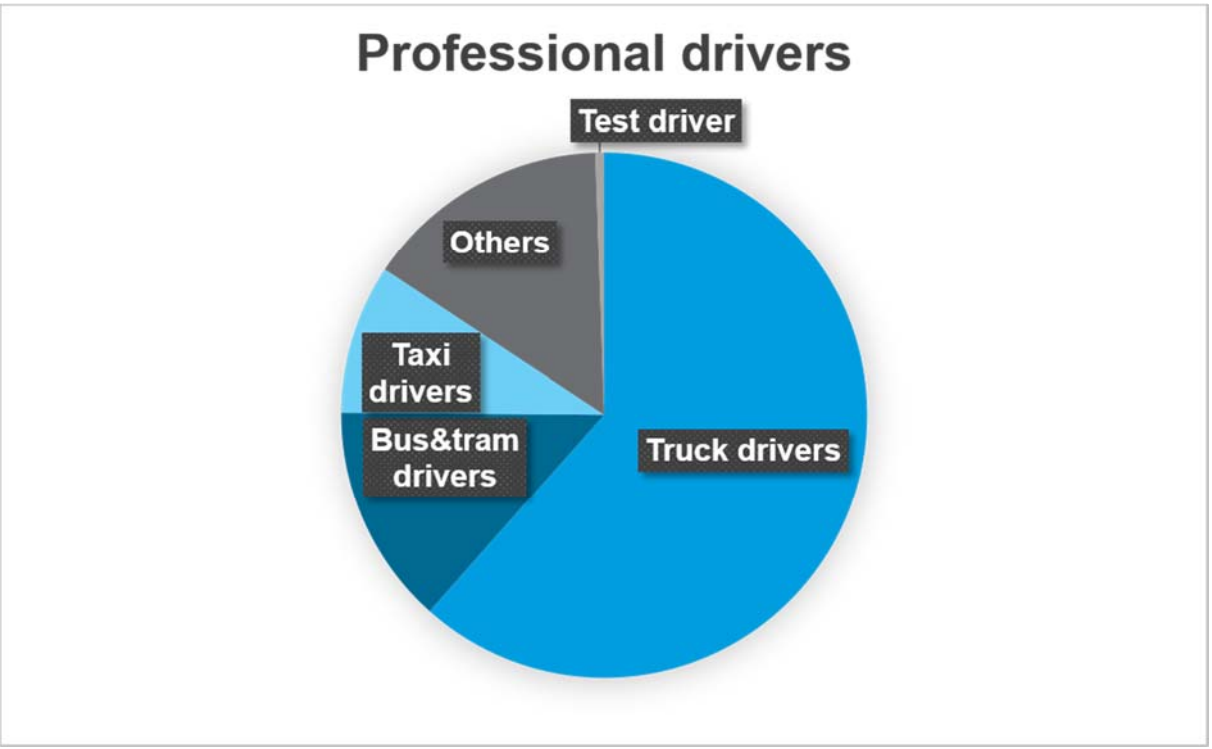

*Figure 2.4: Classification of professional drivers (proportions roughly taken for Germany (Bundesagentur für Arbeit, 2018)*In the context of L3Pilot, the term professional drivers refers to **professional test drivers**. Here, the attribute "professional" means that these drivers are financially compensated for driving. Furthermore, they hold a qualification in driving prototype vehicles and have the ability to "test" certain functions or system attributes through familiarisation and intensive training. Therefore, the following definition is used in this project:

"**Professional (test) drivers** are individuals who drive vehicles as a profession, or as part of their day-to-day work, for remuneration, and have typically extensive driving experience. As part of their training, they have been trained to e.g. handle cars in critical situations. These drivers can be deployed to operate prototype vehicles undergoing road tests."

A professional driver can also act in the same way as a driving instructor, either with duplicate vehicle controls on the passenger side or some mechanism to actively intervene in the driving task, such as an emergency brake button. This driver is usually called a **safety driver** and serves as supervisor and backup in case of critical situations. Such a setup can be mandatory when testing prototype systems in public traffic environments.

While a professional driver can be used for the technical evaluation of an automated driving function, the user-related research questions are focused on the perception of the system from a more general viewpoint and therefore should be addressed mainly with others than professional drivers. In fact, if it is only possible to include professional drivers in an experiment, generalisation of the results to a larger population other than professional drivers would be a challenge.

Another almost similar group of drivers are highly trained company employees, who are also involved in developing and testing the new vehicles and systems. In the tests, they may also act as safety drivers.

Ordinary drivers

A more generalised driver type that covers a wider spectrum of individuals is referred to as ordinary drivers. This type embraces drivers with no specific extreme (low or high) level of driving experience, nor specific training related to the systems or vehicles. For instance, in the L3Pilot description of work, ordinary drivers to be included in the evaluation are characterised as of "age from 20 to 70 years and recent driving experience at least two years, not specially trained". In former discussions, the following definition was proposed:

> "**Ordinary drivers** are individuals who hold a licence granting them permission to drive on public roads, but do not have any additional driving qualifications or permits, such as racing licences, and do not drive or test vehicles as part of their work."

There are two distinctions regarding professional drivers that should be considered. Ordinary drivers do not drive as part of their profession day-to-day and are, therefore, not necessarily highly experienced in the driving task nor trained to test vehicles. They can be naïve regarding the system but will most likely be instructed and familiarised with the ADF to some extent for safety reasons. The subjective evaluation of ordinary drivers is likely reflecting a more general attitude and allows to scale the results to a wider population with higher confidence.

Naïve participants

Considering the novelty of automated driving, most drivers will not yet have experienced an ADF, especially level 3 ADF. In general, a test person not having prior experience of the studied system is considered a naïve participant. It is not expected that the first driving tests of a level 3 automation are performed with completely naïve participants on public roads, due to the risks involved. However, the level of familiarisation is highly relevant when performing subjective evaluation. A harmonisation is desirable (same level of training).

Novice drivers

For the general driving experience, the opposite of a highly experienced driver is a novice driver. This is an individual who recently received their driver's licence and has little driving experience. Novice drivers are known to have a higher crash risk due to lack of ability to assess and handle critical situations. Although the interaction of novice drivers with automation is very interesting, especially the ability to take over the driving task and the influence of automation on gaining driving experience, it is beyond the scope for this project.

Company drivers (non-professionals)

A company driver is an individual recruited internally within the company to perform driving tests or evaluate driving systems occasionally, not as part of their everyday work assignment. Using company drivers in studies and for subjective evaluation is common practice and a reasonable approach considering the easy accessibility of participants.

The required training for company drivers depends on the tested functionality and its technology readiness level and test environment, and on company policy. In some cases, training for AD may require several training courses and tests before a company driver (non-professional driver) is allowed to drive with AD on public roads. These trainings aim at safe testing of the systems in normal traffic.

Two aspects should be considered when working with company drivers: First, they have a different relationship with the product due to their employment. This has an influence on attitude and can bias the subjective evaluation of features depending on the level of identification with the brand. Second, when investigating the handling of features from a customer perspective, it should be ensured that testing systems such as ADAS and ADFs are not only conducted by employees who are involved in the development of, or have a deep understanding of, the system's functional behaviour, as expectation and knowledge about the system's limitations can bias the results.

2.2.5 Selection of driver sample

The driver sample included in the tests will have a significant impact on the data, which will be used for answering research questions on user and acceptance aspects, but also on the technical and traffic evaluation area. Due to safety regulations, vehicle owners have internal requirements about who is allowed to drive the prototype vehicles used in the tests. Pilot plans need to take into consideration these safety requirements, and for the evaluation plan, there is a need to understand the consequences of choices made with respect to the theoretical dimensions described above. For instance, it has to be decided whether a driver who is an employee, has a specific company internal driving licence and works in a department not involved in vehicle development, can be considered an ordinary driver or not. These are further discussed in "Practical guidance to the pilots", chapter 4.2. In addition, a proposal for combining the various driver types for each research question in L3Pilot is given in chapter 4.3.5.

As mentioned, different research questions have different requirements regarding driver type. The same is true for the needed sample size (in terms of number of test participants, driven kilometres and different driving scenarios):

- Especially for research questions related to technical and traffic evaluation, continuous driving with the ADF can be logged with professional test drivers, however, the take-over performance of professional drivers cannot be generalised to the general public.

- Contrary to that, baseline data shall rather reflect the distribution of driving behaviour from a larger, at best representative, driver sample if possible (= between-subjects study) but always within ODD of ADF and in similar circumstances.
- For answering questions on user-related aspects, also driving with ADF should be recorded from a larger, not too homogeneous driver sample.

2.3 Experimental design

2.3.1 Objectives and background

The goal of the experimental design is to enable verification or falsification of the research hypotheses regarding the impacts of AD. The aim is to organise the tests and data collection in such a way that all variables which could bias the results are identified and controlled. The experimental design applied also determines some conditions for the integration of data and results across the pilots.

Four types of variables need to be measured in pilots and differentiated in the analyses (e.g. Shinar, 2017). It is vital that the variable categories are covered and measured during the tests:

- Dependent variables: operationalised as performance indicators (PI) such as distribution of velocity, or frequency of harsh braking and perceived comfort or usefulness, and they are calculated from direct measures from the field tests or by using other means of data collection like questionnaires.
- Independent variables: variables that can be varied systematically, and here they are related to the AD function being in use or available (i.e. driving ADF on or off).
- Control variables: variables related to the driving situation, e.g. road environment and test-participants' type and age. These variables are varied to some degree or kept constant. Furthermore, these background variables can be used to go deeper in the explanation of AD impacts like showing the interactions, e.g. age and use of AD or driving scenario, age and use of AD.
- Confounding variables: variables relevant to describing the circumstances, which cannot be varied systematically but will be part of the data for explanatory purposes, e.g. weather, momentary traffic situation, etc. The identification of the presence of cofounding variables may also be used to judge the quality of the data, and eliminate a part of the data from the analysis in order to reveal the pure impacts of AD.

