Deliverable **D3.4** /

Evaluation plan

Version: 1.0 Final

Dissemination level: PU

Lead contractor: VTT

Due date: 29.02.2020

Version date: 01.07.2020

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

Document information

Authors

Satu Innamaa – VTT

Elina Aittoniemi – VTT

Afsaneh Bjorvatn – SNF

Felix Fahrenkrog – BMW

Johann Gwehenberger – Allianz Deutschland

Esko Lehtonen – VTT

Tyron Louw – University of Leeds

Fanny Malin – VTT

Merja Penttinen – VTT

Roland Schindhelm – BASt

Anne Silla – VTT

Hendrik Weber – ika

Marcel Borrack – Allianz Deutschland

Luigi Di Lillo – SwissRe

Natasha Merat – University of Leeds

Barbara Metz – WIVW

Yves Page – Renault

Elisabeth Shi – BASt

Henri Sintonen – VTT

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

Summary

The L3Pilot project conducts large-scale testing and piloting of highly automated driving functions (ADFs), focusing on SAE Level 3, exposed to different users and mixed traffic environments, including conventional vehicles and vulnerable road users (VRUs) along different road networks. Extensive on-road testing is vital to ensure adequate operating performance for the ADFs, which includes understanding the changes in vehicle operations and traffic dynamics, and users' interaction with, and acceptance of, ADFs. A multidisciplinary methodology tailored to the assessment of automated driving was needed to ensure the success of the pilots in this project and, thus, the overall objective of the methodology subproject (SP3) was to develop the methodology for piloting, testing and evaluating Level-3 ADFs. This Deliverable *D3.4 Evaluation plan* is the fourth and last deliverable of the SP3 Methodology. It defines the methods and data sources for all evaluation areas and research questions. It also reports the lessons learned and good practices found for evaluation of AD studies with real-world pilot data in general.

The methodology described in this deliverable relates to the full chain of assessment, starting from technical assessment of ADF and user evaluation, to adaptation of modelling approaches for impact assessment and scaling up of the impacts to EU level, and ending with choosing the best set-up for socio-economic evaluation of the expected impact. The work started by setting the research questions. They were organised at three levels of detail, with 11 main questions (Level-1) on four evaluation areas: technical & traffic evaluation, user & acceptance evaluation, impact assessment, and socio-economic impact assessment. More detailed questions were developed related to specific components of the questions where appropriate. In all, there were 37 Level-2 and 68 detailed Level-3 research questions. An overview of the interrelations between these evaluation topics is presented, and an evaluation plan including methods and data sources is specified for all the research questions.

Next, this deliverable summarises the process for setting up the experimental procedures for L3Pilot. The procedures were developed to provide a solid base for the evaluation methodology, and to ensure that the results from tests across all pilot sites could lead to an L3Pilot-wide evaluation, taking into account the practical limitations of their implementation. Furthermore, the aim was to harmonise the evaluation criteria by providing detailed recommendations for the pilots with the intention of creating holistic evaluation results of the L3Pilot project. This deliverable provides an update to the earlier practical recommendations given to the pilot sites based on feedback one year after giving the original recommendations.

The work on setting the methods used the best practices of previous projects on ADAS by further developing and adapting the methods to SAE Level 3 automation. For example, due to the nature of ADFs, the prototype systems tested in the project, and the need to ensure the safety of the driver and all surrounding traffic throughout the tests, adaptation and further development of existing methods was needed and implemented. This deliverable presents the final overall evaluation plan for all evaluation areas, i.e. details on how research questions covered by the

project will be addressed, including data sources and methods with possible updates to the topics presented in the previous three deliverables.

Technical & traffic evaluation and user & acceptance evaluation is based on the 20 different ADFs used in the field tests at the 16 pilot sites for driving on a motorway, in an urban environment or for parking. To be able to analyse the overall impact of these functions, that is, independent of the individual implementations of specific vehicle brands, the ADFs will be grouped in the analysis and indicators will be presented such that they are merged across pilot sites. Impact assessment and socio-economic assessment do not address the prototype implementations as they target the potential impacts of ADFs in perspective, when they are in use on a larger scale, and it is expected that the ADFs will be developed further from the ones tested in L3Pilot. Therefore, so-called mature functions are defined to represent such future ADFs.

The objective of the technical & traffic evaluation in L3Pilot is to answer research questions related to the technical performance and the effects resulting from an automated system's behaviour within traffic. Different effects are evaluated by comparing performance indicators. For user & acceptance evaluation, a multifaceted approach is used to form a holistic view of users' behaviours with, and acceptance of, the ADFs. The design of this approach was underpinned by a list of priority research questions. They address user acceptance and trust of the systems, willingness to use and pay for the functionalities, measures of driver state, user risk perception, driver engagement in non-driving related tasks, user behaviour during and after take-over situations, and user motion sickness.

Impact assessment addresses personal mobility, traffic safety, traffic efficiency and the environment. The evaluation plan includes the methods for scaling up of impacts to EU level to ensure a versatile overview of the implications of SAE Level 3 automation in passenger cars. Planned methods include simulations, accident, traffic, road environment and mobility data collection and analysis, etc. The assessment of the socio-economic impacts is concerned with the net welfare gain for the society regarding each of the ADFs in question. Net economic impacts may also be estimated for different stakeholders. The purpose of the socio-economic impact assessment is to clarify whether the benefits of equipping vehicles with ADFs outweigh their costs. A snapshot approach was chosen for quantifying the socio-economic impacts.

Finally, this deliverable reports lessons learned, and good practices found, for the evaluation methodology for AD studies with real-world pilot data and for supplementary studies. The deliverable gives an overview of contributions to the Code of Practice (CoP) from the viewpoint of methodology. The methodology subproject supported the development of the guidelines by specifying requirements and providing input regarding testing, evaluation and impact assessment from a methodology point of view. An additional set of lessons learned are also documented for future automated driving pilots and evaluations.

In L3Pilot, the final details for the methodology will be set during the actual evaluation work in SP7 Evaluation, reflecting the data and results from the pilot experiments. This deliverable concludes the work of SP3 Methodology.

Table of contents

List of figures

List of tables

1 Introduction

1.1 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. Nevertheless, the technology is rapidly advancing and is today at a stage that justifies AD tests in large-scale pilots.

L3Pilot is taking the final steps before the introduction of highly automated vehicles in daily traffic. AD is not achieved simply by integrating more and better technology, but also incorporating user behaviour and needs into the design of AD systems. Therefore, user acceptance is a key factor in the success of AD on the market. In addition, there are many broad legal restrictions and considerations which need to be addressed before AD can be rolled out. Thus, the overall objective of the L3Pilot project is to test and study the viability of AD as a safe and efficient means of transportation, and to gain a knowledge base for exploring and promoting new service concepts to provide inclusive mobility.

The L3Pilot project conducts large-scale testing and piloting of highly automated driving functions (ADFs), focusing on SAE Level 3 (Figure 1.1), exposed to different users and mixed traffic environments, including conventional vehicles and vulnerable road users (VRUs) along different road networks. The goal of the L3Pilot project is to demonstrate and assess the functionality and operation of Level 3 ADFs of passenger cars in real, or close-to-real, contexts and environments as well as evaluate the acceptance of these technologies. The project provides a great opportunity for large-scale on-road testing of automated driving functions, which are not yet available on the market. The engagement of a large number of OEMs, and the implementation of various ADFs in a range of environments (motorway, urban, parking), tested across many parts of Europe, enable a broader view of the potential impacts of automation, compared to evaluations based on a single trial.

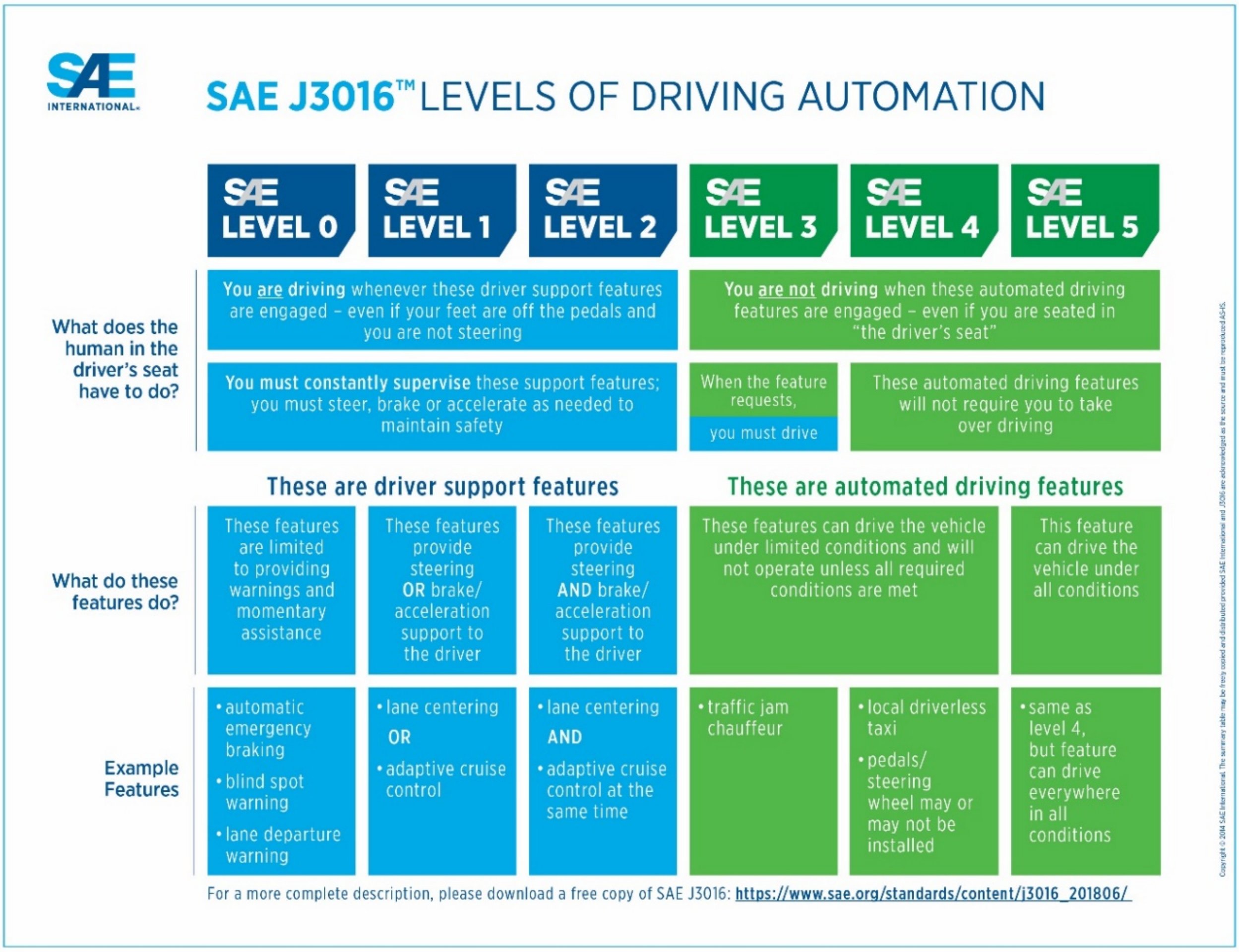


Figure 1.1: SAE Levels of Driving Automation J3016 JUN2018 (Copyright 2018 SAE International).

The key to successful piloting is to ensure that the ADFs used are exposed to variable conditions, but where performance is consistent, reliable, and predictable. This will ensure that test users gain experience with the dedicated features and behaviour of the ADF, which will facilitate studying the factors that lead to accelerating acceptance and adoption of the technology (Figure 1.2).

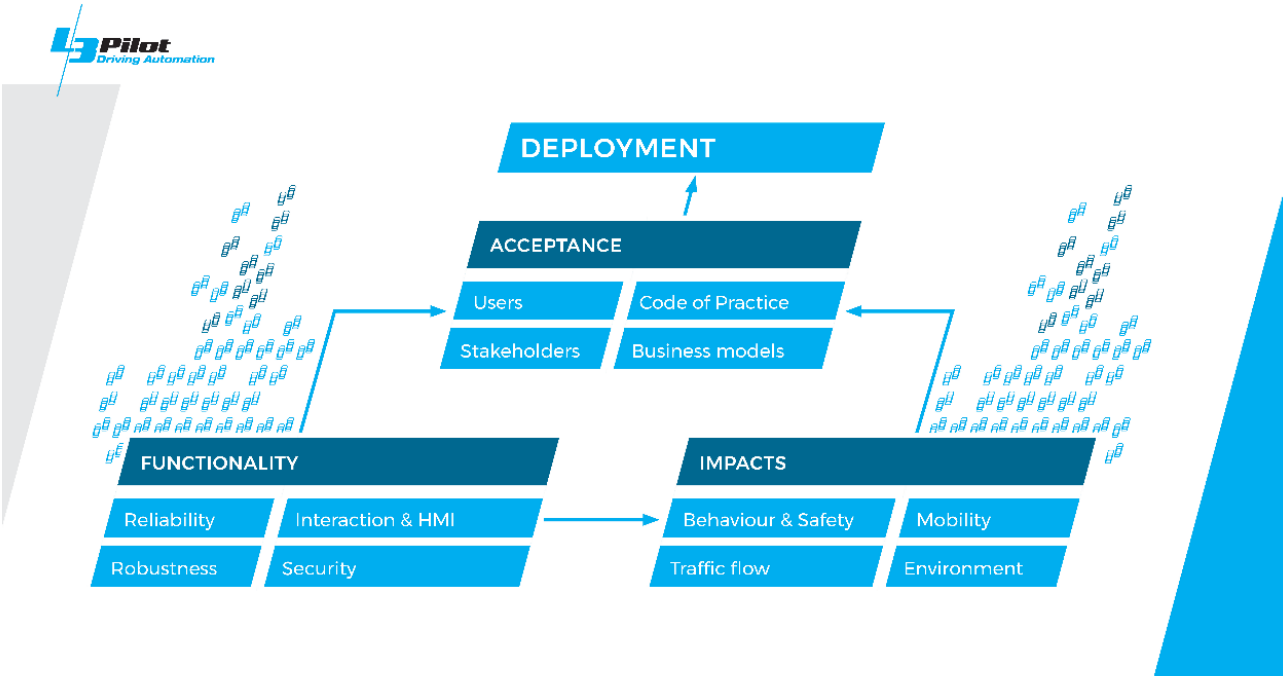


Figure 1.2: L3Pilot approach and mechanism for deployment.

The L3Pilot consortium brings together stakeholders from the whole value chain, including OEMs (original equipment manufacturers), suppliers, academic institutes, research institutes, infrastructure operators, governmental agencies, the insurance sector, and user groups. Since the development of ADFs, especially at SAE Level 3, is relatively well progressed, the aim is not only to pilot the ADFs, but also to study user acceptance and evaluation, reactions, and willingness to use vehicles equipped with such functionalities. This information leads the consortium to create plans for the market introduction of AD.

1.2 Overall methodology in L3Pilot

This chapter is modified from the conference paper published for Transport Research Arena 2020 (Conference cancelled) by Innamaa, Merat, Louw, Torrao & Aittoniemi, 2020.

FESTA methodology (see the latest version of the FESTA Handbook, FOT-Net & CARTRE (2018)) was designed to be applied to field-operational tests (FOTs) with market-ready products. Therefore, it does not necessarily fully apply to studies with prototypical ADFs. Thus, some adjustment of the “V” structure is needed to accommodate testing of prototype functionalities, such as ADFs, in real traffic. The pilot nature of the tests in L3Pilot brings some practical and ethical limitations regarding the use of the automated vehicles and limits any firm conclusions drawn about their implementation in the real world, or their expected impacts. To generate valid conclusions regarding the impacts of the ADFs, principles used to collect the evaluation data and any ensuing conclusions needed to be considered carefully.

The L3Pilot project adapts the original *FESTA V* (Figure 1.3) to better describe the key steps of the project. Specifically, the adaptation made to the original “V” is the following:

- The description of *functions & use cases* involves two separate steps in the original FESTA V. Because in L3Pilot the use case description is linked to the operational design domains (ODD) of the tested ADFs, these two aspects are combined in our adapted process.
- *Research questions & hypotheses* is kept the same as in the original FESTA V.
- In the original FESTA V, defining performance indicators is combined with devising the study design. Defining measures and sensors is a separate step. In the L3Pilot project, the data specification is combined as one step (*performance indicators & measures*) and *study design* is handled as a separate step due to the subproject structure of the project. In this step, L3Pilot also specifies the *evaluation plan*, whereas the original FESTA V does not include an evaluation plan specification at all.
- The original FESTA V goes directly from *measures & sensors* to *data acquisition*. In the adapted version, this is divided into four steps: *test site set-up* as part of preparation, *pre-tests,* and *test* as phases of the actual data collection and *test site wrap-up*, which is conducted after the tests.
- The phases *database* and *data analysis* in the original FESTA V are combined into *data processing* in the adapted version of the V.
- The original FESTA V phase research questions & hypotheses testing is divided into three phases in the adapted V: [evaluation of] technical performance & cybersecurity, user acceptance and driving & travel behaviour.
- The evaluation phase *impact assessment* is specified as *impact on safety, mobility, efficiency and environment* in our adapted V. The last phase of the evaluation is called *socio-economic cost-benefit analysis* in the original FESTA V. In the adapted version, it is called [the evaluation of] *societal impacts* to enable also incorporation of other analysis not considered in the traditional cost-benefit analysis.