The framework for L3Pilot needs to take into account all the different variants of the ADFs tested at all the different pilot sites. Furthermore, some of the pilot sites plan to divide their tests into several phases of data collection, which might differ for instance with respect to:

- Participant type,
- Driving environment,

- Test condition-specific instructions, and
- Presence of a safety driver in the vehicle.

It is recommended that the presentation order of different conditions (baseline and treatment) is varied systematically (e.g. half of the test participants baseline-treatment, the other half treatment-baseline). This is also so-called ABBA-design e.g. counterbalancing, and it is used to control the order of presentation of ADF that might introduce systematic bias to the results. By counterbalancing it is not possible to eliminate the bias, but to balance it over the design so that the possible bias is not belonging systematically to only one of the treatment conditions. Furthermore, the objective is to carefully plan the timing and order of different test conditions, which may have consequences for how many test participants will be needed.

2.3.2 Requirements per research question

The advantages and disadvantages of various driver types, as well as the desired approaches to research questions, were discussed in previous chapters (Chapters 2.1 and 2.2). The summary of these, as well as preferences for each research question, are presented at the end of this chapter (Chapter 2.4).

For some research questions, reference data is required with which driving with ADF active will be compared to baseline driving, e.g. manual driving. The experimental design will be adapted according to the research questions defined in the deliverable D3.1. Generally, to assess the research questions regarding use and user acceptance, baseline measurements do not need to be planned separately but the baseline (manual driving as a reference) is part of the framing of questions. To study impacts on vehicle and road user behaviour, baseline data is a must, as it is for part of the technical performance evaluation.

2.3.3 Definition of baseline

With all approaches to data collection, care must be taken that comparable manual driving in the ODD is available for analysis, especially for the technical and traffic evaluation. Here, a direct comparison between baseline driving and driving with the ADF is planned in order to assess the impact of the function on driving behaviour of the vehicle and on the interaction with other road users.

The selection of a baseline determines the reference to which automated driving is compared. Is automated driving going to be compared with totally manual driving, or manual driving with ADAS support, and does the baseline represent the general public or an experienced and skilful test driver? Furthermore, a question to be answered in the baseline selection is also whether the baseline represents the situation as of today, or the situation when the automated vehicles are entering the market. Consequently, the decisions regarding baseline determine the level of conclusions that can be made.

The results of the comparison to baseline will serve as input to the impact assessment by which the potential implications of the tested functions, for instance on safety and efficiency, will be assessed. Therefore, the baseline data has a direct influence on the results regarding

the derived impact of the functions. For example, one performance indicator could be the speed profile in a road section, and the impact on speed would be calculated by subtracting the baseline values (speed profile) from treatment values. The same concerns other similar types of performance indicators. The schematic picture below (Figure 2.5) indicates that the results of the baseline affect the magnitude of the effect equally to the results of the treatment (AD driving).

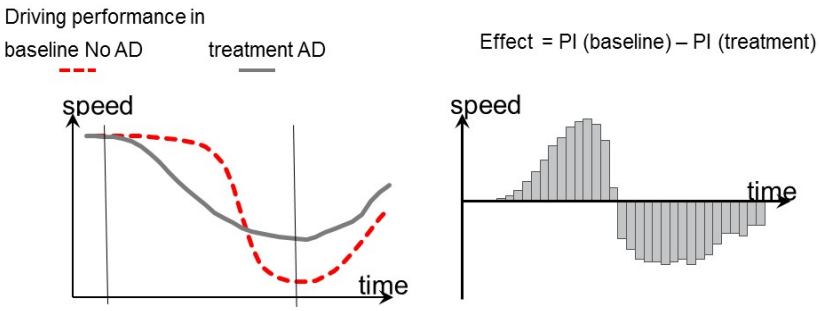

Figure 2.5: Fictive example indicating the role of baseline and treatment (automated driving) data when assessing impacts on speed behaviour of automated driving.

The selection of the baseline from the options listed above as examples is critical for drawing conclusions. Furthermore, it is stressed that everything but the use of ADF (i.e. environment, route, conditions) should be the same in baseline and treatment conditions in this type of comparison to avoid any bias in the results. However, in some cases, specifically in field tests, it may be necessary to compromise this basic rule. This should be considered carefully, motivated well, and made transparent to be able to interpret the results correctly.

To be able to derive reliable conclusions from the comparison of function and baseline driving, the following requirements should be fulfilled:

- The data should contain a sufficient amount of baseline driving covering all relevant driving scenarios within ODD (similar traffic, weather, the same or comparable route etc.). In experimental approaches, it is often the aim to have an equal amount of data for all compared conditions.
- In case baseline data is collected with a group of drivers different to the group used for collecting the treatment data, the data should be generated from a not-too-small sample of ordinary drivers to avoid the results being based on a biased subsample or on a very small number of individual drivers.
- The vehicles in the baseline data collection should be the same or comparable (e.g. car make and model) to the vehicles used for AD data collection, so that a comparable driving style is theoretically feasible for both conditions.
- The logged datasets should be comparable to the data logged with ADF (same or very similar logging system) to ensure that the same data processing and evaluation algorithms and tools can be used for both datasets and that the logging itself does not

cause bias in the results. Note that the data on AD availability (even if not indicated to the user) should be logged also for the baseline.

Besides theoretical issues, practical limitations need to be considered. As an example, it might be the case that the vehicles used in the tests are not available without emergency braking systems or ESP. One option could be that manual driving includes driving with active safety systems that are only active in critical driving situations but other, continuously acting systems like ACC or lane keeping assist are inactive.

Fully manual driving as baseline (L0)

- The difference between fully manual driving and AD is more substantial and effects are easier to show (+)
- May be difficult for safety reasons (especially if not even active safety systems are active) in tests with the general public in particular (needs to be informed carefully and made explicit) (-)
- May be artificial to put test participants in a situation they are no longer used to if they have been driving L1 or L2 on a regular basis (needs to be informed carefully and made explicit) (-)

Manual driving, supported by L1 or L2 as baseline

- Motivated if it is the test drivers' normal driving (+)
- ADAS and active safety systems increase the safety of test participants (+)
- The difference from AD may not as great as with fully manual driving (-)

General public as baseline

- Effects compared to automated driving probably greater than with experienced drivers (+)
- In many cases not an option as only company drivers are allowed to drive the vehicles (-)
- Compensations; insurances? (+/-)

Experienced test drivers as baseline

- Baseline data can be collected with the same fleet as AD (+)
- It is assumed that the effect of AD is smaller than for the general public (-)

2.3.4 Experimental setup

The comparison between driving with ADF active and manual driving can be based on a within- or between-subjects design. A between-subjects design means that the data from the two conditions are logged for two different test participant samples, which may also differ in regard to sample size, driver type etc. A within-subjects design results in collecting data from one sample in which every test participant experiences both conditions (paired data). In case more conditions are of relevance (e.g. different instructions), the number of respective

conditions per test participant can also be larger than two. Given instructions can vary, for instance, regarding indications for usage of the system.

Within-subjects design

- Test participants are the same in baseline and treatment phases and therefore individual features (experience, driving style, gender, age…) are well controlled (+);
- Typically used when sample sizes are relatively small (+);
- Learning or carry-on effects from one test condition to another need to be taken into account (-) and can be at least partly handled by means of counterbalancing.

Between-subjects design

- Data for both conditions can be collected at the same time to control the effects of some circumstantial effects (+);
- The groups may be different regarding the features of individual test participants (-);
- Typically, a larger sample is needed (than in a within-subjects design) to balance the differences between individuals; another option is to select matched pairs into the sample (+/-).

2.3.5 Options for baseline collection

In general, the following two options for baseline collection are preferred:

Option 1:

- Data is collected in an experimental setup with non-professional drivers.
- Here, the baseline (manual driving without ADF active) is collected with test participants driving in the ODD but without having the ADF activated. However, the availability of ADFs needs to be logged in the background (without any indication to the test participant) to make sure that the baseline is collected within the ODD and hence comparable to the treatment data.
- The amount of baseline data collected should be comparable to the amount of data in any of the other conditions. That means that
  - if the experimental setup consists of two conditions (manual driving vs. ADF), baseline data should ideally makes up 50% of collected data;
  - if the experimental setup consists of three conditions (manual driving vs. usage of ADF as much as possible vs. usage of ADF as liked), baseline data makes up 1/3 of collected data;
  - etc.
- This option can be applied in a within- (assuming non-professionals are also driving with the ADF) or between-subjects design.

Option 2:

- Data with manual driving is logged separately from data collection for ADF (e.g. in separate test drives or in a completely different project/test).
- Preferably, non-professional drivers should contribute to the collection of baseline data (manual driving), to ensure that manual driving data reflects as well as possible driving by the general public.
- Driving in ODD should be comparable between both datasets (e.g. same road type, same speed limit, same traffic condition, comparable driving scenarios, if possible similar weather & lighting conditions, same country etc.).
- The sensor setup used in the data analysis should be comparable for both datasets, allowing the analyses of the PIs selected for AD data. Additionally, all signals mandatory for analysis need to be available in both datasets (this especially relates to extra sensors like cameras or sensors measuring rear or side traffic).
- The amount of baseline data needed for the analysis is determined by the number of driving scenarios (selected unit for analysis) required for the analysis.

Figure 2.6: Frequency of different driving manoeuvres per hour of driving, separately for road categories. The figure is based on results reported by Metz et al. (2013) provides some results from a previous project on the frequency of driving scenarios per hour. It needs to be considered that the definition of driving manoeuvres used there is not the same as the one to be used in L3Pilot. Nevertheless, the numbers allow a first estimate of the amount of data needed as baseline.

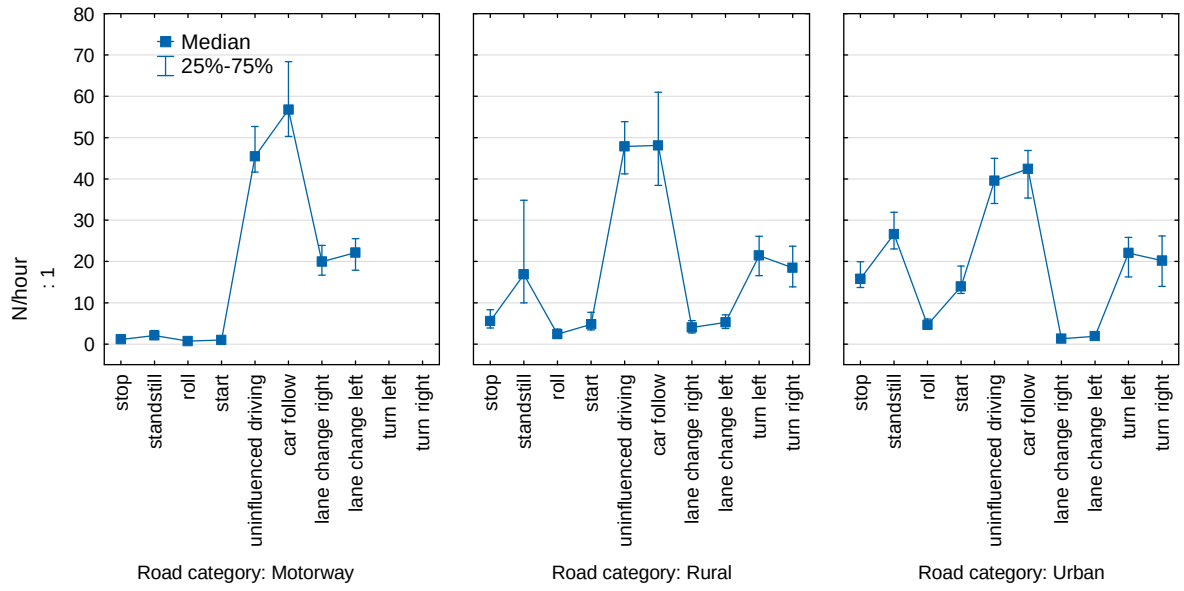

Figure 2.6: Frequency of different driving manoeuvres per hour of driving, separately for road categories. The figure is based on results reported by Metz et al. (2013)

2.3.6 Recommended experimental design

Based on the research questions, and taking into account some of the limitations at the pilot sites, the following experimental designs are recommended in this phase of L3 ADF testing:

- Parking pilot:
  - Controlled study in a closed environment, such as a private parking lot.
  - Within-participants design with two conditions: manual parking vs. usage of ADF.
- Urban pilot:
  - Controlled drives on public roads.
  - Within-participants design or between-participants design with two conditions: manual driving vs. use of ADF.
- Motorway pilot – Design 1:
  - NDS / FOT approach.
  - Within-participants design or between-participants design with two conditions: ADF available vs. ADF not available.
- Motorway pilot – Design 2:
  - Controlled drives on public roads.
  - Within-participants design with three conditions: 1) manual driving, 2) ADF driving, 3) driving where the test participant can decide whether to use ADF or not.

The suggested experimental designs allow answering most of the research questions.

2.4 Recommendations per research question

The recommendations per research question are listed in Table 2.2 and Table 2.3. Research questions in the field of impact evaluation (safety, efficiency, environment and mobility) and socioeconomic impact evaluation will not be answered directly from the data collected in the pilots, but are based on the various results: from research questions related to technical and traffic evaluation, user & acceptance evaluation, and other external data sources. Therefore, they do not have direct requirements on the experimental design and are not considered in these tables. Overall, test track studies are considered to be preferred for parking functions in contrast to motorway and traffic jam pilots and urban pilots. Whenever relevant, the ADF type is indicated separately in the tables. Figure 2.7 shows the approach to developing a harmonised design suited to answering the previously defined research questions.

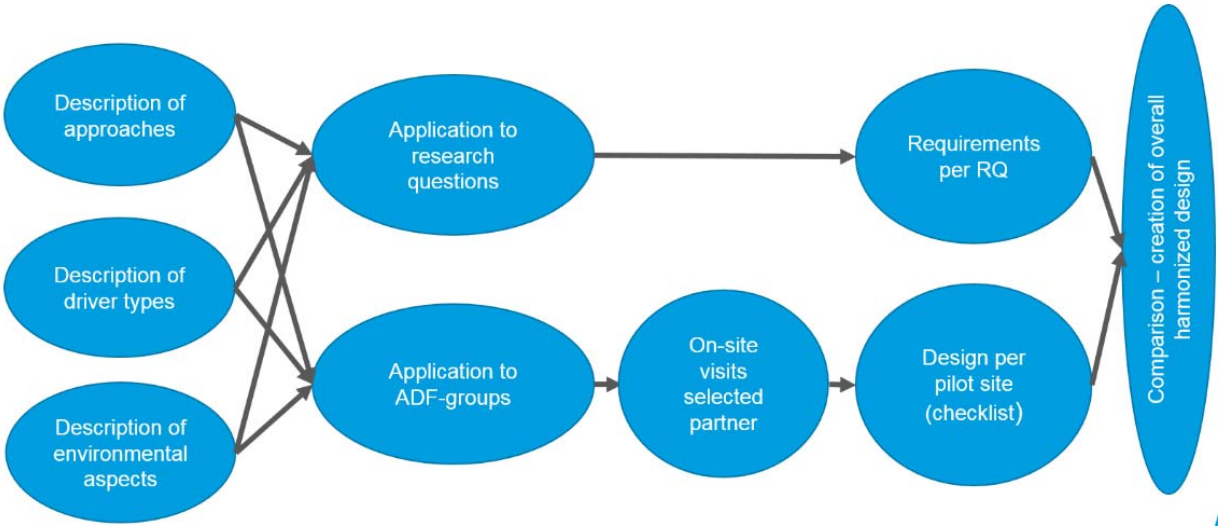

Figure 2.7: Workflow for creating a harmonised design in L3Pilot.

Table 2.2: Recommendations on experimental design per research question – technical and traffic. Abbreviations: (X P) = appropriate approach for parking function, (X) = not desired driver type but useful in the early development phase, BL = required driver type for baseline sample.

RQ Level 1	RQ Level 2	Keyword	Desired participant type					Baseline required	Desired exp. design			Could be addressed by		
			Professional (test) drivers	Company drivers	Ordinary drivers with supervision	Ordinary drivers without supervision	Additional solution: passenger		Driving on test track	Controlled drives on public road	Driving in NDS / FOT	Simulator study	Wizard of Oz	Annual survey
What is the system's technical performance?	How reliable is the system performance in a given driving and traffic scenario?	Reliability of ADF in use cases	X	X	X	X			(X P)	X	X			
	How often and under which circumstances do the ADFs issue a takeover request?	Unexpected takeover requests	(X)	X	X	X			(X P)	X	X			
		Planned take-over requests	X	X	X	X			(X P)	X	X			
What is the impact on own driving behaviour?	Are there any traffic violations while using the ADF?	Traffic violations	(X)	X	X	X, BL		X		X	X			
	How do take-over requests affect driving?	Takeover requests		(X)	X	X			(X)	X	X			
	What is the impact of ADF on vehicle dynamics?	Longitudinal acceleration	X	X	X	X, BL		X	(X P)	X	X			
		Lateral acceleration	X	X	X	X, BL		X	(X P)	X	X			
	What is the impact of ADF on the accuracy of driving?	Precision of manoeuvre	X	X	X	X, BL		X	(X P)	X	X			

RQ Level 1	RQ Level 2	Keyword	Desired participant type					Baseline required	Desired exp. design			Could be addressed by		
			Professional (test) drivers	Company drivers	Ordinary drivers with supervision	Ordinary drivers without supervision	Additional solution: passenger		Driving on test track	Controlled drives on public road	Driving in NDS / FOT	Simulator study	Wizard of Oz	Annual survey
		Vehicle lane position	X	X	X	X, BL		X		X	X			
	What is the impact of ADF on the driven speed?	Speed	X	X	X	X, BL		X	(X P)	X	X			
	What are the impacts of ADF on energy efficiency?	ADF and efficiency	X	X	X	X, BL		X		X	X			
	What is the impact of ADF on the frequency of near-crashes / incidents?	Harsh braking	(X)	(X)	X	X, BL		X	(X P)	X	X			
		Lane departures	(X)	(X)	X	X, BL		X	(X P)	X	X			
	What is the impact of ADF on the frequency of certain events?	Driving manoeuvres	(X)	X	X	X, BL		X	(X P)	X	X			
What is the impact of ADF on the interaction with other road users	What is the impact of ADF on the interaction with other road users in a defined driving scenario?	Distances to other vehicles	X	X	X	X, BL		X	(X P)	X	X			
		Behaviour of surrounding pedestrians	X	X	X	X, BL		X	(X P)	X	X			
		Distance to preceding vehicles	X	X	X	X, BL		X		X	X			
	What are the impacts of ADF on traffic efficiency?	ADF and efficiency	X	X	X	X, BL		X		X	X			

RQ Level 1	RQ Level 2	Keyword	Desired participant type					Baseline required	Desired exp. design			Could be addressed by		
			Professional (test) drivers	Company drivers	Ordinary drivers with supervision	Ordinary drivers without supervision	Additional solution: passenger		Driving on test track	Controlled drives on public road	Driving in NDS / FOT	Simulator study	Wizard of Oz	Annual survey
	What is the impact of ADF on the number of near-crashes / incidents with other road users?	Incidents with other vehicles		(X)	X	X, BL		X	(X P)	X	X			
		Incidents with VRUs		(X)	X	X, BL		X	(X P)	X	X			
What is the impact on the behaviour of other traffic participants?	How does the ADF influence the behaviour of subsequent vehicles?	Behaviour of subsequent vehicles	X	X	X	X, BL		X		X	X			
	How does the ADF influence the behaviour of preceding vehicles?	Vehicles in front of ego-vehicle	X	X	X	X, BL		X		X	X			
			X	X	X	X, BL		X		X	X			
	What is the impact of ADF on the number of near-crashes / Incidents of other traffic participants?	Subsequent vehicles		(X)	X	X, BL		X		X	X			
		Subsequent vehicles		(X)	X	X BL		X		X	X			

Table 2.3: Recommendations on experimental design per research question – user and acceptance. Used abbreviations: (X P) = appropriate approach for parking function, (X) = not desired driver type but useful in the early development phase, BL = required driver type for baseline sample.