Below, each step of the L3Pilot process adapted from FESTA V is described.

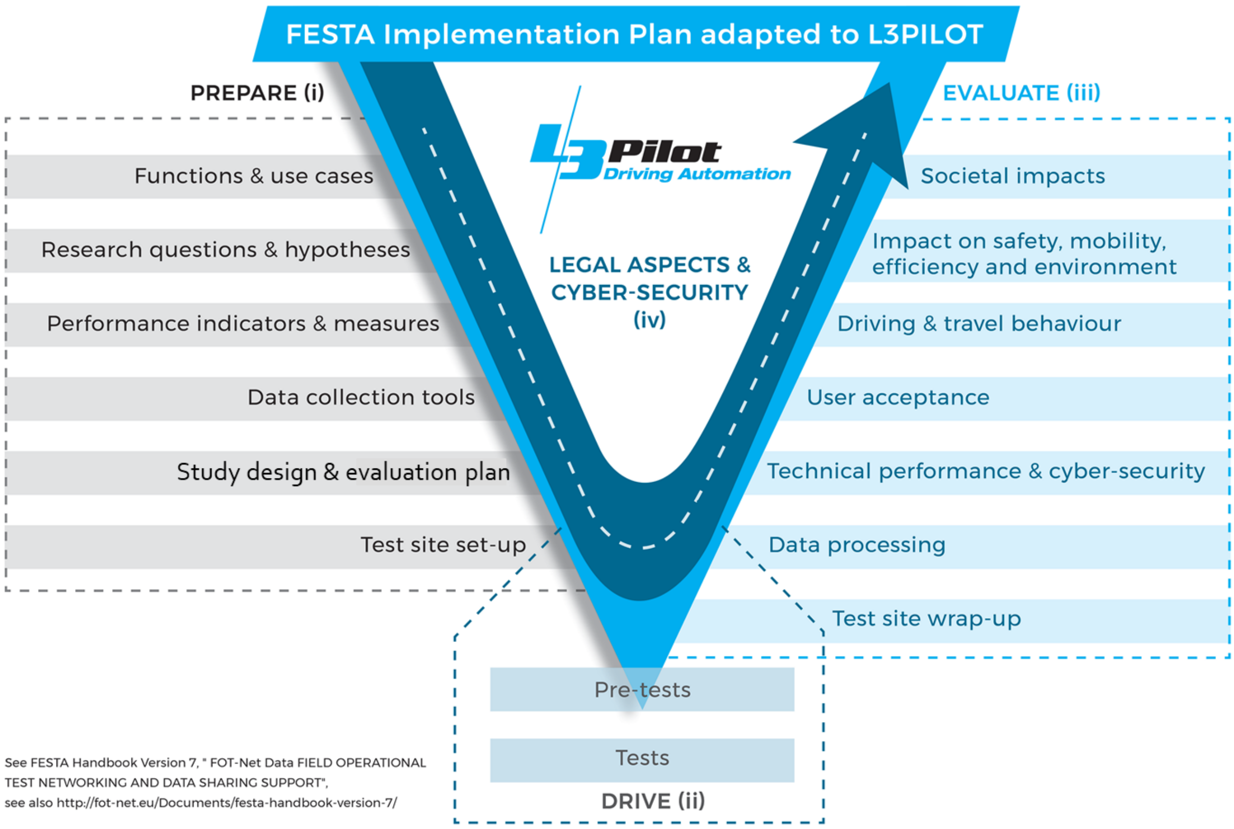


Figure 1.3: FESTA V adapted to the L3Pilot project (updated from Innamaa et al. 2020)

Functions & use cases

In the step *functions & use cases*, (an early) description of the ADF(s) of the test vehicles is made from the user viewpoint:

- What the vehicle is capable of doing (driving tasks included in the ADF)
- In which conditions the ADF is available (ODD specifications according to: road type and condition, weather, traffic situation)
- How the test vehicle communicates the availability of ADFs and requests to the user to resume control (HMI)

Pilots use vehicles with prototype ADFs in their field tests. These are used to collect vehicle technical data as well as to evaluate test users' experience. However, due to this prototype nature, it would be inaccurate to use this data directly to scale up the results to high penetrations on a regional level (e.g. European Union (EU)). Therefore, it is important to define the mature ADFs (i.e. what we expect the ADFs and their ODDs to be like when they are widely penetrated into the market) for the impact assessment. The description of these mature ADFs is done together with the developers of the L3Pilot demonstrator vehicles. Thus, for the evaluation, it is important to describe ADFs in terms of:

- Prototype(s) in real-world pilot,
- Mature ADF(s) for impact assessment,
- The differences between these two.

In practice, it is likely that there are differences between test vehicles: (1) over time within a single test vehicle, if the ADF development continues to unfold during the pilot project, (2) between the prototype ADF and the mature ADF, and (3) between the test environment(s) & conditions and the entire ODD. In addition, there are also differences inherent to changes in the driving code and regulatory framework of the region or country, which have an influence on the way that the test pilot can be carried out. For the evaluation, it is essential to understand and account for all these differences. Otherwise, there is a risk that the reliability of the user & acceptance evaluation and impact assessment are compromised. Therefore, it is an iterative process to update the function and use case descriptions to facilitate accurate evaluations.

See description of the piloted ADFs in L3Pilot Deliverable *D4.1 Description and taxonomy of AD functions* by Griffon, Sauvaget, Geronimi, Bolovinou & Brouwer (2019) and of mature ADFs in Chapter 4.1.1 of this report.

Research questions & hypotheses

The second step of the preparation phase is to define *research questions & hypotheses*. For AD pilots, this starts the same way as for any other study: from the theories behind the evaluation areas and from the (early) descriptions of the functions and use cases. However, the feasibility of the research questions is verified for:

- Data availability – due to protecting intellectual property rights related to these products.
- Test design – due to ethical, safety and regulatory restrictions set for the field experiments on open roads.
- Availability of research tools and methods – as not all previously used tools are automatically fit for assessing the impacts of AD.
- Availability of (time and human) resources (e.g. manual video annotation is time-intensive).

There are plenty of research topics that are of interest and prioritisation is needed. It is nevertheless advised to show the limitations of the evaluation (i.e. what was needed to be left out and how that may affect the overall picture given by the results).

See the process of setting the research questions and causalities between them in Chapter 2, and the lists of research questions for each evaluation area in Chapter 4.

Performance indicators & measures

The third step of the preparation phase is to define the *performance indicators* with which the research questions can be answered, the *derived measures* that are needed to calculate these indicators, and the signals that are needed to calculate these measures. As a single performance indicator may be derived from several alternative derived measures and a single measure may be derived from several alternative signals, a dialogue between the evaluation methodology team and the data providers (pilot sites) is needed to agree on a common list of signals to be collected for the evaluation.

The performance indicators and derived measures were set for the research questions in L3Pilot Deliverable *D3.1 From research questions to logging requirements* by Hibberd, Louw, Aittoniemi, Brouwer, Dotzauer, Fahrenkrog, et al. (2018). These led to a common data format (Hiller et al., 2019) for the signal level data. This format was a product of a dialogue between the evaluation team (data users), the vehicle owners (data providers) and SP5 Pilot tools and data responsible for the tools and data processes. The common data format enables the use of common data processing tools in the evaluations.

Data collection tools

The fourth step, setting up *data collection tools,* is a task for all the pilot sites in the preparation phase and involves locating and developing data collection tools that are capable of logging the data agreed in the previous step – the common data format in the case of L3Pilot (Hiller et al., 2019).

This step also includes the definition and development of databases storing data for different phases of processing and analysis. It may also be necessary for these tools to be able to pseudonymise and anonymise the pilot sites without compromising the validity of evaluation results.

As not all the data is logged from the vehicles, this step also includes planning additional supplementing data collection methods to facilitate answering the research questions. In L3Pilot, this includes questionnaires, Wizard-of-Oz vehicles, driving simulators, etc.

See description of pilot tools for L3Pilot in Deliverable *D5.1 Pilot Tools for L3Pilot* by Nagy, Hiller, Svanberg, Kremer, Luxen, Christen, et al. (2019).

Study design & evaluation plan

The fifth step is setting the *study design & evaluation plan*. Selecting the baseline is an important task in the planning of tests and evaluations. For each research question, the evaluation team must consider whether a baseline is needed and, in that case, what a valid baseline is. Here, a relevant factor for selecting the baseline is the feasibility of the data collection. The inclusion or exclusion of active safety systems and other advanced driver assistance systems (ADAS) and the implications of their inclusion on the impact estimates must also be carefully considered.

An important step in the preparation of the pilot is to agree on the study design to be implemented across pilot sites. This includes selecting the test route and test participants, running the user tests, collecting baseline and treatment data, etc. The criteria for recruiting and selecting the test participants may include e.g. that all test participants should regularly drive, and demographic factors should reflect the driver population of interest, for example the future customer population (depending on evaluation scope).

The introduction of AD will change the mobility ecosystem; therefore, new assessment approaches may be needed. An example of the need for adjustment of FESTA V relates to the FESTA guideline that the "ordinary user" should be able to operate the tested function in real traffic without supervision. However, owing to the practical, safety and ethical issues related to piloting of prototype ADFs, the L3Pilot approach restricts the operation to predetermined test routes. Moreover, the L3Pilot approach needed to be adjusted for the type and role of the test participant (professional driver[1], safety driver, and ordinary[2] driver). For example, when dealing with prototype systems in mixed traffic on open roads, the use of specially trained safety drivers is a requirement for safety, legal, and ethical reasons. Consequently, it is not possible to conduct the same nature of research as in a naturalistic driving study or FOT. This has implications for the scope of user & acceptance evaluations, studies on travel behaviour, and the collection of baseline data. Therefore, the evaluation team needs to anticipate and plan for how to utilise the controlled tests on open roads and supplementing studies to collect the data needed for the evaluation.

As there are differences in the rules and regulations for field testing for each vehicle manufacturer and region, an active dialogue between the evaluation team and the pilot sites is needed during the study planning. Otherwise, the possibility for assessing generalised results across all pilot sites may be difficult.

In this step, the evaluation plan needs to be set for all the research questions in terms of evaluation methods (tools, data needs) to plan how the collected data is used and what additional data or information is needed (external data).

See Chapter 1.2 for the adaptation of FESTA V for L3Pilot, Chapter 4.1.3 for baseline and L3Pilot Deliverable *D3.2 Experimental procedure* by Penttinen, Rämä, Dotzauer, Hibberd, Innamaa, Louw, et al. (2019) for definition of experimental procedures. Recommendations for the AD pilots are updated in Chapter 3.2 of this deliverable. Chapter 4 includes the evaluation plan of all L3Pilot research questions.

Test site setup

For the *test site setup* step, an active dialogue between the evaluation/methodology team and the pilot site is recommended to ensure that the agreed experimental procedures are correctly understood and implemented. In the L3Pilot project, an evaluation partner is dedicated to each pilot site to plan the practical tests together. Each evaluation partner is also responsible for the analysis and processing of the field data at its respective pilot site. Therefore, the details of the field experiments need to be fully understood.

[1] Within L3Pilot, professional (test) drivers are defined as *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, for example, handle cars in critical situations. These drivers can be deployed to operate prototype vehicles undergoing road tests*.

[2] Within L3Pilot, ordinary drivers are defined as *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.*

Pre-tests

The *pre-tests* step involves running all the phases of the project on a small scale to ensure that all the processes and tool chains function as intended. This should include all the steps in the evaluation, related data flows, test participant handling, etc. Verification of all the steps is vital to the success of the project. The possibilities for corrective actions after the start of data collection are limited. Hence, pre-tests are useful to anticipate potential caveats and to validate the designed study prior to the start of the real pilot.

Tests

The *tests* phase involves the actual data collection. The number of vehicles used for data collection is typically less important than how much data is collected (in terms of vehicle kilometres driven, duration of the drives, and the number and type of test users), in which conditions (all dimensions of ODD, representative for scaling up), and how well it represents the phenomena under evaluation (prototype vs mature ADF, users, etc.).

Test site wrap-up

Test site wrap-up includes the delivery of collected data and metadata. In this phase, it is also important to report all the deviations from the plan and any system updates made during the data collection phase (which features were updated when, and what changes might appear in the ADF behaviour and hence impact users' experience).

Data processing

In this phase, the pilot sites handle the *data processing* converting their raw data into the common data format. Dedicated evaluation partners process this data according to commonly agreed principles and tools. They also upload data to a consolidated database to be used later by the other evaluation partners for different evaluation areas (e.g. technical & traffic and user & acceptance evaluation in L3Pilot).

Technical performance & cybersecurity

The evaluation of *technical performance & cybersecurity* aims to understand the system that the users experienced in the field tests. For the evaluation, it is important to know whether the system functioned as specified in the earlier phases of the development process. Any deviations causing unpleasant user experience influence the next phases of evaluation and, therefore, should be communicated to the evaluation team.

See evaluation plan for the technical evaluation in Chapter 4.2 of this deliverable. In L3Pilot, the pilot sites ensure the cybersecurity of the tested systems. Due to the prototype nature of these products, cybersecurity is not evaluated.

User acceptance

In the subsequent phase, the evaluation of *user acceptance* aims to understand users' experience in, and acceptance of, the tested ADF. Challenges for generalisation of the results emerge from the use of professional drivers or vehicle manufacturer employees and from users' potentially limited experience with the tested systems. The role of a safety driver is to ensure safety with the system activated and not to experience the system as a customer. The vehicle manufacturer employees may be more likely to drive a car and have a certain education (e.g. in engineering) than the average population. Thus, it is important to understand the implications that the (potential) non-maturity of the user experience and of the selected test user group have on the results.

See evaluation plan for the user & acceptance evaluation in Chapter 4.3 of this deliverable.

Driving & travel behaviour

Driving & travel behaviour evaluation aims to understand the changes that the introduction and use of ADFs will lead to. These changes should be reflected in the following phases of evaluation. It is also important to understand the implications that the differences between the prototype ADFs and their (potentially limited) testing environments and conditions have on these changes.

See the plan for evaluation of driving behaviour in the *technical & traffic* evaluation in Chapter 4.2 and for evaluation of travel behaviour in the *mobility impact assessment* in Chapter 4.4 of this deliverable. Due to the nature of controlled tests, the assessment of mobility impacts is limited, focusing only on potential implications that the availability of ADFs may have, without the possibility to monitor travel behaviour in the field experiments.

Impact on safety, mobility, efficiency and environment

This phase of the evaluation assesses the *impacts on safety, mobility, efficiency and environment* and scales them up to EU level. These results are needed in the last phase of evaluation, which is the assessment of societal impacts.

The impact assessment phase utilises data or results from the *driving & travel behaviour* evaluation and the *user acceptance* evaluation. Additional inputs include descriptions of the mature ADFs and their ODDs and the penetration rates that will be used in the assessment of *societal impacts*. The impacts are first assessed on the level of a single event (single driving scenario), then on traffic flow and finally on the regional level (EU).

An assessment in an AD pilot faces more uncertainty compared to a traditional FOT of a market-ready product. The sources of that uncertainty include, but are not limited to:

- Uncertainty and disparity of mature ADFs.
- Differences between the tested system and (assumed) mature system.
- Lack of evidence regarding the behavioural adaptation of the other traffic participants (non-users).
- Gap between test situation and future traffic: typically single automated vehicle in today’s traffic vs. high(er) penetration rate of automated vehicles in a flow used to interact with them.
- Future dimensions: uncertainty and disparity in timing market introduction, parallel trends and changes affecting mobility, development phase of other ADAS (influencing inside and outside ODD).

Naturally, the reliability of impact assessment results depends also, for example, on the sophistication of the tools used in the assessment and on the quality and details of available data.

See evaluation plan for mobility, safety, efficiency and environmental impact assessment in Chapters 4.4–4.6 of this deliverable.

Societal impacts

Key factors for the success of the assessment of the *societal impacts* of AD are the accuracy of statistics and how their details meet the needs of this phase of evaluation. If the statistics do not provide information on how many road crashes today in the EU take place on roads and in conditions matching the ODD of the ADFs, or how much CO_2 emissions are contributed by vehicles driving in environments and conditions where ADFs could be used, the uncertainty in the scaling up of the results increases and reduces the reliability of the impact estimates. For the cost-benefit analysis, a further source of uncertainty is the lack of reliable (publicly available) estimates of the costs of these future systems, since they are still under development.

See evaluation plan for scaling up of the impacts in Chapter 4.4 and the plan for the socio-economic impact assessment in Chapter 4.7 of this deliverable.

1.3 Role of the Methodology subproject in L3Pilot

The work in L3Pilot is structured into subprojects following the process proposed by the FESTA methodology. The objectives of the *Methodology* subproject (SP3) in L3Pilot were to:

- Develop a methodology for the piloting, testing and evaluation of ADF 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 wellbeing 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, the subproject on methodology first provided a list of research questions (see L3Pilot Deliverable D3.1 by Hibberd et al., 2018). It then developed innovative and appropriate experimental procedures to collect the data required to answer these questions to ensure that reliable and valid results are achieved from the pilot testing (see L3Pilot Deliverable D3.2 by Penttinen et al., 2019). Finally, it described the methods to be used in the project by SP7 Evaluation (see L3Pilot Deliverable *D3.3 Evaluation methods* by Metz, Rösener, Louw, Aittoniemi, Bjorvatn, Wörle, et al., 2019), meeting the objectives defined above. In D3.3, the original research questions of D3.1 were partly rephrased or improved to facilitate comprehension and to better reflect the chosen evaluation approach.