RQ Level 1	RQ Level 2	Keyword	Desired participant type					Baseline required	Desired exp. Design			Could be addressed by		
			Professional (test) drivers	Company drivers	Ordinary drivers with supervision	Ordinary drivers without supervision	Additional solution: passenger		Driving on test track	Controlled drives on public road	Driving in NDS / FOT	Simulator study	Wizard of Oz	Annual survey
What is the impact on user acceptance & awareness?	Are drivers willing to use an ADF?	Willingness to use	(X)	X	X	X	(X)		(X P)	X	X	X	X	X
	How much are drivers willing to pay for the ADF?	Willingness to pay		(X)	X	X	(X)		(X P)	X	X	X	X	X
	What is the user acceptance of the ADF?	Perceived safety	(X)	(X)	X	X	(X)		(X P)	X	X	X	X	
		Perceived comfort	(X)	(X)	X	X	(X)		(X P)	X	X	X	X	
		Perceived reliability	(X)	(X)	X	X	(X)		(X P)	X	X	X	X	
		Perceived usefulness	(X)	(X)	X	X	(X)		(X P)	X	X	X	X	
		Perceived trust	(X)	(X)	X	X	(X)		(X P)	X	X	X	X	
		Acceptance and system behaviour in unexpected use cases	(X)	(X)	X				(X P)	X	X	X	X	
	What is the impact of ADF on driver state?	Driver stress	(X)	(X)	(X)	X			(X P)	X		X	X	
		Driver fatigue	(X)	(X)	(X)	X			(X P)	X		X	X	

RQ Level 1	RQ Level 2	Keyword	Desired participant type					Baseline required	Desired exp. Design			Could be addressed by		
			Professional (test) drivers	Company drivers	Ordinary drivers with supervision	Ordinary drivers without supervision	Additional solution: passenger		Driving on test track	Controlled drives on public road	Driving in NDS / FOT	Simulator study	Wizard of Oz	Annual survey
		Driver workload	(X)	(X)	(X)	X			(X P)	X		X	X	
	What is the impact of ADF use on driver awareness?	Driver attention to the road & other road users	(X)	(X)	(X)	X		X	(X P)	X	X	X	X	
		Risk perception/ behaviour	(X)	(X)	(X)	X		X	(X P)	X		X	X	
	What are drivers' expectations regarding system features?	Drivers' expectations	(X)	(X)	X	X			(X P)	X	X	X	X	X
What is the user experience?	What is the drivers' secondary task engagement during ADF use?	Drivers' secondary task engagement	(X)	(X)	(X)	X					X	X	X	X
		Drivers' secondary task engagement	(X)	(X)	(X)	X					X	X	X	X
	How do drivers respond when they are required to retake control? (Reaction time, success of takeover)	Takeover performance	(X)	(X)	X	X			(X)	X	X	X	X	
		Takeover performance	(X)	(X)	X	X			(X)	X	X	X	X	

RQ Level 1	RQ Level 2	Keyword	Desired participant type					Baseline required	Desired exp. Design			Could be addressed by		
			Professional (test) drivers	Company drivers	Ordinary drivers with supervision	Ordinary drivers without supervision	Additional solution: passenger		Driving on test track	Controlled drives on public road	Driving in NDS / FOT	Simulator study	Wizard of Oz	Annual survey
	How often and under what circumstances do drivers choose to activate/deactivate the ADF?	Frequency of activation/ deactivation	(X)	(X)	X	X				X	X	X	X	
	What is the impact of ADF use on motion sickness?	Motion sickness		X	X	X	X		(X)	X	X			
	What is the impact of motion sickness on ADF use?	Motion sickness		X	X	X	X		(X)	X	X			

.

3 Coverage of circumstances and implications at EU level

3.1 Experimental environment

Many factors influence driving behaviour. An important group of factors affecting the results can be labelled as the environment in which the data collection takes place. The experimental environment differs depending on whether the data is collected in a virtual or a real environment, whether it is a study on a test track or on a public road, whether the road type is a motorway or a rural road, whether the weather is sunny or rainy, whether there is high or low traffic volume, etc. For this reason, the environment in which the data is collected needs to be described but also decided in case there is flexibility within feasible alternatives.

The effects of the environment on driving can be substantial. As Hogema (1996) showed, the average speed in the fast lane with dry weather and low traffic volume was about 20 km/h faster than driving in the same lane with the same traffic volume but raining (see Figure 3.1). Effects are also found depending on whether a motorway is lighted or not. The average speed on a lighted motorway is higher than on a dark motorway (see Figure 3.2). The average speed decreases when drivers perform a secondary task. Sometimes the effects can also be unexpected.

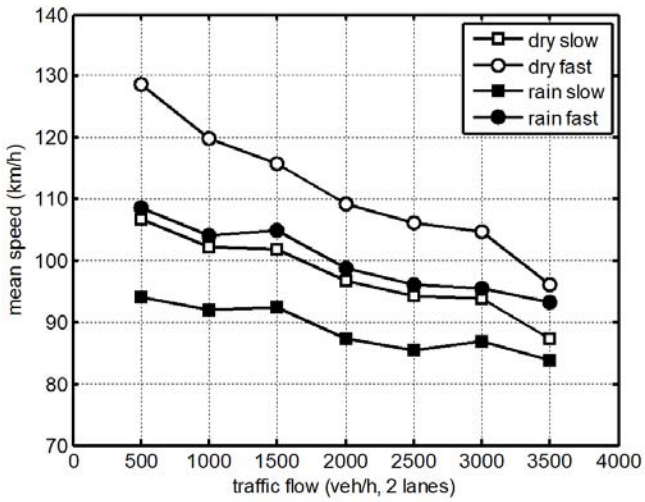

Figure 3.1: Average driving speed on a two-lane motorway under dry and wet conditions and for different traffic intensities (Hogema, 1996).

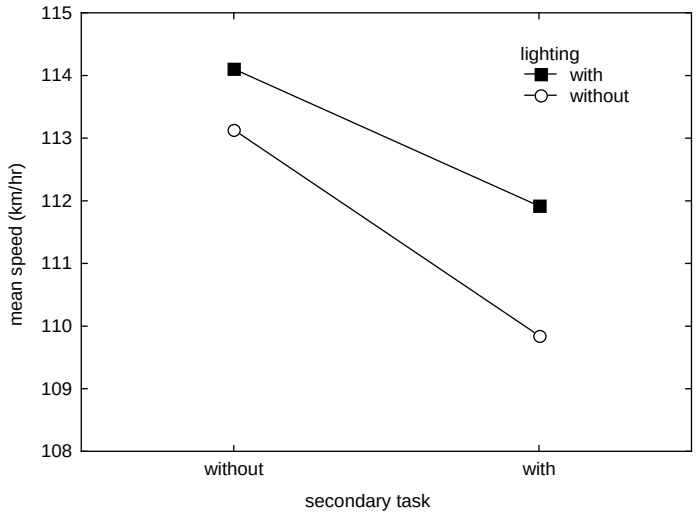

Figure 3.2: Average speed on a motorway with or without lighting and while performing a secondary task or not (Hogema et. al. 2005).

Figure 3.3 shows the average speed of trucks during the day and night for roads with different speed limits. As Figure 3.3 shows, trucks drove faster at night than during the day, which seems to contradict the results of Figure 3.2. However, at night there is also less traffic, which means that trucks can drive for a longer period closer to their speed limit.

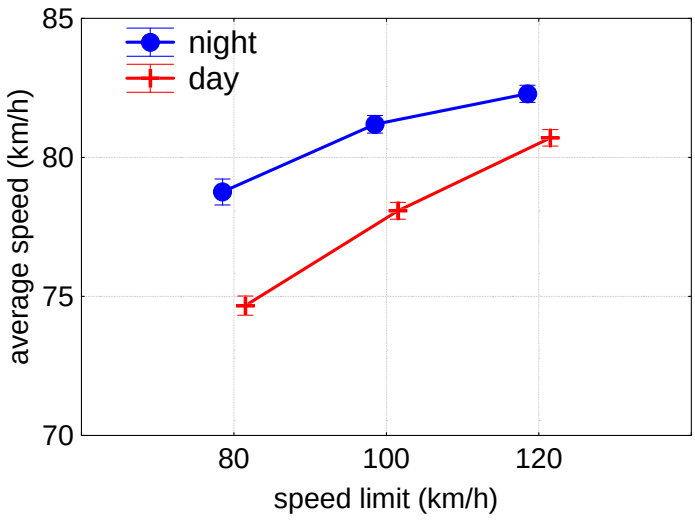

Figure 3.3: Average speed of trucks on roads with different speed limits during day and night (Hogema, 2009).

The above results emphasise that without knowledge of the experimental environment and of the conditions in which the data were collected, it is easy to draw wrong the conclusions. An apparent effect of a system may not be caused by that system, but by the conditions in which the data was collected.