There are close interactions with other subprojects in order to define a methodological approach that is feasible within the project and incorporates special requirements of L3Pilot. In particular, interactions with the following subprojects are relevant:

- Subproject on pilot tools and data (SP5): The data needs for the planned analysis are communicated to and discussed with partners responsible for defining and implementing the common data analysis tools in L3Pilot. The implemented tools are designed to meet the data needs defined by the Methodology subproject.
- Subproject on pilot preparation and support (SP4) and subproject responsible for piloting (SP6): The developed methodology is fitted to the ADFs prepared for testing at the different pilot sites, also considering practical limitations, regarding, for example, the testing environment or legal requirements at the pilot sites.
- Subproject on evaluation (SP7): The feasibility of the proposed methodology is continuously evaluated in terms of the available time and budget.
- Subproject on Code of Practice (SP2): Creation of unified de-facto standardised methods to ensure further development of AD applications.

1.4 Scope of planned evaluations in L3Pilot

The overall goal of L3Pilot is to demonstrate and assess ADFs in close-to-real or real use contexts and environments via on-road tests. There will be four primary areas of analysis:

1. Technical & traffic evaluation, assessing the effects of the ADFs on vehicle behaviour and the surrounding traffic based on data logged directly in the on-road tests.
2. User & acceptance evaluation, assessing users' evaluation and acceptance of, and behaviour while using, the tested ADFs.
3. Impact assessment extrapolates these results and estimates the potential impacts of expected mature ADFs on personal mobility, traffic safety, traffic efficiency and the environment.
4. Socio-economic impact assessment utilises the above analyses to determine monetary values for the estimated effects, weighting expected costs and benefits of the ADFs.

The areas not only address different aspects of the evaluation but also work with different datasets. In Figure 1.4, the blue area marks the datasets that will be used for the different areas of analysis and their topics (e.g. socio-economic impact evaluation will deal with a cost-benefit analysis and will use aggregated data to do so, or user evaluation will look at acceptance by using data from single vehicles and data combined at a fleet level that is per pilot site).

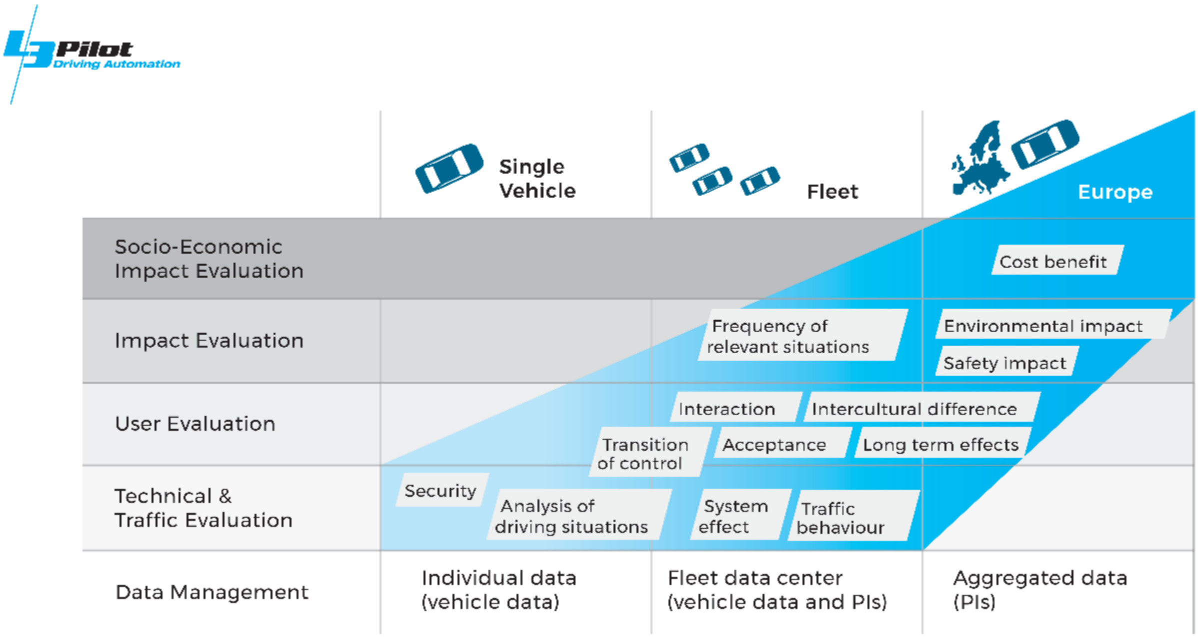


Figure 1.4: Considered evaluation areas and scope of assessment.

Besides answering the related research questions, analyses conducted for the areas 1, 2, and 3 listed above also provide input needed for the analyses in other areas. The process of L3Pilot data analysis (Figure 1.5) starts with data collection during the on-road tests in SP6 Piloting. This data is transferred into a common data format (Hiller, Svanberg, Koskinen, Bellotti, & Osman, 2019) used throughout the project. This data format and the required tools are developed within the project, and conversion to it is a part of the work done at the pilot sites. Evaluation starts with an analysis of the data logged at the pilot sites split into the areas of technical & traffic and user & acceptance evaluation. The analysis for these two areas can be done in parallel and is used to answer related research questions, as well as to derive input for impact assessment. Once that input is available, impact assessment can be carried out. Here the effects that have previously been observed and quantified in technical & traffic and user & acceptance evaluation are used in conjunction with other inputs to estimate the expected impacts of ADFs (e.g. on traffic safety or on traffic efficiency) in the broader context of use. Traffic simulation is the primary method for estimating impacts of ADFs on traffic safety, traffic efficiency and environmental aspects, while personal mobility assessment mainly utilises results from the user evaluation. The results on single scenarios are scaled up to provide an estimated impact at EU level. Based on these estimated impacts, the socio-economic impact is derived.

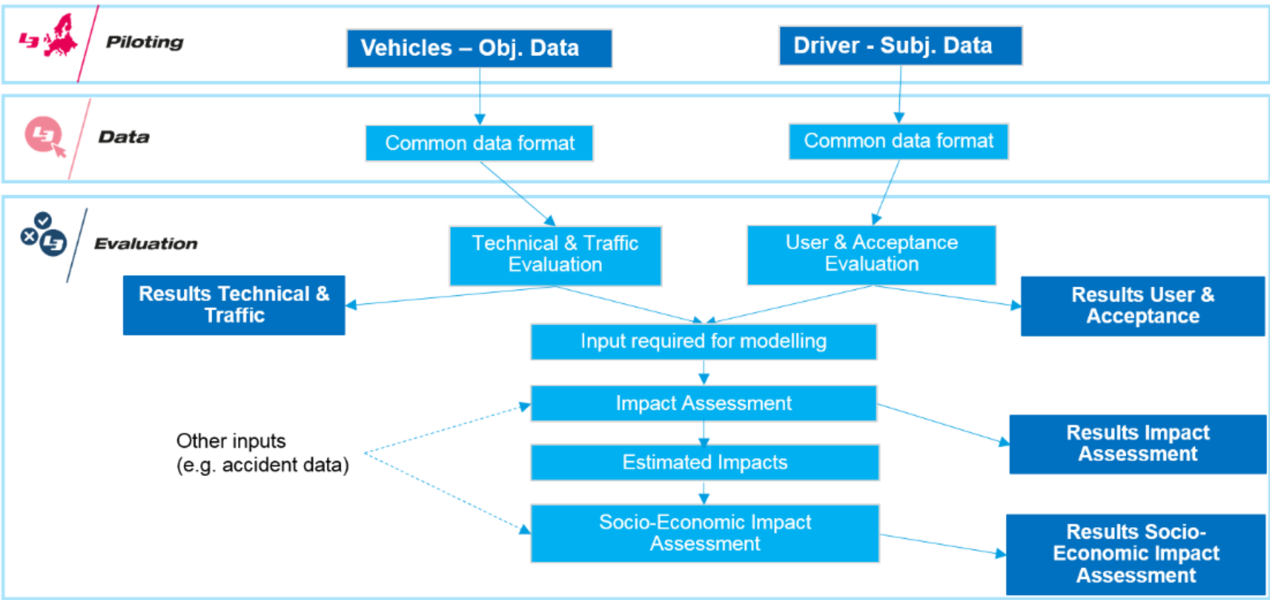


Figure 1.5: Overview of the expected overall process for data analysis.

1.5 Content of this deliverable and its relation to other deliverables

This Deliverable *D3.4 Evaluation plan* is the fourth and last deliverable on methodology in the L3Pilot project. D3.4 presents the final overall evaluation plan for all evaluation areas, i.e. details on how research questions covered by the project will be addressed, including data sources and methods with possible updates to the topics presented in the previous three deliverables. It also reports lessons learned and good practices found for evaluation methodology for AD studies with real-world pilot data and for the supplementary studies. D3.4 is the concluding deliverable of the work conducted in *SP3 Methodology*.

Within the L3Pilot project, there are supplementing research activities addressing some specific topics not directly related to road tests with the pilot vehicles. These include an international survey, driving simulators, and Wizard-of-Oz studies. The supplementary methods are described in this document regarding their contribution to answer the project's research questions in combination with the data from the pilots. For more detailed information on supplementing research, please refer to L3Pilot Deliverable *D7.1 Annual quantitative survey about user acceptance towards ADAS and vehicle automation* by Nordhoff, Beuster, Kessel, Bjorvatn, Ertl, Innamaa, et al. (in preparation) and Deliverable *D7.2 L3/L4 long-term study about user experiences* by Metz, Wörle, Zerbe, Schindhelm, & Bonarens (in preparation).

2 Research questions

This chapter summarises the process for setting up and selecting the research questions for L3Pilot; for more detail see L3Pilot Deliverable D3.1 by Hibberd et al. (2018). This chapter also lists the selected high-level research questions and the causalities between them.

2.1 Research question generation

The generation of research questions was started with an early project review of the ADFs that are likely to be available for testing. The ADF descriptions were used as a guide for the generation, while the precise technical capabilities of the systems were acknowledged later in the process. Therefore, only high-level research questions (Levels 1 and 2) were defined in the first phase, and the focus was on developing them to meet the objectives of the project. This process is described in detail in L3Pilot Deliverable D3.1 by Hibberd et al. (2018).

The top-down approach used for that was the impact area approach recommended by the FESTA Handbook (FOT-Net & CARTRE 2018). The basic principle for generating research questions and hypotheses using this top-down approach lies in a theoretical understanding of the factors that influence the different impact areas. Thus, the process was started with a thorough review of the existing literature to identify the main factors related to different impact areas and knowledge gaps relating to user or driving behaviours when using ADFs. In addition, repeated cycles of ‘research question generation – review – edit and addition’ were conducted to ensure coverage of all major topics with the potential to affect future knowledge, society, or business. This step in the process involved creating a wide-ranging set of research questions, uninhibited by the limitations of any single data collection methodology. These were based on the established literature and had input from project members, based on their experience in previous, related work. The generation of the research questions (Level 1) was structured around the four evaluation areas of L3Pilot: technical & traffic evaluation, user & acceptance evaluation, impact assessment, and socio-economic impact assessment.

The second stage of research question setting involved the development of more detailed questions related to specific components of the higher-level questions, where appropriate. For example, different aspects of user acceptance (Level 2) were elaborated one step further by specifying more detailed questions about related attributes perceived by the driver such as safety, comfort, reliability, usefulness and trust (Level 3).

The top-down approach to setting research questions was supported by a bottom-up check at a later stage, whereby in line with the FESTA methodology, the developed research questions were cross-checked for feasibility based on the data availability within the project, suitable experimental procedure at the pilot sites, and availability of evaluation methods and tools. In the prioritisation of research questions, also test site characteristics, coding and processing demand, ethical constraints, project resources and timescale, and research importance were considered. A scoring system was employed to aid the prioritisation process. Research questions that were rated as highly important and highly feasible were automatically adopted by the project. Research questions that were rated as highly important, but ultimately not feasible, were not adopted. Research questions that were rated as being less important were not generally adopted by the project, irrespective of their feasibility, in order to allocate resources to the areas with the highest priority. More detail can be found in L3Pilot Deliverable D3.1 (Hibberd et al., 2018). An update of the first two levels of research questions after the feasibility check and the Level-3 questions were reported in L3Pilot Deliverable D3.3 by Metz et al. (2019).

The tables below (from Table 2.1 to Table 2.4) list the high-level research questions (Levels 1 & 2) and the detailed research questions (Level 3) selected for evaluation in SP7. The evaluation plan for all the research questions is presented in Chapter 4.

Table 2.1: Research questions for technical & traffic evaluation.

RQ-ID	RQ Level 1	RQ Level 2	RQ Level 3
RQ-T1	What is the ADF's technical performance?	How reliable is ADF performance in a given driving and traffic scenario?	How often is the ADF available in the driving and traffic scenarios of its ODD?
RQ-T2		How often and under which circumstances do the ADFs issue a take-over request?	Does the ADF initiate a take-over request if required by the boundaries of the ADF?
RQ-T5	What is the impact on the driving behaviour of the ADF vehicle?	How do take-over requests affect driving?	How do planned take-over situations affect the driving dynamics of the vehicle?
RQ-T6		What is the impact of ADF on the driving dynamics?	What is the impact of ADF on longitudinal acceleration in defined driving situations?
			What is the impact of ADF on lateral acceleration in defined driving situations?
RQ-T7		What is the impact of ADF on the accuracy of driving?	What is the impact of ADF on precision of manoeuvres?
			What is the impact of ADF on lane-keeping performance in defined driving situations?
RQ-T8		What is the impact of ADF on the driven speed?	What is the impact of ADF on driven velocity in defined driving situations?
RQ-T9		What are the impacts of ADF on energy efficiency?	What is the impact of ADF on energy demand?
RQ-T10		What is the impact of ADF on the frequency of near crashes / incidents?	What is the impact of ADF on the frequency of harsh braking events?
			What is the impact of ADF on the frequency of (unintended) lane departures?
RQ-T11		What is the impact of ADF on the frequency of certain events?	What is the impact of ADF on the frequency of defined driving scenarios?

RQ-ID	RQ Level 1	RQ Level 2	RQ Level 3
RQ-T12	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?	What is the impact of ADF on lateral distances to other vehicles in defined driving scenarios?
			What is the impact of ADF on the behaviour of VRUs (cyclist, motorcyclist, pedestrian) in defined driving scenarios?
			What is the impact of ADF on the behaviour of approaching / crossing pedestrians?
			What is the impact of ADF on car-following behaviour?
RQ-T13		What are the impacts of ADF on traffic efficiency?	What is the impact of ADF on traffic flow at intersections?
RQ-T14		What is the impact of ADF on the number of near crashes / incidents with other road users?	What is the impact of ADF on the frequency of near crashes with other vehicles?
			What is the impact of ADF on near crashes with VRUs?
RQ-T15	What is the impact of ADF on the behaviour of other traffic participants?	How does the ADF influence the behaviour of subsequent vehicles?	What is the impact of ADF on the following behaviour of subsequent vehicles?
RQ-T16		How does the ADF influence the behaviour of preceding vehicles?	What is the impact of ADF on the frequency of defined driving scenarios (e.g. cut-in manoeuvres) of the vehicles in front of the ego-vehicle?
			What is the impact of ADF on the behaviour of preceding vehicles?
RQ-T17		What is the impact of ADF on the number of near crashes / incidents of other traffic participants?	What is the impact of ADF on the frequency of harsh braking events of the subsequent vehicle?
			What is the impact of ADF on the frequency of defined driving scenarios (e.g. very small distance) of the subsequent vehicle?

Table 2.2: Research questions for user & acceptance evaluation.

RQ-ID	RQ Level 1	RQ Level 2	RQ Level 3
RQ-U1	What is the impact on user acceptance & awareness?	Are drivers willing to use an ADF?	
RQ-U2		How much are drivers willing to pay for the ADF?	
RQ-U3		What is the user acceptance of the ADF?	What is the perceived safety of the ADF?
			What is the perceived comfort of the ADF?
			What is the perceived usefulness of the ADF?
			What is the perceived trust of the ADF?
			How does user acceptance differ between ADF types?
RQ-U4		What are drivers' expectations regarding system features?	What is the drivers' overall impression of the system?
RQ-U5		What is the impact of ADF on driver state?	What is the effect of ADF use on drivers' level of stress?
			What is the drivers' level of fatigue while using the ADF?
			What is the drivers' workload while using the ADF?
RQ-U6		What is the impact of ADF use on driver awareness?	What is the effect of ADF use on driver attention to the road/other road users?
			What is the drivers' risk perception while using the ADF?
RQ-U7	What is the user experience?	What is the impact of ADF use on motion sickness?	
RQ-U8		What is the impact of motion sickness on ADF use?	
RQ-U9		What is drivers' secondary task engagement during ADF use?	What secondary tasks do or would drivers engage in during ADF use?
			What is the frequency and duration of drivers' secondary task engagement during ADF use?
RQ-U10		How do drivers respond when they are required to retake control?	How do drivers respond when they are required to retake control in planned take-overs?
			How do drivers respond when they are required to retake control in unplanned take-overs?
RQ-U11		How often and under which circumstances do drivers choose to activate/deactivate the ADF?	How often and under which circumstances do drivers choose to activate the ADF?
			How often and under which circumstances do drivers choose to deactivate the ADF?

Table 2.3: Research questions for impact assessment.

RQ-ID	RQ Level 1	RQ Level 2	RQ Level 3
RQ-I1	What is the impact of ADF on traffic safety?	What is the impact of ADF on the number of accidents in a certain driving scenario / for certain road users?	What is the impact of ADF on the number of accidents in a certain driving scenario?
			What is the impact of ADF on the number of accidents involving other road users such as pedestrians and bikers?
RQ-I2		What is the impact of ADF on accidents with a certain injuries level / damage in a certain driving scenario?	What is the impact of ADF on accidents with fatal injuries in a certain driving scenario?
			What is the impact of ADF on accidents with severe injuries in a certain driving scenario?
			What is the impact of ADF on accidents with slight injuries in a certain driving scenario?
			What is the impact of ADF on accidents with material damages in a certain driving scenario?
			What is the impact of ADF on the rescue chain in terms of preventing injuries?
RQ-I3	What is the impact of ADF on traffic efficiency?	What is the impact on transport network efficiency (throughput) in a certain traffic scenario?	What is the impact of ADF on throughput in a road section or intersection?
			What is the impact of ADF on reliability of travel time?
			What is the impact of ADF on travel times?
			What is the impact of ADF on speed differences between vehicles?
			What is the impact of ADF on network capacity?
RQ-I4	What is impact of ADF on the environment?	What is the impact of ADF on energy demand / pollution in a certain traffic scenario?	What is the effect of the ADF on fuel consumption? (In specific scenarios?)
			What is the effect of the ADF on energy demand?
			What is the impact of the ADF on CO2 emissions?
RQ-I5	What is the impact of ADF on (personal) mobility?	What is the impact of ADF on amount of travel?	What is the impact of ADF on number of trips made?
			What is the impact of ADF on trip distance?
			What is the impact of ADF on trip duration?
RQ-I6		What is the impact of ADF on travel patterns?	What is the impact of ADF on mode choice?
			What is the impact of ADF on timing of trips?
			What is the impact of ADF on the frequency of road type usage? (urban, rural, motorway)
RQ-I7		What is the impact of ADF on quality of travel?	What is the impact of ADF on quality of travel?

Table 2.4: Research questions for socio-economic impact assessment.

RQ-ID	RQ Level 1	RQ Level 2	RQ Level 3
RQ-S1	What is the overall socio-economic impact of L3 ADFs?	What is the net welfare gain for the society?	What is the impact of ADF with respect to direct welfare effects for the society?
			What is the impact of ADF with respect to indirect welfare effects for the society?
RQ-S2		What is the overall socio-economic impact for different groups/stakeholders?	What is the welfare effect in terms of generalised travel costs?
			What are the welfare effects for travellers, producers and the government?

2.2 Causal diagram for the research questions

An overview of the interrelations between different evaluation topics is shown as a causal diagram in Figure 2.1. The diagram does not cover the detailed research questions and partly uses more general terms than in their formulation in order to be easier to understand, staying rather on a high level than going into all the details.

A link between items marked positive indicates a positive relation between them, and a link marked negative indicates a negative relation. In other words, a positive causal link means that the two items change in the same direction; thus, if an item where the link starts decreases, the other item also decreases. Similarly, if the item where the link starts increases, the other item increases as well. An example of this is an increase in quality of travel with AD leading to an increase in the perceived benefits of AD. A negative causal link means that the two items change in opposite directions; thus, if the item where the link starts increases, the other item decreases and vice versa. An example of this is an increase in willingness to use and pay for AD leading to a decrease in amount of travel with other transport modes.

The links between the Level-2 research questions are explained in the text following the illustration.

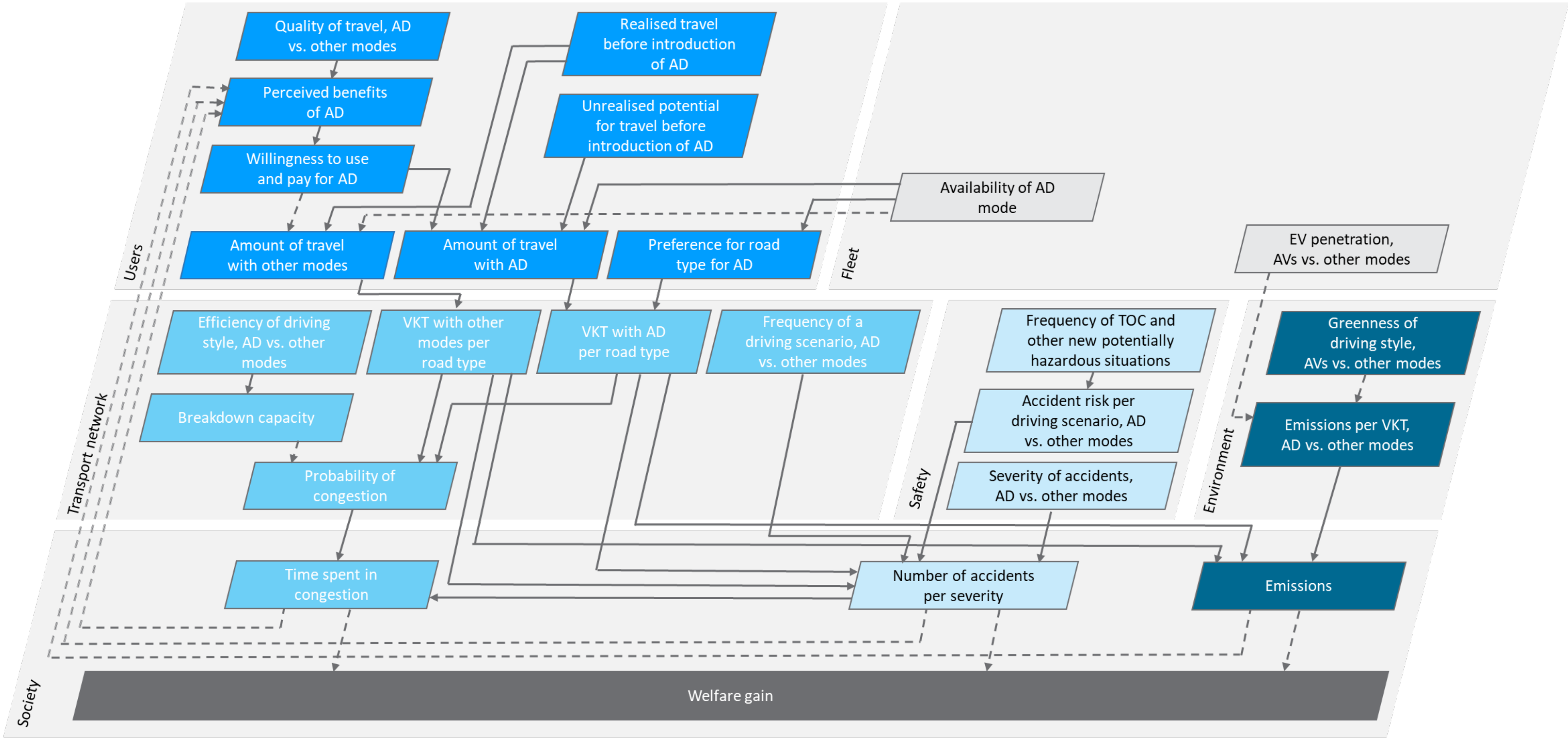


Figure 2.1: Causal diagram of different evaluation topics; positive relations are marked with a solid line and negative relations with a dotted line. Colours indicate different evaluation areas.

The impact on the *quality of travel* (research question RQ-I7) with vehicle automation affects the *perceived benefits* (RQ-U3 & RQ-U4) that potential users see for AD. In L3Pilot, driver state (RQ-U5), risk perception (RQ-U6) and motion sickness (RQ-U7 & RQ-U8) are considered as measures of the quality of travel in addition to the possibilities for secondary task engagement (RQ-U9). *Willingness to use and pay for AD* (RQ-U1 & RQ-U2) is affected by the perceived benefits. Different socio-demographic groups likely see the benefits differently, and their possibilities for paying for AD varies.

The *willingness to use and pay for AD* decreases the *amount of travel* (RQ-I5) *with other modes* (manually driven cars, public transport and active modes, RQ-I6) for the *realised travel* that took place before the introduction of automated vehicles, and increases the *amount of travel with AD*. As the requirements for AD travel differ from other modes, the *unrealized potential for travel* may also lead to new trips enabled by AD. *Amount of travel with AD* also depends on the *availability of AD mode* (ODD), whether the ADF is indeed available when driving within ODD (RQ-T1), and on the preferences of users (activation of system, RQ-U11).

Vehicle kilometres travelled (VKT) per road type with AD and other modes depends on overall *amount of travel* with corresponding modes and on the potential impacts on *preference for road type,* which enables use of AD (RQ-I6).

Impacts on *efficiency of driving style* result from differences between AD and human-driven vehicles in driving dynamics (RQ-T6), driven speed (RQ-T8), frequency of different driving scenarios (RQ-T11), interaction with other road users (RQ-T12), efficiency of traffic flow in intersections (RQ-T13), and in the behaviour of subsequent and preceding vehicles (RQ-T15 & RQ-T16). The efficiency of driving style affects the *breakdown capacity* of the road (traffic volume at which traffic changes from free flow to saturated flow). *VKT* (AD and other modes) and this *breakdown capacity* affect the *probability of congestion,* which is seen in impacts on *time spent in congestion* (RQ-I3).

Difference in *relative accident risk per driving scenario* (RQ-I1) between AD and other modes is affected by effects on the accuracy of driving (RQ-T7), driven speed (RQ-T8), and frequency of near crashes/incidents with and by other traffic participants (RQ-T10 & RQ-T14 & RQ-T17). It is also affected by *frequency of take-over control requests* (RQ-T2) *and other new potentially hazardous situations*. This results from how drivers respond when they are required to retake control (RQ-U10), how often and under which circumstances they choose to deactivate the system (RQ-U11), and from the effects that the take-over has on driving (RQ-T5).

Number of accidents per severity is the result of the effects on *accident risk per driving scenario* (RQ-I2, AD vs. other modes), *frequency of driving scenarios* (RQ-T11) and *severity of accidents* (RQ-I2), as well as of the effects on exposure (*VKT,* RQ-I5).

Greenness of driving style or energy demand per VKT (RQ-T9) results from differences between AD and human-driven vehicles in driving dynamics (RQ-T6) and driven speed (RQ-T8). This greenness, in addition to relative *electric vehicles penetration* among automated vehicles (AVs), affects the impact on *emissions per VKT* (RQ-I4, AD vs. other modes). These emissions per VKT combined with the corresponding VKT with AD and other modes lead to total *emissions* on the network.

Welfare gain (RQ-S1 & RQ-S2) is affected by impacts on the *number of accidents* (per severity) and *time spent in congestion* and *emissions.* All three affect the perceived benefits of potential users, linking them back up in the loop.

3 Experimental procedures

This chapter summarises the process for setting up the experimental procedures for L3Pilot, which was reported in more detail in L3Pilot Deliverable D3.2 by Penttinen et al. (2019). It also provides an update to the practical recommendations given to the pilot sites based on feedback one year after giving the original recommendations.

3.1 Setting up experimental procedure for the pilots

When designing the experimental procedures for a pilot study, one must keep in mind the difference between FOTs of close-to-market products and pilots of prototype systems as mentioned before in this deliverable.

L3Pilot's experimental procedure, which is presented in detail in L3Pilot Deliverable D3.2 (Penttinen et al., 2019), was developed to provide a solid base for the 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 of their implementation. Furthermore, the aim was to harmonise the evaluation criteria by providing detailed recommendations for the pilots with the intention to create holistic evaluation results of the L3Pilot project.

The experimental procedure work was started with an extensive review of available information related to various approaches for data collection, study participants selection, and detailed experimental design. Experimental design covers part of the FESTA-V 'Prepare' phase (Figure 1.3). Specifically, the following topics were addressed: which approach or approaches will be applied for data collection in the pilots (i.e. whether the approach will be subjective, such as surveys, or objective, such as an experimental study or some type of simulation study). Definition of participant types, which is closely related to the approaches – 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 experimental design encompasses several topics (within- versus between-participants design, pre and post measurements, definition of baselines, variable types) and, in practice, determines the framework for both data collection and analysis. The pros and cons of each of these, participant types, approaches, and experimental design, were analysed taking into account the goals and limitations of L3Pilot. Finally, at the end of the theoretical analysis, recommendations for each research question were listed.

To provide the needed practical support for pilots, on-site support visits were made to every pilot by the partner responsible for 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 on the preliminary plans, and several points critical to the success of piloting and evaluation were checked during the visits.

Finally, a set of practical recommendations was listed for the pilot sites to finalise their preparations for the pilot tests. These were reported in L3Pilot Deliverable D3.2 (Penttinen et al., 2019), and the following Chapter 4.2 of this report includes an update to the recommendations based on the feedback and remarks by the pilot sites when the pilots were already ongoing (February–March 2020).

3.2 Update of the recommendations with remarks from the pilots

The following chapter includes the updated lists of recommendations with some remarks from the pilots. In all, pilots appeared to agree that these recommendations are good goals, even if some of them were not fully applicable to all sites.

3.2.1 Test participants

A. *Use the best alternative class of test participants allowed by your company rules, country legislation, ethical aspects, and other limiting factors. The preference of test participants is in the order:*

   1) *Externals (ordinary drivers 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 in the pilot site questionnaires, based on indirect user experience (being on board and seeing ADFs in use).*

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

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

E. *The sample sizes should be close to 100 participants or preferably more per site (not if only professional safety drivers are used).*

The most challenging participant-related recommendations at this point are the recommendations to include (even as passengers) ordinary drivers into the pilot. The other challenging recommendations, which may be more realistic in the future when getting towards market-ready products, are the sample size and demographic factors. If only professional or highly trained drivers can be used, a large group with representative demographics cannot be achieved. This limitation was already recognised in the original recommendations. It should also be borne in mind that recommendation *C* may become obsolete in the future when targeting the customers of robot taxis.

3.2.2 Planning of test routes

G. *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.*

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

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

The recommendation of the route length was left quite open in this development phase. It was recognised that there are, for example, company set limits for individual driving time with the pilot vehicles, and this, naturally, limits the length of the test route. Longer driving times and hence routes are expected when entering the FOT phase later in the AD development.

3.2.3 Baseline data collection

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

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

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

O. *Baseline data should be collected within ODD.*

As seen above, the baseline data collection recommendations were slightly different for each environment, being strictest for urban and parking. A good additional remark from the pilots was related to longer drives (on motorways) and potential fatigue of the driver; if the driver drives for a long time or many times in a row, the order of the baseline and treatment drives should be alternated to make sure drivers are as alert on baseline drives as on treatment drives.

3.2.4 Metadata collection and pilot data

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

Q. *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.*

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

If the tests include a second person in the vehicle during the drive, then a tool with which that person can flag for example special events, and save needed metadata related to traffic, weather etc. was seen as useful and hence implemented in many pilot sites. Naturally, this approach is not applicable if there is only a single test driver in the vehicle and secondary tasks are not allowed.

3.2.5 Performing the tests – instructions

Performing the tests can be divided into several phases: 1) *First part of the pilot site questionnaire*: Can be completed in the recruitment phase. The same questionnaire is used independently of participants being drivers or passengers of the test vehicle. 2) *Instructions*: Test participants are given basic information on the test vehicle, ADF and its use (including ODD) and their task during the test drives/rides. The basics of the project aims are introduced as well.

S. *In spite of professional test driver groups, all test participants should act (imagine the situation) as if they had purchased a new automated vehicle and would be using it for their own trip. The test participants are advised to act as normally as possible, as they would do in that hypothetical situation – despite, of course, being aware of participating in a study. Accordingly, all unnecessary interactions with the test driver during the tests should be minimised even if there is another person in the vehicle.*

T. *The test participants are aware that 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.*

The information that needs to be given to the test drivers is naturally highly dependent on the driver type (professional test driver, …, ordinary driver). For members of the development team much less information is needed, whereas for ordinary drivers/passengers it is important to highlight that what they experience/see in this phase of piloting is not yet a final product.

3.2.6 Performing the tests – actual test drives

The actual test drives have two phases: 1) *Familiarisation with the vehicle*: In case the test participants are the drivers of the vehicle, they need to be given sufficient opportunity to familiarise themselves with the vehicle and systems before the data collection starts. The user interface and ADF will be explained before the pilot drive, as well as the other conditions during the drive (role of the safety driver, other people on board etc.). This familiarisation phase should not be used as a baseline. 2) *Actual test drives*: Data collection for both the baseline and the treatment.

U. *The task of test participants is to drive the route as indicated by the navigator. As soon as the vehicle indicates that AD is available, they can accept the AD mode. In case the AD option is not utilised by the test drivers, they can be encouraged to use it.*

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

If secondary tasks are allowed, a good remark is to instruct drivers to put, for example, their phone or tablet in such a place that they can safely reach it while driving. This should be done before starting the drive, not during the drive.

3.2.7 Questionnaires after the test

Second part of pilot site questionnaire: This is filled in right after the last test ride. The same questionnaire is used independently of participants being drivers or passengers of the test vehicle. Those test drivers who drive for significantly long periods may fill in the questionnaire periodically.

3.3 Remarks on experimental procedures in L3Pilot

In all, pilots appeared to agree that the given recommendations were good goals and guidelines, even if some of them were not fully applicable to all pilot sites (yet). The most challenging recommendation for practical implementation in the pilots concerned test participants. Everyone agreed on the recommendation from a scientific perspective. However, since the ADFs are still prototypes and their safety must be ensured, for most pilot sites the safety concept foresees the use of professional safety drivers. In this case, it is acknowledged that the demographic factors of this group do not necessarily reflect the driver population of the future customer or user base. It is highly appreciated that some pilot sites were allowed to recruit also ordinary drivers as test participants.