With respect to the experimental environment, different factors can be distinguished (see also: FOT-Net 2017)

- Road type,
- Weather conditions,
- Traffic conditions (intensity and other road users), and
- Geographical location.

The information on a number of these factors can be collected before the pilots start, especially if a fixed test route is used or a driving simulator experiment is performed. Other information needs to be collected during the pilots or afterwards in the analysis phase. If collected after the tests, it must be verified that the required information is then still available.

Table 3.1 lists the information that needs to be collected regardless of whether the approach to data collection is a pilot on a public road, a pilot on a test track, a driving simulator study, etc. This information will be collected as part of metadata and descriptive data for all the pilots.

Table 3.1: Information needed on experimental environments.

Road
Road type
Urban/rural
of lanes
Separation of driving directions/central reservation
Lane width
Type of marking left/right
Speed limit (static and dynamic)
Road surface (dry/wet)
Intersections yes/no
Type of intersection (right of way/traffic lights/zebra crossings/bicycle crossing)
Roadworks (yes/no)
Weather and light conditions
Day/night
Lighting
Visibility (good/reduced visibility)
Precipitation (yes/no)
Wind (normal/strong wind/strong gusts)
Traffic conditions
Other road users (cars/trucks/bicycles/pedestrians etc.)
If bicycles, then location of the bicycles (same road/bicycle path on the road/separate bicycle path)
Traffic density (quiet/normal/rush hour)
Speed of traffic flow (if available from external sources)
Adjacent vehicles
Oncoming vehicles
Passenger (road tests only)
Geographical location
Flat/hilly
Curvy/straight
For simulators only
Moving base (yes/no)
(Full) cabin (yes/no)
Drift (yes/no)

4 Practical guidance for the pilot sites and remarks for the evaluation

4.1 Aim and process of practical support

The experimental procedure synchronises and harmonises the planned tests for different pilot sites to ensure the smooth conduction of harmonised evaluation across all sites. Providing practical support in planning and implementing the experimental procedures at the pilot sites has helped by identifying critical items of the test set-up, and how to ensure that each pilot meets all the requirements set by the evaluation without compromising rules and regulations related to testing AD.

To provide the needed support for the pilots, physical support visits were made to every pilot site by the responsible partner of the experimental procedure. The visits also involved the responsible evaluation partner and the pilot site responsible people. The visits were planned for discussion on, and review of, pilot test plans. Feedback was given to the preliminary plans, and several points critical for the success of piloting and evaluation were checked during the visits. This procedure has shown to be vital for the success of field tests in earlier FOTs, and was also considered valuable for the success of L3Pilot. Provision of the D3.2 and visits to the pilot sites are the main contributions to the process of pilot site support. However, to ensure the continuation of the support throughout the pilot planning and execution, the cooperation between experimental procedure responsible partner and pilot sites will continue, and the information needed on the sites will be complemented based on follow-ups (emails and telcos) and identification of needs of the pilots in the coming phases of L3pilot.

The following chapter (Chapter 4.2) summarises the recommendations to the pilot sites (identified from *A to X*). It also provides remarks regarding experimental procedures from the practical implementations of the tests for the evaluation team (Chapter 4.3).

4.2 Recommendations for the pilot sites

4.2.1 Test participants

Use the best alternative class of test participants allowed by your company rules, country legislation, ethical aspects and other limiting factors.

A. *The preference of test participants is in the order:*

   1) *externals (general public or some specific user/customer group),*

   2) *employees with no or little additional training on driving and no prior knowledge of tested ADFs,*

   3) *highly trained or professional safety drivers.*

B. *If externals or employees of the OEM (without specific training), however, cannot be allowed to drive the test vehicle, it is recommended that they participate in the study by joining the test rides as a passenger and by filling the user questionnaires, based on indirect user experience (being on board and seeing ADFs in use).*

Drivers without extensive additional training are preferred especially for driving the baseline over highly trained or professional drivers, as they better represent average drivers in traffic.

For the selection of the test participants (except professional safety drivers), use the following selection criteria:

C. *All test participants should drive regularly (in their daily life).*

D. *Demographic factors should reflect the driver population of the future customer population (depending on evaluation scopes). Therefore, balance between female and male participants should be taken care of. Include all age groups, also young (<25) and old (60+) drivers if possible. The samples should preferably be balanced even by selecting both male and female participants in all three age groups.*

Regarding the number of test participants needed, Janssen & al. (2004) recommend a sample of at least 26 participants if the study has even a single between-participants variable (comparison of group means), while for a within-participants study 18 participants would be a minimum. These estimates assumed a medium size effect, and moderate probabilities for both detecting an effect (power 80%) and the significance level (0.05). Based on these figures, and with the intention to cover several user groups representing the general public, the following is recommended:

E. *The sample sizes would be close to 100 participants or preferably more per site.*

It is acknowledged that the number of professional drivers is going to be small. Therefore, in addition to the number of test participants, the amount of the data is going to be assessed based on km's and estimates of frequency of the relevant driving scenarios in the route.

4.2.2 Planning of tests

Technology Readiness Level (TRL) of the ADFs, company rules, national legislation and ethical considerations (such as traffic safety) set the limits to the tests. The tests are divided into baseline data collection (driving without AD) and treatment data collection (AD in use within ODD). A part of the treatment data collection can also be the participant's choice (AD available within ODD, drivers can choose when to activate it) after test participants have had a period to drive with ADF activated. All instructions on the ADF usage given to the test participant need to be recorded to the metadata of the tests.

F. *If the aim is to find out whether the driver is willing to use the ADF (participant's own choice), the instructions of the usage should be kept neutral in this respect.*

The test routes are selected and planned based on the functionalities of the ADF and respective ODD. Critical for the route/location selection is how often the targeted driving scenarios occur on the route, and how to guarantee a sufficient amount of comparable data for all of them.

G. *For highway chauffeur and urban chauffeur, it is recommended that the routes are relatively long to have more realistic user experience (What is "long" is left open. It depends on the environment but should preferably be assessed together with the selected partner.)*

H. *Parallel to ADF calibration and in the pre-piloting an initial estimate of frequency of the driving scenarios should be made (together with the selected partner).*

(Too) high environment complexity should be avoided (increases variance of the data) if not specifically targeted at some pilot site.

I. *A practical approach is to check with the local road operator the status of potential construction works on the planned test route, and avoid locations with planned road works when selecting the test route(s).*

Whether the baseline data needs to be collected in same or in similar roads varies by type of AD and environment.

J. *In urban environments the baseline data should be from the same test route as the AD data.*

K. *In motorway environments the baseline data should be from the same or similar environment (in terms of number of lanes, speed limit, proportion of heavy traffic, density of intersections).*

L. *For parking, the baseline and AD data should be collected in the same place.*

M. *For all environments, the baseline data should be collected in similar traffic as the AD data. In addition, weather and lighting conditions should be as similar as possible.*

In practice, the last point usually means baseline and treatment data collection during same/comparable days of the week and hours of the day.

N. *The variation in the traffic conditions should be checked before starting the actual tests (e.g. during pre-piloting).*

Baseline data can be collected with the same test vehicle or with another similar vehicle (without the ADF) and with the same logging system. This collection of baseline data needs to be planned separately — the periods when driving outside ODD do not qualify as good baseline data.

O. *The use of ADAS systems in the baseline data must be carefully considered and noted to the metadata.*

It is remarked that also for the baseline data, a log of availability of AD within ODD is needed as background (in theory even if not activated) to ensure the comparability of the baseline and ADF drives. In practice, this may limit use of other than the automated vehicles in baseline data collection.

P. *The order of baseline and treatment data collection should vary across test participants (BL-TR / TR-BL) if feasible.*

This minimises the impacts of systematic differences happening over time between the two phases of data collection influencing the results. The same concerns any other test conditions sensitive to presentation order, such as changing the roles of two drivers in test rides.

In principle, the best approach would be an equal amount of data for the baseline and treatment. In practice, the treatment data can dominate the data collection, because part of

the intended treatment ride may also turn out to be driving without AD activation. The test participant, however, must be aware whether AD is available, and an unsuccessful treatment condition is in most cases not valid for a baseline. Treatment data can also dominate in studies, where user experience during AD is the main focus. Nevertheless, the baseline data should be sufficient to avoid biases.

> Q. *Design the use of resources carefully to guarantee a sufficient baseline data (regarding both representativeness and amount) to study the research questions intended to covered.*

A complete metadata is vital for data analyses.

> R. *Plan the during-the-test metadata collection and updating procedure for the test conduction plan so that the recordings reflect the actual tests, not just plans. Record the test conduction plan as part of the metadata.*

Validate the test plans by carefully implementing all steps in the pre-piloting phase including data collection and creation of the metadata.

> S. *Provide a small set of complete data (including video data, metadata and possibly external data sources) to the selected partner for pre-piloting of the analyses.*

In case of a second person (safety driver/observer) on board, it is suggested to design a tool (sheet/app) to record observations / tag driving scenarios together with the selected partner.

4.2.3 Performing the tests

Performing the tests can be divided into five phases (excluding the preparation and wrap-up phases of the test site):

1. Pre-test user questionnaire: Can be completed in the recruitment phase. The same questionnaire is used independently of the participant being a driver or passenger of the test vehicle.
2. Instructions: The test participant is given basic information on the test vehicle, ADF and its use (including ODD) and his/her task during the test drives/rides. The basics of the project aims are introduced too.

T. *In spite of professional test driver group, all test participants should act (imagine the situation) as they would have purchased a new automated vehicle and would use the AV on their own trip. The test participants are advised to act as normally as possible, as they would do in that hypothetical situation - in spite that they are of course aware they are participating in a study. Following from this all unnecessary interaction with the test driver during the tests should be minimized even if there is another person in the vehicle.*

U. *The test participants are aware they will be interviewed regarding the user experience. No specific attention should be shown to measuring the driver behaviour with a logging system. However, in case of questions, one should answer truthfully*.