Another methodological challenge concerns the representativeness of the pilot tests with respect to the traffic environment. As there are only a small number of urban pilots in L3Pilot, and urban environments are so diverse, it is also acknowledged that the representativeness of urban ADF results may be limited. For the motorway environment, with more pilots and less variation in the road layout the representativeness of the results is expected to be better.

Overall, the pilot results should be taken as indicative, and the findings should be confirmed with a larger field operational test once the technology is more mature.

4 Evaluation plan for all evaluation areas

This chapter provides an evaluation plan for each research question together with a summary of and updates to the methods described in L3Pilot Deliverable D3.3 (Metz et al. 2019).

4.1 General assessment principles

4.1.1 Evaluated automated driving functions

Technical & traffic and user & acceptance evaluation

Technical & traffic evaluation and user & acceptance evaluation is based on the 20 different ADFs used in the field tests at the 16 pilot sites for driving on motorway, in urban environment or for parking (see L3Pilot Deliverable D4.1 by Griffon et al., 2019). To be able to analyse the overall impact of these functions, that is, independent of the individual implementations of specific vehicle brands, the ADFs will be grouped in the analysis and indicators will be presented such that they are merged across pilot sites.

For both evaluation areas, the grouping of the prototype implementations will be done based on the ODD of individual functions. A basic distinction is made between three different environments where the ADF is used: motorway, urban, and parking. In the urban environment, all ADFs are evaluated together as *urban ADF*. All ADFs that perform the parking are evaluated as *parking ADF*. In the motorway environment, there are two ODDs for the ADFs: *Traffic jam ADF* can drive in conditions where traffic density is high and speeds are below 60 km/h; while *motorway ADF* can drive also in free-flow conditions with ADF's maximum speed up to 130 km/h.

In the technical & traffic evaluation, the driving behaviour of the ADFs will be analysed. Separate analyses will be made for the traffic jam situations and for flow with higher speed on motorways. The grouping will be done independently of the full ODD of the tested function, always using data logged within a specific traffic condition. This means that data from a traffic jam ADF and a motorway ADF will be merged for parts driven in traffic jam conditions regardless of the system used. Driving behaviour is analysed separately also for the urban ADF and the parking behaviour for the parking ADF.

For the user & acceptance evaluation, the driver's evaluation of the tested ADFs as a whole is investigated. It is expected that the evaluation of a tested ADF is based more on the overall ODD and less on the behaviour in a single driving situation. Therefore, motorway ADF covering the full speed range and traffic jam ADF driving only in congested conditions will be differentiated, in addition to addressing separately the urban ADF and parking ADF.

In both technical & traffic and user & acceptance evaluations, the outcomes of individual pilots will be anonymously merged, which will allow generic results per ADF type to be presented to understand the overall implications of ADFs and not of individual implementations. Furthermore, the process developed for the merge guarantees data privacy and confidentiality within the project.

Impact assessment and socio-economic assessment

Impact assessment and socio-economic assessment do not address the prototype implementations that are tested in the technical & traffic and user & acceptance evaluations. The reason is that the impact assessment analyses the potential impacts of ADFs in perspective, when they are in use on a larger scale. It is expected that the ADFs will be developed further from the ones tested in L3Pilot. Therefore, so-called *mature functions* and their ODD are defined to represent such future ADFs.

The mature function descriptions were developed in cooperation with ADF developers. Thus, they take into account the knowledge within L3Pilot and represent ADFs that are considered mature enough to be used on roads by ordinary customers. It is important to note in this context that the defined mature ADFs do not represent any particular L3Pilot ADF tested at any of the pilot sites. Rather, they provide a generic description of what these ADFs may look like when adopted by users on a large scale.

All mature functions keep the vehicle in lane and hold a safe distance to vehicles in front. Lane changes can be performed automatically. All mature ADFs operate both in daylight and at night, and in good weather conditions or in light or normal rain. However, heavy rain, snow, fog and extreme weather conditions, as well as icy or snowy road surfaces, are outside their ODD. The ADFs can handle small gaps in lane markings. At the end of the ODD, a take-over request is sent to drivers and they are required to take control of the vehicle.

- *The mature motorway ADF* requires a road infrastructure that ensures a clear separation between opposing directions of traffic and visible lanes and road markings. The mature motorway ADF can start operating after the vehicle has merged onto the motorway and ends before or when the vehicle leaves the motorway. The ADF can handle weaving without ramps. Driving in toll station areas is not included.
  - *The mature traffic jam ADF* is a sub-function of the motorway ADF and works in the same ODD, except that it operates only in congested (traffic jam) conditions and reaches speeds of up to 60 km/h. The mature motorway ADF function also includes the mature traffic jam function. They are treated separately because they are expected to become available on different time horizons.
- *The mature urban ADF* operates on urban roads at speeds of up to 50 km/h. It requires lane markings or clear curbs on both sides of the lane. A form of markings is also needed for handling street-side parking, bicycle lanes etc. – either as lane markings or clearly defined on a high definition (HD) map used by the vehicle. It is expected that HD maps are especially important for the deployment of early generations of the mature urban ADF. These maps also enable the vehicle to be rerouted in particular conditions, for example, if roadworks are detected on the planned route. The ADF can drive in signalised and non-signalised intersections. It can handle simple roundabouts; take-over is requested only for the complex ones.
- *The mature automated parking ADF* has two functionalities: home zone parking and public parking. Home zone parking is intended for use on private parking grounds, while public parking

covers parallel and perpendicular parking in public parking spaces such as street-side parking and parking lots. Both functionalities can deal with static objects and other slow-moving traffic participants.

- *The mature home zone parking* can handle the actual parking manoeuvres and drive on the private driveway to reach the parking spot. This functionality requires pre-training of the trajectory from the driveway entrance to the dedicated parking spot. In home zone parking, the user can be outside the vehicle and is not required to monitor the manoeuvres.
- *The mature public parking* functionality requires the driver to be inside the vehicle and monitor the parking manoeuvre. Markings or parked cars are needed to indicate the available parking space. The ADF does not require pre-training of the trajectory.

4.1.2 Assessment scenarios

To ensure that all areas of analysis are harmonised and build on each other, a common understanding is needed of the basic methodological principles chosen in L3Pilot. The basic approach of how driving will be analysed and described is common to all methods. These common definitions ensure that methods set for generalising and upscaling of the effects can build on the results derived from the experiments.

To capture the influence of ADF on driving behaviour and its potential effects on traffic safety, traffic efficiency, and the environment, comparable sections of driving with and without ADF need to be examined. In controlled experimental approaches, this comparability is achieved by experimental protocol through which moderating factors are controlled, and unwanted situational variance is minimised. In the on-road tests of L3Pilot, this approach is difficult because of uncontrollable variation in on-road situations. Furthermore, there will be variations between pilot sites in the driving environment and test setup.

To come to reliable results, a control of moderating factors is planned by analysing the logged on-road data by *driving scenarios*. A driving scenario is a short period of driving defined by its main driving task (e.g. car following, lane change) or triggered by an event (e.g. an obstacle in the lane). A *driving situation* represents a single segment in time that is assigned to a certain driving scenario. Driving situations within different driving scenarios differ fundamentally, whereas situations of the same driving scenarios are similar. Furthermore, for impact assessment *traffic scenarios* are used in addition to the driving scenarios. These scenarios have a broader horizon than the driving scenarios and cover a specific road section with certain traffic characteristics. Table 4.1 provides a summary of the definitions used for different scenario types.

Table 4.1: Definition of the different types of scenarios used in L3Pilot (updated from definitions in L3Pilot Deliverable D3.3 by Metz et al., 2019)

	Definition
Driving scenario	Driving scenarios describe the development of a situation within a traffic context in which at least one actor performs a (pre-)defined action and/or the driving scenario is triggered by a (predefined) event. The action or event is specified without the definition of concrete parameters. The influenced actor may either be the ego vehicle (e.g. performing a lane change or a minimum risk manoeuvre) or another traffic participant (e.g. a lane change in front of the ego vehicle). The event triggering the driving scenario can be a change in road infrastructure (e.g. an end of lane or a change in speed limit) or an external obstruction (e.g. an obstacle on the road). *An example: a lane change.*
Driving situation	A driving situation is a specific instance of a driving scenario (e.g. a lane change) but with specific parameters. Thus, a driving situation describes in detail a situation that can be simulated and analysed. *An example: a lane change at 60.8 km/h with a second vehicle driving at a distance of 10 m behind the ego vehicle in the adjacent lane and with a velocity of 65.0 km/h.*
Traffic scenario	Traffic scenarios describe a larger traffic context by covering a longer period of time and longer road sections with certain traffic characteristics. One traffic scenario may include different (not predefined) driving scenarios. *An example: a 3-lane motorway section of length 10 km with 2 motorway entrances and exits, a speed limit of 130 km/h, traffic volume of 4 000 vehicles/h/direction, 10% of heavy vehicles and a time period of 1 hour.*

Driving scenarios are the basic unit of analysis concerning driving behaviour (see L3Pilot Deliverable D3.3 by Metz et al., 2019 for details). A list of driving scenarios is created aiming to cover driving on motorways and in urban areas and avoiding overlaps. Within these scenarios, the behaviour in baseline and treatment is compared. Within the scope of the impact assessment, additional driving scenarios are considered, focusing on certain events which only happen during AD or transition phases.

In technical & traffic evaluation, all time-series data logged during the on-road tests are divided into *driving situations*, which all belong to one of the defined *driving scenarios*. Multiple driving situations of one driving scenario can occur within a single log of driving data. Performance indicators are defined for each driving scenario, and they describe driving behaviour in the scenario in a meaningful way. Indicators are calculated for every driving situation identified in the data. By comparing the values of the performance indicators in baseline and in treatment for all situations of a driving scenario, the impact of ADF on driving behaviour in the defined driving scenario is derived. In the impact assessment, driving scenarios and related results are used to set up simulation models.

Traffic scenarios are relevant for impact assessment. Within a traffic scenario, several driving scenarios can occur in different combinations. Traffic scenarios have a wider focus than driving scenarios and cover a certain road section and the traffic on it. Except for the general characteristics, the traffic is not predefined but is a result of the individual behaviour of the traffic participants within the traffic scenario. Traffic scenarios include both road infrastructure characteristics, such as number of lanes, speed limits, motorway entrances and exits, and traffic characteristics, such as penetration rate of the ADF, fleet composition and traffic volume.

4.1.3 Baseline and treatment

For evaluation purposes, a comparison is required for the situation before the introduction of the ADF in focus (called *baseline* in the following) with the situation after their introduction (called *treatment*). In other words, to assess the potential effects, driving with the ADF is compared to reference driving behaviour in order to derive how the ADF changes driving behaviour, traffic safety, traffic efficiency, etc. Therefore, not only driving with the ADFs (treatment) needs to be described; the baseline needs to be defined as well for driving without the respective ADF.

In theory, there are several options of how a baseline scenario could be defined. In L3Pilot, the baseline will be based on *an approximation of the traffic today,* since the requirements for the use of e.g. a future ADAS baseline would be beyond our possibilities. It is also more reliable to base the assessment on existing statistics and not to increase the uncertainty by also attempting to predict the future baseline situation. Treatment scenarios are created for how the situation would change if a proportion of the vehicles in the world of today had been operated with ADFs. In this way, it will be possible to capture the pure impacts of implementing ADFs. We call this a "snapshot" approach (see Chapter 4.7.1).

In practice, the relevance of different systems and the feasibility of making the assessment lead to slightly different baseline definitions for each impact area. Therefore, depending on the area of analysis within L3Pilot, the baseline to be used varies. A summary over all evaluation areas is given in Table 4.2. Specification of different penetration rates for safety impact assessment is given in Table 4.3. Current vehicle stock penetration rates for autonomous emergency braking (AEB) and forward collision warning (FCW) were estimated based on Eurostat statistics on the vehicle stock and new passenger car registrations (Eurostat 2018a & 2018b), Öörni's (2016) estimates on the EU28 penetration rates in 2015, and Frost & Sullivan's (2018) estimates on the share of newly sold vehicles equipped with different systems in Western Europe in 2017 (see details in L3Pilot Deliverable D3.3 by Metz et al. 2019). We acknowledge that the penetration rates of different ADAS will increase in parallel with the Level 3 ADFs. However, in line with the snapshot principle of not trying to predict the future and the implications of all parallel changes in transport system and travel behaviour, we simplify the situation for having the same share of non-ADF vehicles equipped with ADAS in the treatment scenario as in the baseline scenario. All ADF vehicles are assumed to have also AEB and FCW (available everywhere).

Table 4.2: Overview of baseline and treatment used in different areas of evaluation.

Evaluation area	Baseline	Treatment
Technical & traffic	Manual driving (SAE 0) (in ODD)	SAE L3/L4 (in ODD)
User & acceptance	Current situation (users' own car)	SAE L3
Safety impact assessment	Manual driving without active safety systems Manual driving with active safety systems AEB and FCW	Variation of SAE L3 penetration rate (in ODD) and active safety systems (everywhere)
Efficiency and environmental impact assessment	Manual driving (in ODD)	Variation of SAE L3 penetration rate (in ODD), equal to the ones in safety impact assessment
Socio-economic impact assessment	Today's situation (Latest statistics)	Variation of SAE L3 penetration rate (in ODD), equal to the ones in safety impact assessment

Table 4.3: Penetration rates for safety impact assessment.

Name	Penetration rate for L3Pilot mature ADF	Penetration rate for AEB and FCW
Baseline 1	0%	0%
Baseline 2	0%	7.3%
Treatment 1	5%	12.3% (5% ADF + 7.3% baseline)
Treatment 2	10%	17.3% (10% ADF + 7.3% baseline)
Treatment 3	30%	37.3% (30% ADF + 7.3% baseline)
Treatment 4	100%	100% (100% ADF)

4.2 Technical and traffic evaluation

4.2.1 Overall approach

The objective of the technical & traffic evaluation in L3Pilot is to answer research questions related to the technical performance and the effects resulting from an automated system's behaviour within traffic (Table 2.1). In this chapter, an overview is given on how those research questions are answered quantitatively considering relevant measures.

Different effects are evaluated by comparing performance indicators. These indicators are numerical values which can be compared between baseline and treatment. In order to derive performance indicators from the signals gathered during the pilot, a first step is to determine derived measures, which will then be aggregated to indicators. Performance indicators are scalar values that can be evaluated and compared between the two conditions. Values considered may be statistical parameters such as minima, maxima, mean or standard deviation. Furthermore, the frequency of certain events per time or distance is compared.

A distinction is made between performance indicators which are evaluated per trip, such as the frequency of certain events or scenarios. Other indicators are evaluated per driving situation. In this case, a trip will generate multiple values per performance indicator to be considered for evaluation due to instances of the driving scenario appearing multiple times during a trip. For all performance indicators, a mapping between the indicators and driving scenarios has been generated, showing which performance indicators deliver meaningful values in which scenarios (see the mapping in L3Pilot Deliverable D3.3 by Metz et al., 2019).

In the following, the performance indicators are listed per research question, structured by the top-level research questions. The corresponding tables show whether the merging in the consolidated database is done for the given performance indicator per trip or per situation. All performance indicators listed below in Chapter 4.2.2 will be applied to motorway, traffic jam and urban ADFs. For urban ADFs, additional performance indicators will be considered. They are described in Chapter 4.2.3.

4.2.2 Evaluation plan for motorway, traffic jam and urban ADFs

<u>RQ-T1 – RQ-T2: What is the ADF's technical performance?</u>

To evaluate the technical performance of the ADFs, three main indicators are considered. The percentage of time with system available gives an indication of how much of a trip can be covered by the ADF. The duration of sections with ADF available indicates how consistently the system can be operated in automated mode. Furthermore, the number of take-over requests could be evaluated, indicating how consistent the ODD of the systems is and whether take-over requests are issued when required. However, the pre-series status of the systems needs to be considered. Reasons why systems cannot be activated or issue a take-over request may be due to study design, legal requirements or properties of the environment. All performance indicators considered for the evaluation of the system's technical performance are listed in Table 4.4.

Table 4.4: *Performance indicators evaluated for research questions RQ-T1 & RQ-T2.*

RQ-ID	Performance indicator	Unit	Merged per
RQ-T1	%time (ADF available)	-	Trip
	duration of sections with ADF available	min	Section with ADF available
RQ-T2	number(take-over requests) / h	1/h	Trip
	number(planned take-over requests) / number(total take-over requests)	-	Trip

RQ-T5 – RQ-T11: What is the impact on the driving behaviour of the ADF-vehicle?

Performance indicators describing the driving behaviour of the ADF are evaluated by means of the research questions RQ-T5 to RQ-T11. For this, longitudinal and lateral accelerations (a_x, a_y) as well as velocity (v) of the vehicle are considered. Performance indicators derived from these signals are the minima (min), maxima (max), mean and standard deviation (sd). Furthermore, the number of standstill events is analysed and defined as sections in which the velocity is below 0.2 km/h. An overview of all the measures to be merged between the pilot sites is given in Table 4.5.

Table 4.5: Performance indicators evaluated for research questions RQ-T5 – RQ-T11.