3. Familiarisation with the vehicle: In case the test participant is the driver of the vehicle, they need to be given sufficient opportunity to familiarise with the vehicle and systems before the data collection starts. The user interface and automated functions will be explained before the pilot drive, as well as other conditions (role of the safety driver, other people on board etc.) This familiarisation phase should not be used as baseline.

4. Test drives: Data collection for both the baseline and the treatment.

V. *The task of the test participant is to drive the route as indicated by the navigator. As soon as the vehicle indicates that automated driving is available, s/he can accept the automated driving mode. In case the automated driving option is not utilized by the test driver, they can be encouraged to use it.*

W. *In case secondary tasks are allowed during automated driving the test driver should be informed about this option explicitly. However, a neutral approach is recommended (not tempt or encourage too much).*

5. Post-test user questionnaire: This is filled in right after the last test ride. The same questionnaire is used independently of the participant being a driver or passenger of the test vehicle. Those test drivers who drive for significantly long periods are recommended to fill in the questionnaire periodically.

4.3 Remarks of the pilot plans for the evaluation

4.3.1 Approaches for data collection

Nearly all pilots for motorway, traffic jam and urban ADFs will include experimental tests on public roads. Furthermore, in most pilots, the roads are pre-defined, and several drives will be conducted on the same roads. For closer-to-market functions, it is possible to include more variability on the driving routes than for the earlier prototypes.

All pilots will include some kind of baseline. A few will also have a full experimental design including baseline, ADF drive and one additional drive where the driver can select whether to use the ADF or not.

In most of the pilots, there is a safety driver on board the vehicle. The safety driver acts as a primary driver of the vehicle (in the driver seat) or as a back-up/secondary driver in the passenger seat. Back-up safety drivers may have different controls, varying from the full set of controls, including steering wheel on the passenger side, to the “AD-switch-off button”. The (secondary) safety driver may also act as a test operator, checking that all the systems related to AD and logging are functioning properly.

In a few pilots, the test operator may also operate the additional data logger, designed to log the environmental aspects, such as weather and traffic situation, and possibly also some interesting events during the drive with a dedicated interface integrated into the general data logging system.

Non-driving related tasks are not allowed in most cases. In a few cases, this is due to national legislation, but in others it is due to e.g. specific company policies or insurances. Hence, the experimental design does not include details related to secondary tasks.

4.3.2 Participants

The participant type in the pilots depends highly on at least three aspects:

- the readiness of the function (prototype or closer to market introduction),
- internal company policy (if externals or even internals without specific training are allowed to drive the AD test vehicles with AD on),
- and country-specific legislation related to testing of level 3 ADFs on public roads.

Most of the pilots will start with a well-trained or even professional driver being the responsible (safety) driver of the vehicle, either in the driver seat or in the passenger seat with access to some level of double controls to either switch off the system or even take over control of the vehicle if needed. Later, as the development proceeds, the number of participants with less L3 ADF training is expected to increase.

At least the following participant categories will be included in the pilots:

- fully professionals (driving as a profession; possibly also involved in the AD-development);
- non-professionals (company employees) but with (several) training courses & detailed knowledge of the ADF;
- non-professionals (company employees) with little or no extra licence requirements; ordinary (external) drivers (general public).
- In addition, a few pilots have plans to include passengers as pilot test participants, when the professional or trained driver is driving the vehicle.

When analysing the feasibility of the research questions, the participant type, as well as taxonomy (which ADFs are similar enough), have been considered.

4.3.3 Test environments

The test environments for motorway and traffic-jam functions are always dual carriageway roads with a proper separation of the driving directions. The number of lanes in each direction is mainly three, but varies between two and four. In addition, there may be entrance or exit lanes at different intervals.

Speed limits on the motorways (for motorway chauffeur testing) vary between 70 km/h up to 130 km/h. Additionally, there may be lane-specific or even variable speed limits on the routes. The speed limits are recognised by the vehicles' systems. When analysing the impacts of ADF on driving speed, it is important to compare the driving scenarios in similar speed areas and traffic volume for baseline and ADF drives. If the average speed of the traffic flow is higher than the speed limit, it would also be important information to record this from external sources if possible.

Urban environments include driving at typical urban speeds, variability of road users including vulnerable road users, different intersection types, changing lanes, and turning to the right and left. The speed limit in the tests varies from 30 km/h up to 50 km/h. The driving speed can be much lower than this, depending on the traffic situation.

Parking will be tested on private grounds, at least at the beginning of the pilot period.

All the pilots include driving in good weather, and also in slight rain. Heavy rain, flooding, heavy snowfall and icy roads are excluded. A few pilots will also collect driving data in darkness. Roadwork areas and toll stations are outside the current ODD for all pilots. In addition, both motorway and traffic-jam chauffeurs are available only after the driver has manually entered the motorway and other requirements for the ODD are met. Merging (from a ramp) is excluded.

4.3.4 ADFs included in the pilots from the users' perspective

All piloted ADFs have similar logic for activation: often an indicator (e.g. light) to indicate that the function is available and a switch/switches to the driver to activate the ADF. To take over, the driver can either add torque to the steering wheel, brake, accelerate or use the switch(es) to turn the function off.

The taxonomy groups the ADFs based on their functionality and use. This grouping is presented in more detail in L3Pilot deliverable D4.1. The taxonomy groups the ADFs that are similar enough to be handled as "one ADF" when analysing user acceptance and technical and traffic-related research questions. The objective is to have enough data (e.g. at least two pilots) for each research question / ADF combination to allow assessment of this.

4.3.5 Feasibility of research questions from the experimental procedures viewpoint

Chapter 2.4 listed the desired participant type and approach for each technical, traffic and user-acceptance related research question. This analysis was revisited when the actual pilot plans, including available driver types, were clarified for each pilot site during support visits.

It should be noted that the results present the status as confirmed during the test site visits. It is possible that the design of the pilots will change by the start of the testing activities.

Table 4.1 shows an example of the number of pilots per driver type for each research question for the traffic jam function, and Table 4.2 for motorway chauffeur, taking into account the available driver type, and proposing which driver/test participant groups may be combined for each research question.

Note that for many research questions it is suggested that the non-trained group be analysed separately; hence, there is only one pilot collecting information from this type of drivers. One possible solution could be to analyse these user & acceptance research questions together with non-trained drivers who experienced the motorway pilot.

Table 4.3 lists the user-acceptance research questions in a similar way and proposes the driver groups to be combined for the traffic jam pilot and Table 4.4 for the motorway pilot. Since there are fewer pilots planned for urban or parking ADFs, extensive analysis of the research question has not yet been conducted.

Table 4.1: Number of pilots with different driver type for technical and traffic related research questions from an experimental procedure point of view for traffic jam ADF.

RQ Level 1	RQ Level 2	Keyword	# of pilots with the driver type			Baseline needed	Comments
			Professional (test) drivers	Trained company drivers	Ordinary drivers with supervision		
What is the system's technical performance?	How reliable is system performance in a given driving and traffic scenario?	Reliability of ADF in Use cases	4	2	1		No need to separate the user groups.
	How often and under which circumstances does the ADF issue a takeover request?	Planned takeover requests	4	2	1		No need to separate the user groups.
What is the impact on own driving behaviour?	What is the impact of ADF on vehicle dynamics?	Longitudinal acceleration	4	2	1	x	No need to separate the user groups.
	What is the impact of ADF on vehicle dynamics?	Lateral acceleration	4	2	1	x	No need to separate the user groups.
	What is the impact of ADF on the accuracy of driving?	Precision of manoeuvre	4	2	1	x	No need to separate the user groups.
		Vehicle lane position	4	2	1	x	No need to separate the user groups.
	What is the impact of ADF on the driven speed?	Speed	4	2	1	x	No need to separate the user groups.
	(What are the impacts of ADF on energy efficiency?)	(ADF and Efficiency)					Feasibility depends on other issues than participant type.
	What is the impact of ADF on the frequency of near-crashes / incidents?	Harsh braking	4	2	1	x	All pilots have a safety driver to prevent critical situations. Needs to be taken into account in analysis.

RQ Level 1	RQ Level 2	Keyword	# of pilots with the driver type			Baseline needed	Comments
			Professional (test) drivers	Trained company drivers	Ordinary drivers with supervision		
	What is the impact of ADF on the frequency of near-crashes / incidents?	Lane departures	4	2	1	x	All pilots have a safety driver to prevent critical situations. Needs to be taken into account in analysis.
	What is the impact of ADF on the frequency of certain events?	Driving manoeuvres	4	2	1	x	No need to separate the user groups.
What is the impact of ADF on the interaction with other road users?	What is the impact of ADF on the interaction with other road users in a defined driving scenario?	Distance to other vehicles	4	2	1	x	No need to separate the user groups. Feasibility depends on whether this can be measured.
		Distance to preceding vehicles	4	2	1	x	No need to separate the user groups.
	What is the impact of ADF on the number of near-crashes / incidents with other road users?	Incidents with other vehicles	4	2	1	x	All pilots have a safety driver to prevent critical situations. Needs to be taken into account in analysis.
		Incidents with VRUs	-	-	-		Not applicable to traffic jam function.
What is the impact on the behaviour of other traffic participants?	How does the ADF influence the behaviour of subsequent vehicles?	Behaviour of subsequent vehicles	4	2	1	x	No need to separate the user groups. Feasibility depends on whether this can be measured.
	How does the ADF influence the behaviour of preceding vehicles?	Vehicles in front of the ego-vehicle	4	2	1	x	No need to separate the user groups.
	What is the impact of ADF in near-crash incidents on other traffic participants?	Subsequent vehicles (harsh braking)	4	2	1	x	No need to separate the user groups. Feasibility depends on whether this can be measured.
		Subsequent vehicles (small distances)	4	2	1	x	No need to separate the user groups. Feasibility depends on whether this can be measured.