RQ-ID	Performance indicator	Unit	Merged per
RQ-T5	min(a_x)	m/s²	Situation
	max(a_x)	m/s²	Situation
	sd(a_x)	m/s²	Situation
RQ-T6	max(abs(a_y))	m/s²	Situation
	sd(a_y)	m/s²	Situation
RQ-T8	mean(v)	km/h	Situation
	max(v)	km/h	Situation
	sd(v)	km/h	Situation
RQ-T8, RQ-T9	number(v<0.2km/h) / h	1/h	Trip
	mean(duration of sections with (v < 0.2km/h))	s	Trip
RQ-T7	sd(position in lane)	m	Situation
	mean(position in lane)	m	Situation
RQ-T9	mean(energy consumption) / 100 km	kWh/100km	Trip
RQ-T10	number(harsh braking events) / h	1/h	Trip
RQ-T11	number(driving scenario: X) / h	1/h	Trip
	time(in driving scenario) / time(in ODD)	-	Trip

The impact on energy efficiency cannot appropriately be evaluated by looking at fuel consumption or consumption of energy stored in the battery during the on-road pilots. Furthermore, the pre-series equipment for operating the function may have higher energy consumption than what is to be expected for vehicles in series production, which cannot be told apart from the effects resulting from the differences in behaviour between a human driver and an ADF. In Rösener, Sauerbier, Zlocki, Eckstein, Hennecke, Kemper, et al. (2019), an approach considering the energy demand has been applied to assess the effect of AD in terms of energy efficiency. The approach considers that the fuel consumption is proportional to the energy demand per driven distance (neglecting different efficiencies of operating states of an internal combustion engine). Thus, the energy demand E_{demand} per distance s is taken as a surrogate measure for assessing the impact of ADF on energy efficiency. The total resistance force F_{total} considers the air resistance F_{air}, the rolling resistance F_{roll} and the resistance from acceleration F_{accel}. Resistance resulting from gradients F_{grad} can be neglected, since due to study designs it cannot always be guaranteed that the elevation profile will be the same for baseline and treatment and it may introduce greater errors.

$$\frac{E_{demand}}{s} = F_{total} = F_{air} + F_{roll} + F_{accel} + F_{grad}$$

with:

$$F_{air} = 0,5 \cdot c_w \cdot A \cdot \rho_{air} \cdot v^2$$

$$F_{roll} = m_{veh} \cdot g$$

$$F_{accel} = e_i \cdot m_{veh} \cdot a \quad \forall\, a > 0, else\ F_{accel} = 0$$

The vehicle parameters air resistance c_w, cross section A, vehicle mass m_{veh} and mass factor e_i are set to be in line with the definition of the mature pilot functions. The fact that for negative accelerations the resulting resistance force is set to zero is a simplification of overall energy conversion processes in the vehicle. If the acceleration is negative, the resistance would be negative as well, thus generating power that could be fed into the vehicle again. This energy can, however, never be reused by the vehicle to a full extent. In the case of braking, most energy is dissipated via the brakes. Even for the battery of an electric vehicle, the amount of energy recuperated is comparatively small. Thus, the equations above picture a reasonable simplification of a vehicle's energy consumption and conversion processes. Differences in energy consumption between treatment and baseline may result from slower overall speeds resulting in a smaller air resistance, from shorter or smoother stretches of acceleration, or less braking due to a more anticipatory driving style.

RQ-T12: What is the impact of ADF on the interaction with other road users?

To assess the effect of AD on the interaction with other road users, typical performance indicators considered are the time headway (THW) and time-to-collision (TTC). For scenarios with a low overall velocity, the distance to the vehicle in front is also considered a relevant measure. Performance indicators evaluated per research question are listed in Table 4.6.

Table 4.6: Performance indicators evaluated for research question RQ-T12.

RQ-ID	Performance indicator	Unit	Merged per
RQ-T12	mean(THW)	s	Situation
	sd(THW)	s	Situation
	min(THW)	s	Situation
	min(TTC)	s	Situation
	mean(long. dist $_{\text{Lead Vehicle}}$)	m	Situation
	sd(long. dist $_{\text{Lead Vehicle}}$)	m	Situation

RQ-T14: What is the impact of ADF on the number of near crashes / incidents with other road users?

To evaluate the research question RQ-T14, the incident criteria are evaluated for the different pilot sites. As described in L3Pilot Deliverable D3.3 (Metz et al., 2019), there are four classes of incidents which characterise the incident, which may further be divided into sub-classes:

- Distance-based incidents with sub-categories
  - Front incidents
  - Side incidents
  - Rear incidents
- Vehicle dynamic-based incidents with sub-categories
  - Incidents due to longitudinal acceleration
  - Incidents due to lateral acceleration or yaw rate
- Lane and road departures

Furthermore, there are three levels of incidents classifying their severity: (1) increased risk, (2) crash-relevant situation, and (3) near crash.

Considering the likelihood of a crash on a motorway and the overall driven distance in L3Pilot, crashes are not expected to occur during piloting operations. Therefore, the frequency of the combinations of the different incident types and severity levels are evaluated. In order to verify that the automatically detected incidents are in fact incidents, video validation will be applied for all incidents detected.

RQ-T15 – RQ-T17: What is the impact of ADF on the behaviour of other traffic participants?

To evaluate the effect of the ADF on other traffic participants, similar performance indicators as for evaluating the behaviour of the ADF are considered for the preceding and subsequent vehicle. Furthermore, the frequencies of cut-ins by other vehicles and the frequency of a vehicle changing lanes in front of the ego vehicle are evaluated as listed in Table 4.7.

Table 4.7: Performance indicators evaluated for research questions RQ-T15 - RQ-T17

RQ-ID	Performance indicator	Unit	Merged per
RQ-T16	number(cut in) / h	1/h	Trip
	number(lane change $_{lead\ vehicle}$) / h	-	Trip
	mean($v_{lead\ vehicle}$)	km/h	Situation
	sd($v_{lead\ vehicle}$)	km/h	Situation
RQ-T15	mean($THW_{rear\ vehicle}$)	s	Situation
	sd($THW_{rear\ vehicle}$)	s	Situation
	min($THW_{rear\ vehicle}$)	s	Situation
	min($a_{x\ rear\ vehicle}$)	m/s²	Situation
RQ-T17	number($a_{x\ rear\ vehicle}$ < threshold) / h	-	Trip
	number($THW_{rear\ vehicle}$ < threshold) / h	-	Trip
	number($TTC_{rear\ vehicle}$ < threshold) / h	-	Trip
	number(long. distance $_{rear\ vehicle}$ < threshold) / h	-	Trip

4.2.3 Evaluation plan, additional indicators for urban ADFs

Pilot data

Driving automation in the urban domain can be seen as a more complex challenge compared to motorways. Some reasons for this are that urban areas include more various types of road users – especially VRUs – and level crossings, which creates more factors to be considered when planning and executing manoeuvres in the urban domain. This also requires additional considerations to enable a successful evaluation of urban ADFs.

Performance indicators and driving scenarios considered for ADFs on motorways largely represent a subset of performance indicators to be considered for urban driving. *Free driving*, *following a lead vehicle*, *approaching a lead vehicle*, *lane change* and *cut-ins* are scenarios taking place in both environments. However, whereas for motorways approaching or following a vehicle slower than 60 km/h is considered approaching a traffic jam or driving in a traffic jam, this condition cannot be used for urban environments. For the evaluation in L3Pilot, only generic *following a lead object* and *approaching a lead object* are considered instead. Furthermore, intersection scenarios and manoeuvres are to be considered for the following research questions:

RQ-T11: What is the frequency of defined driving scenarios?

Whereas for motorway functions the denominator to evaluate the frequency is typically time or distance since the vehicle is following a single road most of the time, in an urban environment for the driving scenarios at intersections, the manoeuvres *crossing*, *left turn* and *right turn* as well as the related scenarios describing interactions with other traffic participants (e.g. *turning with laterally moving object*) need to be considered. The number of manoeuvres taking place during a trip varies per pilot site due to the chosen route, and, therefore, also the occurrence of the various intersection scenarios per time is dependent on the route. Thus, to evaluate the frequency of certain driving scenarios, the number of manoeuvres is to be used as the denominator. The frequency (*f*) of the driving scenario *right turn with lead object* is thus calculated as:

$$f(right\ turn\ with\ lead\ object) = \frac{n(right\ turn\ with\ lead\ object)}{n(right\ turn)}$$

All relevant combinations between the manoeuvre at an intersection and the related driving scenarios are given in Table 4.8.

Table 4.8: Additional performance indicators for urban ADF evaluated for research question RQ-T11.

RQ-ID	Performance Indicator	Unit	Merged per
RQ-T11	number(approaching laterally moving object) / h	-	Trip
	number(approaching laterally moving VRU) / h	-	Trip
	number(crossing with laterally moving non-VRU object) / number(crossing)	-	Trip
	number(crossing with static object) / number(crossing)	-	Trip
	number(crossing with lead object) / number(crossing)	-	Trip
	number(crossing with laterally moving VRU) / number(crossing)	-	Trip
	number(turning with static object) / number(turning)	-	Trip
	number(turning with lead object) / number(turning)	-	Trip
	number(turning with laterally moving object) / number(turning)	-	Trip
	number(turning with laterally moving VRU) / number(turning)	-	Trip
	number(turning right with static object) / number(turning right)	-	Trip
	number(turning right with lead object) / number(turning right)	-	Trip
	number(turning right laterally moving object) / number(turning right)	-	Trip
	number(turning right laterally moving VRU) / number(turning right)	-	Trip
	number(turning left with static object) / number(turning left)	-	Trip
	number(turning left with lead object) / number(turning left)	-	Trip
	number(turning left laterally moving object) / number(turning left)	-	Trip
	number(turning left laterally moving VRU) / number(turning left)	-	Trip

RQ-T12: What is the impact of ADF on the interaction with other road users in a defined driving scenario?

For urban ADF, it is necessary to make a distinction between the different object classes. The research questions differentiate between vehicles and VRUs. The distinction is made by means of the object classes considered in the common data format (Hiller et al., 2019) as listed in Table 4.9. The resulting performance indicators to be compared are listed in Table 4.10.

Table 4.9: Classification of object types in common data format for definition of VRUs.

Object type	Traffic participants included
Vehicle	• Car • Truck
VRU	• Motorcycle • Bicycle • Pedestrian
Not considered	• Not applicable • Other • Unknown

Table 4.10: Performance indicators evaluated for research question RQ-T12.

RQ-ID	Performance Indicator	Unit	Merged per
RQ-T12	min(distance $_{VRU}$ per turning, crossing & approaching)	m	Situation
	min(TTC $_{VRU}$ per crossing & approaching)	s	Situation
	min(distance $_{Non\text{-}VRU\ object}$ per turning, crossing & approaching)	m	Situation
	min(TTC $_{VRU}$ per crossing & approaching)	s	Situation

A further measure to be evaluated separately for urban ADFs is making a distinction between incidents with vehicles and VRUs, which relates to research question RQ-T14 (see Chapter 4.2.2). The overall evaluation procedure is similar to the evaluation of motorway functions. Parameters chosen to detect incidents are adjusted to traffic in the urban domain.

AIM Mobile Unit data

The Application Platform for Intelligent Mobility - AIM Mobile Units as described in L3Pilot Deliverable D3.3 (Metz et al. 2019) allow for a different evaluation approach of certain research questions in terms of performance indicators, which are more suitable to be recorded from a bird's eye point of view on an intersection. For this reason, the AIM Mobile Units are used to analyse behaviour at an intersection, which may be incorporated in automated vehicle development, and to compare automated vehicle behaviour with manual driving.

In order to evaluate the effects between baseline and AD, every automated vehicle trajectory is matched with two trajectories of manual driving. For each automated vehicle trajectory, two baseline trajectories will be extracted within a time window of 20 minutes (i.e. within 10 minutes before and 10 minutes after the automated vehicle passed through the intersection). The rationale for this is to ensure matching on external conditions (e.g. time of day, weather condition, traffic density, etc.).

Many performance indicators evaluated using vehicle data from the pilot sites can also be analysed by means of the AIM intersection. These indicators are listed in Table 4.11. Additional performance indicators that can typically only be studied from a bird's eye point of view are listed in Table 4.12. Typical performance indicators to be analysed are post-encroachment time, gap time, and time to accident. Vehicle dynamics performance indicators can be analysed similarly considering their mean, standard deviation, and extrema.

Table 4.11: Performance indicators evaluated using vehicle data that can be covered by AIM.

RQ-ID	Performance indicator	Unit	Merged per
RQ-T5	min(a_x)	m/s²	Situation
	max(a_x)	m/s²	Situation
	sd(a_x)	m/s²	Situation
RQ-T6	max(abs(a_y))	m/s²	Situation
	sd(a_y)	m/s²	Situation
RQ-T8	mean(v)	km/h	Situation
	max(v)	km/h	Situation
	sd(v)	km/h	Situation
RQ-T8, RQ-T9	number(v<0.2km/h) / h	1/h	Trip
RQ-T7	sd(position in lane)	m	Situation
	mean(position in lane)	m	Situation
RQ-T12	min(TTC)	s	Situation
	min(distance $_{\text{VRU}}$ per turning, crossing & approaching)	m	Situation
	min(TTC $_{\text{VRU}}$ per crossing & approaching)	s	Situation
	min(distance $_{\text{Non-VRU object}}$ per turning, crossing & approaching)	m	Situation
	min(TTC $_{\text{VRU}}$ per crossing & approaching)	s	Situation

Table 4.12: Additional performance indicators evaluated with the AIM method.

RQ-ID	RQ Level 3	Performance indicator
RQ-T7	What is the impact of ADF on precision of manoeuvres?	• sd to predefined points of the manoeuvre
	What is the impact of ADF on lane keeping performance in defined driving situations?	• sd to fitted trajectory
RQ-T12	What is the impact of ADF on the behaviour of VRUs (cyclist, motorcyclist, pedestrian) in defined driving scenarios?	• post-encroachment time • gap time • time to accident
	What is the impact of ADF on the behaviour of approaching / crossing pedestrians?	• mean(v) • min(v) • max(v) • sd(v) • distribution(v) • mean(a_y) • min(a_y) • max(a_y) • sd(a_y) • distribution(a_y)
	What is the impact of ADF on car-following behaviour?	• temporal and spatial distribution of distances
RQ-T13	What is the impact of ADF on traffic flow at intersections?	• loss time • q [veh/min]
RQ-T14	What is the impact of ADF on the frequency of near crashes with other vehicles?	• post encroachment time • min(TTC)
	What is the impact of ADF on near crashes with VRUs?	• post encroachment time • min(TTC)
RQ-T15	What is the impact of ADF on the following behaviour of subsequent vehicles?	• post encroachment time • min(TTC)

Research question RQ-T13 – “What is the impact of ADF on traffic flow at intersections?” – can only be studied using the bird’s eye view of the entire intersection, since it is necessary to consider the behaviour of multiple subsequent vehicles. Applicable performance indicators are time loss and number of vehicles passing the intersection per minute.

4.2.4 Evaluation plan for parking ADFs

Parking ADFs require different evaluation approaches than the other ADF types evaluated in L3Pilot. Whereas many of the research questions defined on a project level can also be applied to parking ADFs, different driving scenarios and performance indicators need to be considered.

Driving scenarios in which the systems operate can be divided into the following:

- Trajectory learning
- Approaching a parking spot
- The parking manoeuvre
- Emergency manoeuvres

Take-over requests are considered in a similar manner as for motorway and urban ADFs. Similarly to motorway and urban ADFs, those events will be analysed to evaluate the technical performance of the system by means of their frequencies. Furthermore, the frequencies of successful manoeuvres will be analysed as a means of assessing technical performance.

Further insight into how the system operates technically, as well as velocities and accelerations in the longitudinal and lateral directions, will be analysed considering their extrema and mean and standard deviations (sd). Table 4.13 gives an overview of which research question is analysed for which scenario, and the applicable performance indicators are listed. *Trajectory learning* is a scenario that will not be analysed by means of performance indicators. However, a qualitative analysis per pilot site may be feasible.

Table 4.13: Research questions and performance indicators evaluated for parking ADFs per driving scenario.

RQ-ID	RQ Level 3	Entire ODD	Approaching parking spot	Parking manoeuvre	Emergency manoeuvre	Expected take-over request	Unexpected take-over request	Performance indicators
RQ-T1	How often is the ADF available in the driving and traffic scenarios of its ODD?	x	(x)	(x)				• number(manoeuvre with final parking position) / number(manoeuvre with ADF active)
RQ-T2	Does the ADF initiate a take-over request if required by the boundaries of the ADF?				x	x		• number(take-over request) / number(parking manoeuvre) • number(ODDLim_TOR) / number(ODD_Lim)
RQ-T5	How do planned take-over situations affect the driving dynamics of the vehicle?				x	x	x	• detailed results from TOC rating
RQ-T6	What is the impact of ADF on longitudinal acceleration in defined driving situations?	x	x	x				• min(a_x) • max(a_x) • sd(a_x)
	What is the impact of ADF on lateral acceleration in defined driving situations?	x	x	x				• max(abs(a_y)) • sd(a_y)
RQ-T7	What is the impact of ADF on precision of manoeuvres?	x		x				• mean(manoeuvre duration) • evaluation of parking position by experimenter • distribution(chosen position, desired position)
RQ-T8	What is the impact of ADF on driven velocity in defined driving situations?	x	x	x				• mean(v) • max(v) • sd(v)

4.2.5 Evaluation procedure

The first step of the technical & traffic evaluation is the calculation of derived measures and performance indicators, which is done by research partners selected by the vehicle owners (*selected partners*). The calculation of such measures is done by running the scripts developed in L3Pilot. The supply of data to the selected partner is typically dependent on the requirements of the vehicle owners. Data is delivered using the common data format.