Table 4.2: Number of pilots with different driver types for technical and traffic related research questions from the experimental procedure point of view for motorway pilot.

RQ Level 1	RQ Level 2	Keyword	# of pilots with the driver type					Baseline needed	Comments
			Professional (test) drivers	Trained company drivers	Company employees (non-trained) with a safety driver	External driver with a safety driver	Company employee (non-trained) without safety driver		
What is the system's technical performance ?	How reliable is system performance in a given driving and traffic scenario?	Reliability of ADF in Use cases	3	3	1	1	1		No need to separate the user groups.
	How often and under which circumstances does the ADF issue a takeover request?	Planned takeover requests	3	3	1	1	1		No need to separate the user groups.
What is the impact on own driving behaviour?	What is the impact of ADF on vehicle dynamics?	Longitudinal acceleration	3	3	1	1	1	x	No need to separate the user groups.
	What is the impact of ADF on vehicle dynamics?	Lateral acceleration	3	3	1	1	1	x	No need to separate the user groups.
	What is the impact of ADF on the accuracy of driving?	Precision of manoeuvre	3	3	1	1	1	x	No need to separate the user groups.
		Vehicle lane position	3	3	1	1	1	x	No need to separate the user groups.
	What is the impact of ADF on the driven speed?	Speed	3	3	1	1	1	x	No need to separate the user groups.
	(What are the impacts of ADF on energy efficiency?)	(ADF and Efficiency)	3	3	1	1	1		Feasibility depends on other issues than participant type.

RQ Level 1	RQ Level 2	Keyword	# of pilots with the driver type					Baseline needed	Comments
			Professional (test) drivers	Trained company drivers	Company employees (non-trained) with a safety driver	External driver with a safety driver	Company employee (non-trained) without safety driver		
	What is the impact of ADF on the frequency of near-crashes / incidents?	Harsh braking	3	3	1	1	1	x	All pilots have a safety driver to prevent critical situations. Needs to be taken into account in analysis.
	What is the impact of ADF on the frequency of near-crashes / incidents?	Lane departures	3	3	1	1	1	x	All pilots have a safety driver to prevent critical situations. Needs to be taken into account in analysis.
	What is the impact of ADF on the frequency of certain events?	Driving manoeuvres	3	3	1	1	1	x	No need to separate the user groups.
What is the impact of ADF on interactions with other road users	What is the impact of ADF on the interaction with other road users in a defined driving scenario?	Distance to other vehicles	3	3	1	1	1	x	No need to separate the user groups. Feasibility depends on whether this can be measured.
		Distance to preceding vehicles	3	3	1	1	1	x	No need to separate the user groups.
	What is the impact of ADF on the number of near-crashes / incidents with other road users	Incidents with other vehicles	3	3	1	1	1	x	All pilots have a safety driver to prevent critical situations. Needs to be taken into account in analysis.
		Incidents with VRUs	3	3	1	1	1		Not applicable for motorway.
What is the impact on the behaviour of	How does the ADF influence the behaviour of subsequent vehicles	Behaviour of subsequent vehicles	3	3	1	-	1	x	No need to separate the user groups. Feasibility depends on if this can be measured. If a company car following the ego-vehicle as a safety car, then this RQ excluded.

RQ Level 1	RQ Level 2	Keyword	# of pilots with the driver type					Baseline needed	Comments
			Professional (test) drivers	Trained company drivers	Company employees (non-trained) with a safety driver	External driver with a safety driver	Company employee (non-trained) without safety driver		
other traffic participants?	How does the ADF influence the behaviour of preceding vehicles?	Vehicles in front of the ego-vehicle	3	3	1	1	1	x	No need to separate the user groups.
	What is the impact of ADF on near crashes/incidents of other traffic participants?	Subsequent vehicles (harsh braking)	3	3	1	-	1	x	No need to separate the user groups. Feasibility depends on whether this can be measured. If a company car following the ego-vehicle as a safety car, then this RQ excluded.
		Subsequent vehicles (small distances)	3	3	1	-	1	x	No need to separate the user groups. Feasibility depends on whether this can be measured. If a company car following the ego-vehicle as a safety car, then this RQ excluded.

Table 4.3: Number of pilots with different driver type for user and acceptance related research questions from an experimental procedure point of view for traffic jam pilot.

RQ Level 1	RQ Level 2	Keyword	# of pilots with the driver type			Comments
			Professional (test) drivers	Trained company driver	Company employee (not trained) with a safety driver	
What is the impact on user acceptance & awareness?	Are drivers willing to use an ADF?	Willingness to use	4	2	1	Interpretation by driver group essential.
	How much are drivers willing to pay for the ADF?	Willingness to pay				Willingness to pay is mostly interesting from the externals - and hence is proposed to be excluded from the RQs for company employees.
	What is the user acceptance of the ADF?	Perceived safety	4	2	1	Interpretation by driver group essential.
		Perceived comfort	4	2	1	Interpretation by driver group essential.
		Perceived usefulness	4	2	1	Interpretation by driver group essential.
		Perceived trust	4	2	1	Interpretation by driver group essential.
	What is the impact of ADF on driver state?	Driver stress	4	2	1	Interpretation by driver group essential. In this RQ profs and trained separately.
		Driver fatigue	4	2	1	Interpretation by driver group essential. In this RQ profs and trained separately.
		Driver workload	4	2	1	Interpretation by driver group essential. In this RQ profs and trained separately.
	What is the impact of ADF use on driver awareness?	Driver attention to the road & other road users	4	2	1	Interpretation by driver group essential.
		Risk perception/behaviour	4	2	1	Interpretation by driver group essential.
	What are drivers' expectations regarding system features?	Drivers' expectations	4	2	1	Interpretation by driver group essential. Expectations depend highly on previous knowledge level of the system features.

RQ Level 1	RQ Level 2	Keyword	# of pilots with the driver type			Comments
			Professional (test) drivers	Trained company driver	Company employee (not trained) with a safety driver	
What is the user experience?	What is drivers' secondary task engagement during ADF use?	Drivers' secondary task engagement			1	Can be analysed only if secondary tasks are allowed overall.
		Drivers' secondary task engagement			1	Can be analysed only if secondary tasks are allowed overall.
	How do drivers respond when required to retake control? (reaction time, success of takeover)	Takeover performance	4	2	1	Interpretation by driver group essential.
		Takeover performance	4	2	1	Interpretation by driver group essential.
	How often and under which circumstances do drivers choose to activate/deactivate the ADF?	Frequency of activation/deactivation				Feasibility depends on instructions given to the driver. If they have an opportunity to select themselves, then this is a feasible RQ. If feasible, then interpretation by driver group is essential.
	What is the impact of ADF use on motion sickness?	Motion sickness		2	1	Motion sickness not expected for professional drivers.
	What is the impact of motion sickness on ADF use?	Motion sickness		2	1	Motion sickness not expected for professional drivers.

Table 4.4: Number of pilots with different driver type for user and acceptance related research questions from an experimental procedure point of view for motorway pilot.

RQ Level 1	RQ Level 2	Keyword	# of pilots with the driver type						Comments
			Professional (test) drivers	Trained company drivers	Company employees (non-trained) with a safety driver	External driver with a safety driver	Company employee (non-trained) without safety driver	Passengers as test participants	
What is the impact on user acceptance & awareness?	Are drivers willing to use an ADF?	Willingness to use	3	3	1	1	1	x	Interpretation by driver group essential.
	How much are drivers willing to pay for the ADF?	Willingness to pay				1		x	Rather to be asked only from externals (main source of information annual survey).
	What is the user acceptance of the ADF?	Perceived safety	3	3	1	1	1	x	Interpretation by driver group essential.
		Perceived comfort	3	3	1	1	1	x	Interpretation by driver group essential.
		Perceived usefulness	3	3	1	1	1	x	Interpretation by driver group essential.
		Perceived trust	3	3	1	1	1	x	Interpretation by driver group essential.
	What is the impact of ADF on driver state?	Driver stress	3	3	1	1	1		Interpretation by driver group essential.
		Driver fatigue	3	3	1	1	1		Interpretation by driver group essential, for professionals' fatigue RQ only if driving longer periods of time.

RQ Level 1	RQ Level 2	Keyword	# of pilots with the driver type						Comments
			Professional (test) drivers	Trained company drivers	Company employees (non-trained) with a safety driver	External driver with a safety driver	Company employee (non-trained) without safety driver	Passengers as test participants	
		Driver workload	3	3	1	1	1		Interpretation by driver group essential.
	What is the impact of ADF use on driver awareness?	Driver attention to the road & other road users	3	3	1	1	1		Interpretation by driver group essential. Feasibility depends on whether there is e.g. eye tracking.
		Risk perception/behaviour	3	3	1	1	1	x	Interpretation by driver group essential. Also take into account whether there is an additional safety driver.
	What are drivers' expectations regarding system features?	Drivers' expectations	3	3	1*	1	1*	x	Interpretation by driver group essential. Overall, company internals and externals may have different level of knowledge. Hence, 1* could be combined.
What is the user experience?	What is drivers' secondary task engagement during ADF use?	Drivers' secondary task engagement			1	1			To be checked which pilots allow secondary tasks overall.
		Drivers' secondary task engagement			1	1			To be checked which pilots allow secondary tasks overall.
	How do drivers respond when required to retake control? (Reaction time, success of takeover)	Takeover performance, reaction time	3	3	1*	1	1*		Expected to see different takeover performances for different driver groups. 1* could be combined.
		Takeover performance, success or takeover	3	3	1*	1	1*		Expected to see different takeover performances for different driver groups. 1* could be combined.