The calculation of the performance indicators is done as a two-staged-process: In a first step, all parameters that can be calculated directly are determined. This also includes applying the detection of the driving scenarios. Scenario detection is implemented in a semi-automated manner: the output of the scenario detection marks scenarios for which the driving scenario type cannot be determined with sufficient certainty. These scenarios are then checked manually in the second step and labelled with the correct driving scenario type. When all scenarios have been determined, scripts to calculate the remaining derived measures and performance indicators are run.

It is expected that algorithms for detecting the scenarios may be improved over the course of validating scenarios manually. This will allow reducing the number of scenarios to be labelled manually. Improvements to the toolchain will be made available to all pilot sites, using a shared version control system. The output of the work per selected partner is a list of performance indicators per trip per pilot sites.

The output of the data processing toolchain will consist of several performance indicators per trip, which are accompanied by metadata about the system state:

- ADF active
- ADF available (but not active)
- ADF not available (but not baseline condition)
- Baseline

This structuring allows consideration of different study designs, where some pilot sites will have professional drivers in the driver's seat and others will have ordinary drivers in the passenger seat. Performance indicators with the system information are then processed by *aggregation scripts*, which collect the relevant data per trip file and structures the data in a form suitable for upload to the consolidated database.

The consolidated database is the central source of data for analysis. Analysts in the evaluation work packages have access to data stored in the database. However, they are not able to detect which pilot site the data belongs to. Primary results of the analysis will be generated by analysing the entire contents of the database per performance indicator. Some performance indicators may not be delivered by certain pilot sites – either because the system design does not allow recording signals needed for certain performance indicators in a decent enough quality, or because the study design does not allow gathering results for some pilot sites. For instance, some pilot sites will let their system execute protected lane changes guided by a vehicle driving behind the ego vehicle, ensuring that no fast-approaching cars are too close in the target lane.

The contents of the consolidated database can be presented online in table format for quick checks. For further analysis, the data are downloaded from the consolidated database to be analysed in detail and processed for presentation in the final deliverables. Analysis tools used in this processing step are typically SSPS and MATLAB.

The preferred format for presenting the results is histograms. Each driving situation or trip will yield one value to be assigned to a bin of the histogram. Distributions for baseline and treatment are represented in the same diagram, highlighting the detailed differences between human and ADF.

Besides the graphical representation, a statistical hypothesis test is carried out per performance indicator (alpha level 5%) to test for effects of the treatment condition. The automated system (treatment) and manual driving as baseline are considered as two independent groups. The hypotheses per research question are listed in L3Pilot Deliverable D3.1 (Hibberd et al., 2018). The typical null hypothesis is that there is no difference in the considered performance indicator between treatment and baseline. Where applicable, parametric t-tests (in case of two independent means) are to be applied. Non-parametric tests such as the Mann-Whitney U test are considered if certain properties of the available data do not support the use of parametric methods, although the expected sample size should ensure robust performance of the parametric procedures.

With large sample sizes, as are to be expected with piloting data, hypothesis tests may almost always state a significant effect; the meaningfulness of such a value needs to be further evaluated. Therefore, the effect size d of Coe (2002) will be given along with the result of the hypothesis test, which allows for meaningful interpretation of the effect between baseline and treatment.

$$d = \frac{\mu_{treatment} - \mu_{baseline}}{\sqrt{\frac{\sigma_{treatment}{}^2 + \sigma_{baseline}{}^2}{2}}}$$

The effect size may also serve as a means for comparing performance indicators that are not suitable to be merged as situations, since it is possible to present the effect meaningfully without displaying performance indicators which are not feasible for sharing.

During the phase before data delivery from the pilots, the data processing and analysis tools are tested in depth. The central aim of the testing and improvement is that the developed toolchain works equally well for all pilot sites and the way the results are generated are consistent across all pilot sites. Improvements to the developed software tools will be tracked and distributed among all analysis partners via a shared code repository. Individual adjustments to issues that only affect a single pilot site may be tracked and checked for consistency using code branches.

In order to test the entire toolchain for evaluation before the end of the pilot, two main test runs are executed. This will ensure that the data evaluation runs smoothly and that the results generated are meaningful, while also ensuring that testing of the toolchain is executed with the same data by multiple evaluation partners. Prior to a testing loop, data that has been processed by the selected analysis partners will be uploaded to the consolidated database in order to have the largest possible amount of data available for testing.

For the first testing loop, evaluation partners and selected partners focus on performance indicators which are not dependent on the detection of complex driving scenarios. Thus, performance indicators for driving scenarios *following a lead object* and *free drivin*g will be considered. Furthermore, the scenario and incident detection scripts are tested in detail and improved. The second testing loop will focus on the detection of all driving scenarios and detection of incidents, as well as on the derivation of performance indicators from driving scenarios.

After the end of the piloting phase, the remaining recorded data is delivered to the selected partners. The data is then processed and possible minor changes to the toolchain are made if necessary. Performance indicators from the time series data are then uploaded to the consolidated database and analysed on a consortium level. The visualisation and interpretation of the results will be distributed among analysis partners. Finally, the results will be discussed during multiple workshops among all analysis partners and pilot sites.

4.3 User and acceptance evaluation

4.3.1 Overall approach

The L3Pilot project uses a multifaceted user & acceptance evaluation approach to form a holistic view of users' behaviours with, and acceptance of, the ADFs. The design of this approach was underpinned by a list of priority research questions (Table 2.2). The research questions were organised into a number of key themes, including user acceptance and trust of the systems, willingness to use and pay for the functionalities, measures of driver state (stress, distraction, fatigue, workload), user risk perception, driver engagement in non-driving related tasks, user behaviour during and after take-over situations, and user motion sickness. These were translated into 11 research questions (Level 2), which were then developed into 21 detailed research questions (Level 3).

The project has sought to employ a number of quantitative and qualitative data collection methods to provide data to answer these research questions. These methods are centred primarily around pilot studies, including pilot site questionnaires, videos of the driving scene, recordings of drivers' head, hands, and posture during the pilot, and vehicle-based data. Data is also collected from supplementary studies, including driving simulator and Wizard-of-Oz studies, and a large-scale international survey. Each of these methods of data collection is introduced and discussed in the sections below.

The range of methods we have employed in L3Pilot will naturally produce a great deal of data. For the most part, data contributes directly to answering the core L3Pilot research questions. However, the experiments may produce additional data and results that are not linked to the core L3Pilot questions. The additional research questions are too numerous and detailed to capture within Table 2.2. Therefore, they are listed and discussed in Chapter 4.3.3, which includes an overview of the scope of the three evaluation deliverables within SP7 as guidance for the correct evaluation result report.

4.3.2 Data sources and methods

Pilot site questionnaires

One of the primary sources of data for the user & acceptance evaluation within L3Pilot is a pilot site questionnaire, which gathers subjective data from participants at the 14 different pilot sites (for the full questionnaire, see L3Pilot Deliverable D3.3, Metz et al., 2019). This is a unique contribution of the L3Pilot project, as participants will have had real-world experience with these ADFs. In contrast, previously, subjective data was collected from participants either with experience only in simulated environments (cf. Madigan, Louw, & Merat, 2018) or with no hands-on experience at all (cf. Kyriakidis, Happee, & de Winter, 2015).

As mentioned in Chapter 4.1.1, there are four different types of ADFs within L3Pilot, operating in three driving environments. Therefore, three different pilot site questionnaires were designed, one for each environment, with function-specific questions for ADFs operating in each environment. This method allows us to collect responses that are context and ADF specific.

The questionnaire is split into two parts. The first is administered before the pilot drives commence and includes questions related to socio-demographic factors (age, gender, country of residence, education level, employment status, income, and family size), vehicle use and purchasing decisions, driving history, in-vehicle system usage, activities while driving, trip choices, and mobility patterns. The data collected in the first part will be used to create different user groups for the evaluation, and to understand the impact of various socio-demographic factors on participants' acceptance and perception of the ADFs.

The second part of the questionnaire is administered immediately after the pilot drive concludes, or after the final pilot drive if a participant participates in more than one drive. It examines test participants' reactions regarding their experience while using the particular ADF, including acceptance, safety and comfort, among others. To examine whether participants felt that they would change any of their behaviours should they have access to that particular ADF in their daily lives, they were re-asked questions about vehicle use and purchasing decisions, driving history, in-vehicle system usage, engagement with activities while driving, trip choices, and mobility patterns. The questions in this section are phrased to address the specific ADF under investigation, the only exception being motorway and traffic jam ADFs, which utilise the same questions because they have similar ODDs.

As an optional additional section, where feasible, users' controllability and performance during and after a take-over is evaluated mid-drive, following any need to resume manual control from the ADF. For this analysis, drivers are asked immediately after a take-over scenario to rate the criticality of the preceding situation as a whole, using a ten-point scale to judge the criticality of the situation, ranging from harmless (1) to uncontrollable (10). The scale is based on Neukum, Lübbeke, Krüger, Mayser, & Steinle (2008), and allows a direct comparison of drivers' own evaluation of the take-over and the post-drive evaluation by expert raters. This data will only be collected for ordinary drivers and at pilot sites where the safety protocol permits mid-drive evaluations.

All pilot sites and supplementary studies were sent guidance on the administration and implementation of the pilot site questionnaire, to ensure consistency across sites (for more information, see Annex 1). In addition, all sites were sent guidance on what information to give participants about the experiment, ADF operation, safety instructions and permitted actions, ethics, and confidentiality (for more information, see Annex 2).

Video-based analysis

Video-based data has been used extensively in naturalistic and on-road driving studies to collect information about drivers' behaviour (van Nes, Bärgman, Christoph, & van Schagen, 2019). For example, Carsten, Hibberd, Bärgman, Kovaceva, Pereira Cocron, Dotzauer, et al.'s 2017 analysis of UDRIVE video data showed that car drivers were involved in distracting activities for approximately 10% of their driving time. Fridman, Langhans, Lee, and Reimer (2016) used video data to estimate driver gaze location from videos of drivers' faces, and then applied those deep-learning-based computer-vision approaches to data from 129 participants in a real-world AD study (Fridman et al., 2019).

Similar techniques will be used within L3Pilot to assess drivers' operation and use of the ADF. Video recordings can be used in conjunction with deep-learning-based computer-vision approaches to accelerate analysis in this context, providing information such as the type and frequency of users' engagement with non-driving related tasks, as well as providing knowledge about driver state (e.g. fatigue) and body pose. Finally, it is possible to use video-based data to verify the frequency of ADF activation and deactivation, and to identify the situations in which drivers prefer to drive manually, and why. However, the suitability of video analysis is determined primarily by the driver type, and the freedom given to drivers regarding whether they are permitted to activate and deactivate the ADF. For example, professional (test) drivers are not permitted to perform a non-driving task, so this particular analysis would only be relevant for ordinary drivers, in situations where this activity is permitted.

To evaluate the controllability and safety of take-over situations within L3Pilot, a video-based procedure is used, with the same scale described in the previous section. The take-over controllability rating (TOC-rating, Naujoks, Widemann, Schömig, Jarosch and Gold, 2018, www.toc-rating.de/en) was developed as part of the German research initiative KO-HAF (www.ko-haf.de). It provides a uniform, and easy to understand, approach to evaluating take-over situations. TOC-rating was developed to provide a more holistic assessment of take-over situations that goes beyond vehicle parameters, such as deviation of speed or lateral control, but also considers traffic violations (such as missing safety-related glances or absent indicator use) as well as the observed emotions of the driver. One advantage of TOC-rating that makes it especially suitable for the needs of L3Pilot is that a common and standardised rating can be compared across situations, drivers and pilot sites.

In L3Pilot, vehicle-based data is mainly used to answer research questions in the area of technical & traffic evaluation. However, it can also be used to address some user-related topics, since there are scenarios where the driver interacts with the system and controls the vehicle – for example, examining how often, and under what circumstances, drivers choose to activate and deactivate the

ADF. Vehicle-based data can also be used to analyse take-over situations, thus supplementing the TOC-rating with more specific indicators, such as take-over times, or more specific vehicle controllability measures, including the magnitude of lateral or longitudinal accelerations (cf. Louw, Markkula, Boer, Madigan, Carsten & Merat, 2017).

International survey

The pilot site questionnaires provide a unique opportunity to understand how drivers who have used the ADFs perceive and accept them. However, the pilot sites have unavoidable differences in terms of the participant population and experimental design. This means that the sample size available for analysing some questions or topics may not have the statistical power for drawing robust conclusions about the interactions of various socio-demographic factors, user groups and their influence on acceptance of conditionally automated driving (Price, Daek, Murnan, Dimmig, & Akpanudo, 2005). For example, since some questions related to acceptance are tailored to the specific ADF being tested, responses across all ADFs cannot be combined in the evaluation. Therefore, the sample size for some questions may be less than 100. Consequently, a much larger sample is needed to allow analyses of socio-demographic factors. To address this shortcoming, L3Pilot is conducting a two-phase international questionnaire study on ~24,000 car drivers, to investigate the factors that influence acceptance of SAE Level 3 AD (i.e. conditionally automated vehicles) in this group of ordinary drivers.

The first phase of this survey was administered between April 2019 and April 2020 and included 78 items. To develop a global perspective, the survey included respondents from eight European countries (UK, Sweden, France, Germany, Italy, Hungary, Finland, and Spain) and eight non-European countries (China, USA, Brazil, India, South Africa, Turkey, Japan, and Indonesia), with each country represented by 1,000 respondents. These countries were chosen because they represent different geographical regions in Europe and on the other continents and have large car markets. The survey addresses factors such as user acceptance, mobility, privacy, trust, perceived safety, willingness to pay and use, technology readiness, experience with road vehicle automation, and knowledge of ADFs. The survey also includes an adapted version of the Unified Theory of Acceptance and Use of Technology (UTAUT) questionnaire, initially developed by Venkatesh, Morris, Davis and Davis (2003), which provides a comprehensive synthesis of research to model technology acceptance. In L3Pilot, this survey will be used to understand the factors that influence respondents' acceptance of Level 3 AD and to build a more comprehensive model that incorporates aspects such as technology readiness and socio-demographic factors, including age, gender, driving experience, country of residence, education, household, employment status, and income.

The second phase of this exercise will include an updated survey, modified based on experience and the results of phase one, and will be administered in the last quarter of 2020 to respondents from a subset of the 16 countries used in the first phase. Where feasible, the survey uses items from the pilot site questionnaire so that we can examine the extent to which real-world exposure to the ADFs affects respondents' perceptions by comparing the results from the survey and from the data collected in the pilot tests.

Driving simulator and Wizard-of-Oz studies

Driving simulators offer a safe and controllable setting to conduct studies that are either logistically more difficult, or more dangerous, to address in the real-world pilots. Two separate studies are being conducted in high-fidelity driving simulators, and there are also two studies employing the Wizard-of-Oz methodology (Mok, Sirkin, Sibi, Miller, & Ju, 2015) to collect supplementary data on user acceptance and behaviour. Since it takes time and experience for users to build trust in automation (Muir, 1994), it can be argued that drivers' behaviour and interaction with the ADF after the first usage might be different compared to multiple exposures over a prolonged period. With these more controllable approaches, it is also possible to investigate changes in drivers' behaviour in take-over situations, or during interactions with different system functionalities, whereas this would be too dangerous during the pilots. These supplementary studies also continue the emerging research on driver state during AD (Beggiato, Hartwich, & Krems, 2019) and non-driving task engagement (Gold, Berisha, & Bengler, 2015). This is possible because some participants in these studies are free to use the ADF as they like, unlike in the pilots where the freedom to perform non-driving tasks during AD is limited to only some pilot sites for safety or legal reasons.

More specifically, simulator study 1 investigates in an urban road scenario how the time headway adopted by an automated vehicle during car-following situations influences driver behaviour and driver state in subsequent manual driving, including car-following behaviour. Simulator study 2 is a longitudinal study with six driving sessions per driver in a high-fidelity driving simulator. The study investigates behavioural changes when using an ADF repeatedly, take-over performance, and the impact of ADS on driver state with two different system implementations. (For more information, see L3Pilot Deliverable D7.2 (Metz et al., in preparation)).

Wizard-of-Oz study 1 includes three sub-studies. In the first one, the test participant switches between automated driving and manual driving on a public road. The second one is performed on a test track to study the take-over performance and driver's response during a conflict. The third one is similar to the second but includes the influence of intoxication. In Wizard-of-Oz study 2, ordinary drivers used an alleged ADF on public motorways three times. There was a break of one week between the drives. Video records of the subjects' behaviour during the drives are supplemented with questionnaires on trust and acceptance before and after each drive. (For more information, see L3Pilot Deliverable D7.2 (Metz et al., in preparation)).

4.3.3 Evaluation plan

First, Table 4.14 provides an overview of the performance indicators that will be used to address each detailed research question (Level 3), as well as the specific research questions that each data collection methodology will be contributing to. As mentioned above, this table shows that the questionnaire-based data forms a large part of the user & acceptance evaluation, with additional quantitative data provided by simulator and Wizard-of-Oz studies, as well as an international user survey. For references to items in the pilot site questionnaire or video coding, please refer to the full questionnaire and video codebook in L3Pilot Deliverable D3.3 (Metz et al., 2019). For references to items in the international user survey, please refer to the full survey in L3Pilot Deliverable D7.1 (Nordhoff et al., in preparation). For references to specific performance indicators related to simulator and Wizard-of-Oz studies, please refer to their detailed methodologies in L3Pilot Deliverable D7.2 (Metz et al., in preparation). Second, Table 4.15 provides an overview of adapted hypotheses for behavioural adaptation set for each research question and how they are addressed in the simulator and Wizard-of-Oz studies.