RQ Level 1	RQ Level 2	Keyword	# of pilots with the driver type						Comments
			Professional (test) drivers	Trained company drivers	Company employees (non-trained) with a safety driver	External driver with a safety driver	Company employee (non-trained) without safety driver	Passengers as test participants	
	How often and under which circumstances do drivers choose to activate/deactivate the ADF?	Frequency of activation/deactivation	3	3	1	1	1		Feasibility of this RQ depends highly on instructions to drivers. Could be best assessed if the drivers have a drive with voluntary AD usage.
	What is the impact of ADF use on motion sickness?	Motion sickness		3	1	1	1	x	Professional drivers not expected to face motion sickness.
	What is the impact of motion sickness on ADF use?	Motion sickness		3	1	1	1	x	Professional drivers not expected to face motion sickness.

5 Summary and outlook

The report has two main focuses. On one hand it discusses different aspects of the experimental procedure from the scientific or more theoretical perspective. On the other hand it brings the methodological principles addressed into AD pilot context and provides practical guidance for the pilots. The report consists of chapters including overall recommendations for approaches to data collection, participants for the field tests, experimental design, and experimental environments. Additionally, one chapter (4.2) is dedicated to recommendations for the pilots and remarks of the pilot plans for the evaluation in L3Pilot.

The theoretical work for this deliverable was allocated to the methodology experts of several partners in the project. Finally everything was summarised, and several internal commenting rounds were organised to achieve a common view of this challenging task of creating the experimental procedure for ADF pilot project. In addition, a crucial part of the work defining the recommendations for the pilots was the visits to the pilot sites conducted in the second half of 2018. The input from the test sites and their practical possibilities and limitations for on- the-road testing is included in this report.

The report continues the earlier methodology work started in L3Pilot project and presented in Hibberd et al. [2] and Innamaa et al. [3], which covered the theoretical basis for the L3Pilot evaluation framework, research questions generation process with actual research questions, and logging needs associated to the research questions.

After finalising this deliverable, methodology work continues in L3Pilot with the detailed definition of evaluation methods (D3.3). The feasibility of research questions from the viewpoint of data availability will be added then. Additionally, practical support for the pilot sites during implementation of the experimental procedures will continue with the selected partners, with dedicated pilot site-specific notes on experimental procedures tailored for them. The final deliverable for the methodology (D3.4 ‘Evaluation plan’) will include any needed updates on experimental procedure principles and recommendations. In addition, the detailed pilot execution plan will be reported in deliverable D6.1.

6 Conclusions

The experimental procedure is determined for each research question formulated in the beginning of the project. All critical decisions of the experimental procedure, such as choosing the optimal group of test participants, assessing the need for a separate baseline data collection phase, selecting the test environments, are dependent of and vary by research questions and research hypotheses formulated in the project. Furthermore, individual pilots have somewhat different roles with respect to which research questions they are suitable for and want to focus on. This can be seen as an advantage as together the twelve pilots have the potential to show a versatile picture of AD and its potential impacts. Having said that, it is stressed that one main goal of designing the experimental procedure was to increase harmonisation between the individual pilots, in particular, when studying similar ADFs, issues and topics by several pilots. In addition, it should be taken care that all research questions assessed as important will be covered by the consortium.

This report presents the recommendations for experimental procedures for L3Pilot-project pilot sites. It takes into account both theoretically preferred ways to collect the data, and practical limitations in the current L3AD on-the-road testing in European countries. Our intention was to apply the best practices and good principles described in the methodology literature. Experimental procedures to be carried out at the pilot sites are described in such a way that the data collection will allow L3Pilot project-level evaluation at a later stage of the project.

We are aware that while providing the overall description of the experimental procedures, and practical guidance for the actual pilot sites, we needed to make some compromises due to practical reasons and limitations - what is currently feasible from several perspectives, not least from the safe open-road testing of new technologies. For instance the user of professional drivers is not the optimal solution from purely theoretical point of view, but needed step to ensure safety before it is feasible to include general public in on-the-road FOTs or NDS.

In this document, we focused on what would be the most recommended solutions. At the same time we included more detailed recommendations taking into consideration currently known boundaries.

One important issue discussed in this deliverable, is the remarkable difference between FOTs of close-to-market products, and pilots of systems on earlier technology-readiness levels. In an AD pilot, satisfactory levels of field tests are controlled tests with a safety driver and OEMs' employees. This procedure is very different to FOTs, where ordinary drivers use the system as part of their daily lives. Thus, in a pilot study, field tests produce indicative estimates of impacts, while further assumptions need to be made on market-ready versions, and their use utilising other sources of information to complement the field measures in evaluation. In a FOT, one can expect more direct proof of impacts from the field measurements.

Methodology work continues in L3Pilot next with detailed definition of evaluation methods. The feasibility of research questions from data availability viewpoint will be added too. Additionally, the practical support for the pilot sites in implementation of the recommended experimental procedures will be continued by the selected evaluation partners, and dedicated detailed pilot site specific notes on experimental procedures tailored for them. Final checking of methodology, including experimental procedure principles and recommendations will be made for the final deliverable for the L3Pilot methodology.

References

Bundesagentur für Arbeit (German Ministry of Labour), (2018) Statistik der Bundesagentur für Arbeit, Tabellen, Beschäftigte nach Berufen (KldB 2010) (Quartalszahlen), Nürnberg, September 2018.

Caird, J. K., & Horrey, W. J. (2011). Twelve practical and useful questions about driving simulation (pp. 5-1). CRC Press, Boca Raton, Fla.

Dotzauer, M., Stemmler, E., Utesch, F., Bärgman, J., Guyonvarch, L., Kovaceva, J., Tattegrain, H., Zhang, M., Hibberd, D., Fox, C., Carsten, O., (2012) UDRIVE deliverable 42.1 Risk factors, crash causation and everyday driving of the EU FP7 Project UDRIVE (www.udrive.eu)

Edison, S. W., & Geissler, G. L. (2003). Measuring attitudes towards general technology: Antecedents, hypotheses and scale development. Journal of Targeting, Measurement and Analysis for Marketing, 12(2), 137–156. https://doi.org/10.1057/palgrave.jt.5740104

European Commission. (2018). Professional drivers. Retrieved June 13, 2018, from https://ec.europa.eu/transport/road_safety/users/professional-drivers_en

European Parliament and the Council of the European Union. DIRECTIVE 2003/59/EC OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL (2003).

Fahrenkrog F., Wang, L., Rösener, C., Sauerbier, J. & Breunig, S. (2017), Impact analysis for supervised automated driving applications, AdaptIVe Deliverable D7.3, 2017.

FOT-Net (2016). FESTA Handbook, version 6.

Hibberd D, Louw T, Aittoniemi E, Brouwer R, Dotzauer M, Fahrenkrog F, Innamaa S, et al. (2018). From Research Questions to Logging Requirements. L3Pilot Deliverable D3.1.

Hogema, J.H. (1996). Effects of rain on daily traffic volume and on driving behaviour (Report TM-96-B019). Soesterberg, The Netherlands: TNO Human Factors Research Institute.

Hogema, J.H., Veltman, J.A., & Van 't Hof, A. (2005). Effects of motorway lighting on workload and driving behaviour. G. Underwood (Ed.), Traffic & Transport Psychology - Theory and application. Proceedings of the ICTTP 2004.

Hogema, J. (2009). AOS data analyse (Rapport TNO-DV 2009 IN 297). Soesterberg: TNO Defensie en Veiligheid.

Janssen, W., Nodari, E., Brouwer, R., Plaza, J., Östlund, J., Keinath, A., Toffetti, A., Alonso, M., Rimini-Doering, M., Portouli, V., Horst, D., Marberger, C., Vega, H., Hherri, C., (2008). Specification of AIDE methodology. AIDE project deliverable A2.1.4. Available online: http://www.aide-eu.org/pdf/sp2_deliv_new/aide_d2_1_4_summary.pdf

Kelly, K., & Preacher, K. J. (2012). On effect size. Psychological Methods, 17(2), 137-152.

Krueger, R. A., & Casey, M. A. (2014). Focus groups: A practical guide for applied research. Sage publications.

Lietz, H., Petzoldt, T., Henning, M., Haupt, J., Wanielik, G., Krems, J. & Noyer, U. (2011). Methodische und technische Aspekte einer Naturalistic Driving Study. FAT-Schriftenreihe, (229).

Martinussen, L. M., Hakamies-Blomqvist, L., Møller, M., Özkan, T., & Lajunen, T. (2013). Age, gender, mileage and the DBQ: The validity of the Driver Behavior Questionnaire in different driver groups. Accident Analysis and Prevention, 52, 228–236. https://doi.org/10.1016/j.aap.2012.12.036

Metz, B., Landau, A., Hargutt, V., & Neukum, A. (2013). *Naturalistic Driving Data - Re-Analyse von Daten aus dem EU-Projekt euroFOT* (FAT-Schriftenreihe Nr. 256). Berlin: Forschungsvereinigung Automobiltechnik e.V.

PReVENT, 2009. Code of Practice for Design and Evaluation of ADAS, version 5.0. Preventive and Active Safety Applications; Integrated Project Contract number FP6-507075

Reason, J. T., Manstead, A., Stradling, S G., Baxter, J., Campbell, K., 1990. Errors and violations on the road – a real distinction. Ergonomics, 33 (10/11), 1315-1332.

Roesener, C., Hennecke, F., Sauerbier, J., Zlocki, A., Kemper, D., Oeser, M. & Eckstein, L. (2015), A Traffic-based Method for Safety Impact Assessment of Road Vehicle Automation, Workshop Fahrerassistenzsysteme, Walting, 2018.

Shinar, D. 2017. Traffic Safety and Human Behaviour. Emerald publishing.

Zuckerman, M. 1994. Behavioral expressions and biosocial bases of sensation seeking. Cambridge, UK: Cambridge Univ. Press.