Table 4.14: Overview of the plan for user & acceptance evaluation, including performance indicators for each research question, and for each method.

RQ	RQ Level 2	All pilot sites, pilot site questionnaire (D7.3)	International Survey (D7.1)	Simulator 1 (D7.2)	Wizard-of-Oz 1 (D7.2)	Simulator 2 (D7.2)	Wizard-of-Oz 2 (D7.2)
		Reporting after single or final exposure to ADF				**Reporting changes after multiple exposures to ADF**	
RQ-U1	Are drivers willing to use an ADF?	TJM.32.a, U.32.a, P.27.a	f17a-f17d, f16_10, f16_12, f16_14, f16_16, f16_23,	Bespoke Questionnaire		Questionnaire: TJM.32.a Recorded data of actual usage of ADF (vehicle data)	TJM.32a, vehicle data: recordings of 'driver in control'
RQ-U2	How much are drivers willing to pay for the ADF?	TJM.32.d, TJM.32.e,TJM.32.f, U.32.d, U.32.e, U.32.f, P.27.e, P27.f, P.27g, TJM.37, U.37, P30	f18a-f18d			Questionnaire: TJM.32.d, TJM.32.e,TJM.32.f,	
RQ-U3a	What is the perceived safety of the ADF?	TJM.28, TJM.32.c, TJM.32.z, U.32.c, P.25, P.27d	f19b_2, f19b_3, f19b_2, f19b_6, f19b_8	Bespoke Questionnaire		Questionnaire: TJM.28, TJM.32.c, TJM.32.z	Questionnaire: TJM.28/32c
RQ-U3b	What is the perceived comfort of the ADF?	TJM.31, TJM.32q, U.31, U.32.q, P.25		Bespoke Questionnaire		Questionnaire: TJM.31, TJM.32q,	Questionnaire: TJM.31
RQ-U3c	What is the perceived usefulness of the ADF?	TJM.30, TJM.32.g, TJM.32.p, U.32.g, U.32.p, P.26, P.27.n, P.27.q	f16_1, f16_2, f16_5, f16_22	Bespoke Questionnaire		Questionnaire: TJM.30, TJM.32.g, TJM.32.p	Questionnaire: TJM.32g
RQ-U3d	What is the perceived trust of the ADF?	TJM.32.o, TJM.32.l, U.32.o, U.32.l, P,27.o, P.27.m	f19a_4 - f19a_12, f19c_4 - f19c_12, f19d_4 - f19d_12	Bespoke Questionnaire	Open questions Pilot site questionnaire Extent of secondary task engagement	Questionnaire: TJM.32.o, TJM.32.l	Questionnaire Secondary task engagement Eye-tracking: monitoring gazes

RQ	RQ Level 2	All pilot sites, pilot site questionnaire (D7.3)	International Survey (D7.1)	Simulator 1 (D7.2)	Wizard-of-Oz 1 (D7.2)	Simulator 2 (D7.2)	Wizard-of-Oz 2 (D7.2)
RQ-U3e	How does user acceptance differ between ADF types (urban, motorway, traffic jam, parking)?	TJM.32.b, TJM.32.k, TJM.32.m, U.32.b, U.32.k, U.32.m, P.27.c	Yes – all items from RQ3 can be compared				
RQ-U4	What is drivers' overall impression of the system?	TJM.32.p, TJM.32.u, TJM.v, TJM.w, U.32.p, U.32u, U.32,v, P.27.q		Questionnaire		Questionnaire: TJM.32.p, TJM.32.u, TJM.v, TJM.w,	
RQ-U5a	What is the effect of ADF use on drivers' level of stress?	TJM.32.j, U.32.j, P.27.k		Heart Rate Variability Electrodermal activity		Questionnaire: TJM.32.j,	
RQ-U5b	What is drivers' level of fatigue while using the ADF?	TJM.32.t, U.32.t			Eye-tracking: **PERCLOS *KSS	Questionnaire: TJM.32.t, *KSS, Eye-tracking: **PERCLOS	
RQ-U5c	What is drivers' workload while using the ADF?	TJM.32.h, TJM.32.i, U.32.h, U.32.i, P.27.i, P.27.h		Heart Rate Variability Electrodermal activity		Questionnaire: TJM.32.h, TJM.32.i	TJM32.h
RQ-U6a	What is the effect of ADF use on driver attention to the road/other road users?	TJM.32.r, U.32.r, P.27.r		Eye gaze distribution	Video coding of secondary task engagement Eye gaze distribution	Questionnaire: TJM.32.r, Eyes-on-road time, Frequency and duration of visual checks	
RQ-U6b	What is drivers' risk perception while using the ADF?	TJM.32.s, U.32.s, P.27.s		Eye gaze distribution	Video coding of secondary task engagement Eye gaze distribution	Questionnaire: TJM.32.s, eyes-on-road-time	

RQ	RQ Level 2	All pilot sites, pilot site questionnaire (D7.3)	International Survey (D7.1)	Simulator 1 (D7.2)	Wizard-of-Oz 1 (D7.2)	Simulator 2 (D7.2)	Wizard-of-Oz 2 (D7.2)
RQ-U7	What is the impact of ADF use on motion sickness?	TJM.29.a, U.29.a					
RQ-U8	What is the impact of motion sickness on ADF use?	TJM.29.b, U.29.b					
RQ-U9a	What secondary tasks do or would drivers engage in during ADF use?	TJM.32.n, (TJM.33& TJM.21), U.32.n, (U.33&U.21)	f14, f14b_1-f14b_11		Video coding of secondary tasks engagement (See codebook in D3.3)	Questionnaire: TJM.32.n, (TJM.33& TJM.21)	TJM.33 Video coding of secondary tasks engagement (See codebook in D3.3)
RQ-U9b	What is the frequency and duration of drivers' secondary task engagement during ADF use?	TJM.33, U.33			Video coding of secondary tasks engagement (See codebook in D3.3)	Questionnaire: TJM.33, U.33 Video coding of secondary tasks engagement (See codebook in D3.3)	TJM.33 Video coding of secondary tasks engagement (See codebook in D3.3)
RQ-U10a	How do drivers respond when they are required to retake control in planned take-overs?	Questionnaire: TJM.32.x, TJM.32.α, TJM.32.β, U.32.w, U.32.z.U.32.α, TJM.38, TJM.39a-39.f, U.38, U.39a-39.f TOC-Rating		Time to resume control Time headway Heart Rate Variability Electrodermal activity	Video coding of hands on wheel TOC-Rating based on video Steering input Brake input	Questionnaire: TJM.32.x, TJM.32.α, TJM.32.β, TJM.38, TJM.39a-39.f, TOC-Rating based on video, reaction times after TOR (e.g. eyes on road time, hands on wheel time)	Video coding of secondary tasks engagement (See codebook in D3.3)
RQ-U10b	How do drivers respond when they are required to retake control in unplanned take-overs?	Questionnaire: TOC-Rating			Video coding of surprise reaction and hands on wheel Steering input Brake input		Video coding of secondary tasks engagement (See codebook in D3.3)

RQ	RQ Level 2	All pilot sites, pilot site questionnaire (D7.3)	International Survey (D7.1)	Simulator 1 (D7.2)	Wizard-of-Oz 1 (D7.2)	Simulator 2 (D7.2)	Wizard-of-Oz 2 (D7.2)
RQ-U11a	How often and under which circumstances do drivers choose to activate the ADF?					Time with ADF activation in different conditions (e.g. traffic jam, low traffic density, speed limit)	Video coding: classification of different conditions, Automation activation profile
RQ-U11b	How often and under which circumstances do drivers choose to deactivate the ADF?						Video coding: classification of different conditions, Automation activation profile

*KSS = Karolinska Sleepiness Scale

**PERCLOS = Percentage of eye closure

Table 4.15: Overview of adapted hypotheses for behavioural adaptation addressed in simulator and Wizard-of-Oz studies. Simulator study Sim1 and Wizard-of-Oz study WoZ1 expose each user once to ADF while studies Sim2 and WoZ2 expose them multiple times.

RQ-ID	Hypotheses for behavioural adaptation	Sim1	WoZ1	Sim2	WoZ2
RQ-U1	Willingness to use increases with increasing experience with ADF.				
RQ-U3	Perceived safety increases with increasing experience with ADF.	x		x	x
	Perceived comfort increases with increasing experience with ADF.	x		x	
	Perceived reliability increases with increasing experience with ADF.	x		x	
	Perceived usefulness increases with increasing experience with ADF.	x		x	
	Trust increases with increasing experience with ADF.	x		x	(x)
	Critical take-over scenarios will have a decreasing effect on trust in the ADF.	x		x	x
RQ-U5	Over AD usage time, drivers experience less stress.			x	
	After a familiarisation period, drivers will become drowsy more rapidly.			x	(x)
	Over AD usage time, drivers experience less workload.	x		x	
RQ-U6	With increasing experience, attention to other road users will become more dependent on the driving scenario.	x		x	
RQ-U4	With increasing experience, understanding of the system increases.	x		x	x
RQ-U9	Secondary tasks engaged in will vary with increasing experience with ADF.	x	x	x	(x)
	Secondary task interaction increases with increasing experience with ADF.			x	x
RQ-U10	Take-over performance increases with increasing experience with ADF.		x	x	x
RQ-U11	Pattern of system activation will become more dependent on driving scenario with increasing experience with ADF.		x	x	
RQ-UE1	After ADF use, manual driving behaviour changes.			x	x

Since there is no single deliverable dedicated to reporting on all the results of the user & acceptance research questions, Table 4.16 outlines in detail how these will be addressed across the three respective deliverables as part of the L3Pilot evaluation.

Table 4.16: Reporting of user & acceptance evaluation results in L3Pilot Deliverables

Deliverable	Scope & Evaluation plan
D7.1 Annual survey about user acceptance towards ADAS and vehicle automation	L3Pilot Deliverable D7.1 reports the results from the international survey. *Scope*: The safety, usability and acceptance of conditionally automated vehicles depend on managing a comfortable and efficient take-over situation (Nordhoff et al., submitted). Monitoring the behaviour of the human driver is one possible way to making sure that the take-over situation is safe and efficient. This raises the question whether human drivers permit the collection of sensitive behavioural data by automated vehicles. Data privacy concerns were rated as one of the main concerns people have with automated cars (Regan, Cunningham, Dixit, Horberry, Bender, Weeratunga, et al., 2017, Schoettle & Sivak, 2014). They originate from transmitting data to the government, vehicle developers, and insurance companies without notice, using data against people, or hacking data (Zhang, Tao, Qu, Zhang, Lin & Zhang, 2019). We think that the effective functioning of automated vehicles depends on driver monitoring, and hence, on the consent of human drivers for automated vehicles to collect and use this highly sensitive behavioural data. For this reason, D7.1 examines the willingness of human drivers to share data with an automated vehicle, as well as how the construct willingness to share data interacts with other key technology constructs (e.g. UTAUT constructs). The literature on automated vehicle acceptance has also shown that a person's technology readiness is a key determinant of acceptance. Therefore, the technology readiness as well as its interaction with key technology constructs will be investigated. *Evaluation*: Data from the international survey will be analysed using a range of basic and advanced statistical techniques. The techniques used will depend on the specific research question and data being analysed. For example, as with the pilot site questionnaires, some items within the international survey are often constructed in such a way as to include their own baseline. Participants may be asked to estimate the potential change that using a particular ADF may have based on a current behaviour. In these cases, descriptive analysis (sample demographics) and Chi-squared analysis (comparison different ADFs and demographic groups) will be used. There are also examples where more complex analysis will be needed. For example, structural equation modelling will be used to analyse the UTAUT constructs, Latent class analysis based on Bayesian inference criteria will be used to cluster and analyse personal mobility patterns, while ordinal regression analysis will be used to analyse which factors predict intention to use ADFs.
D7.2 L3/L4 long-term study about user experiences	L3Pilot Deliverable D7.2 reports the results of simulator and Wizard-of-Oz studies. *Scope*: With increasing vehicle automation, the role of the driver changes fundamentally. On a technological level, vehicle automation is progressing rapidly. Still, there are challenges regarding the interaction between human drivers and automated systems. Another important aspect that has to be considered is that drivers may change their behaviour due to the automation. This phenomenon is referred to as *behavioural adaptation*. Part of the research on user-related questions in L3Pilot deals explicitly with the effects of behavioural adaptation. For instance, the impact of increasing experience with an ADF on acceptance, trust and usage is studied. The questions addressed are mostly linked to the common L3Pilot research questions but look at the change over time. However, there are also specific research questions on behavioural adaptation. *Evaluation*: Since the simulator and Wizard-of-Oz studies are more controlled and planned than the ADF pilots, the evaluation of data from these will typically include more traditional parametric statistical techniques. For example, using mixed-ANOVAs to compare responses between different experimental groups, at different time points, and after exposure to different experimental conditions. The supplementary studies also use the pilot site questionnaire, so the analysis of that data will follow the details described below.

D7.3 Pilot Evaluation results	L3Pilot Deliverable D7.3 reports on all evaluations conducted on data collected at the pilot sites, including on-road and test track studies for the four tested ADFs. *Scope*: As part of that, D7.3 includes the evaluation of the pilot site questionnaires, which will be conducted and reported per research question (Level 3). D7.3 will also consider differences between ADFs, which will primarily include the comparison of responses to the same questions in the pilot site questionnaires made for different ADFs. Where feasible and relevant, D7.3 will also include comparisons of responses between different testing types (i.e. open-road pilot test, test track, driving simulator, Wizard-of-Oz) and user types (i.e., professional (test) driver, non-professional driver, passenger). Since there is no single output for all user-based evaluations within L3Pilot, D7.3 will contain a summary of results for each user-related research question, across all methods described above. *Evaluation*: Items within the pilot site questionnaire are typically constructed in such a way as to include their own baseline. Participants may be asked to estimate the potential change that using a particular ADF may have based on a current behaviour. In these cases, descriptive analysis (sample demographics) and Chi-squared analysis (comparison of different ADFs and demographic groups) will be used.

4.4 Mobility impact assessment

4.4.1 Overall approach

L3Pilot uses a multidisciplinary mobility assessment approach and framework to define the potential mobility impacts of ADF, and further to answer questions about potential impacts on actual travel exposure. In the mobility impact assessment, we focus on the three points of view described in the FESTA Handbook (FOT-Net & CARTRE, 2018): amount of travel, travel patterns, and journey quality (Figure 4.1), which were translated into three research questions (Table 4.17).

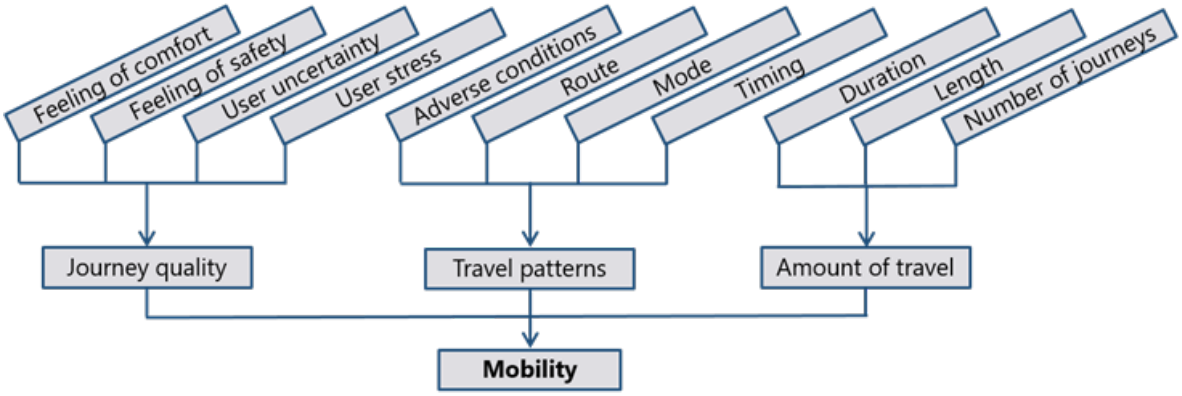


Figure 4.1: Mobility model for impact assessment (Innamaa, Axelson-Fisk, Borgarello, Brignolo, Guidotti, Martin Perez, et al., 2013). Edited.

The overall approach to mobility impact assessment within L3Pilot has four major steps, which are presented in Figure 4.2. First, the current realised mobility behaviour is described. Second, the scope of the impacts is identified by analysing on which proportion of current trips ADFs would be available, and who would be the first to start using them. Third, in the mobility impact assessment, the potential impacts of ADFs will be assessed for all the research questions. Fourth, in the scaling-up, some of the potential impacts will be translated into quantitative estimates relative to the current travel behaviour (actual trips made), which we call realised mobility. Scaling-up is an important step especially considering the implication of the mobility impacts on efficiency, environment and safety, as changes in mobility change travel exposure. For example, if the number of kilometres travelled by car increases, so does exposure, influencing the net safety impact of ADFs.

The mobility impact assessment will be performed primarily assuming that all the tested ADFs will be in use. Where applicable, the impacts of specific ADFs will be estimated. When determining the scope for mobility impact, we will estimate the proportion of roads where specific ADFs could be available. This will help us understand the impact potential of each of the ADFs.

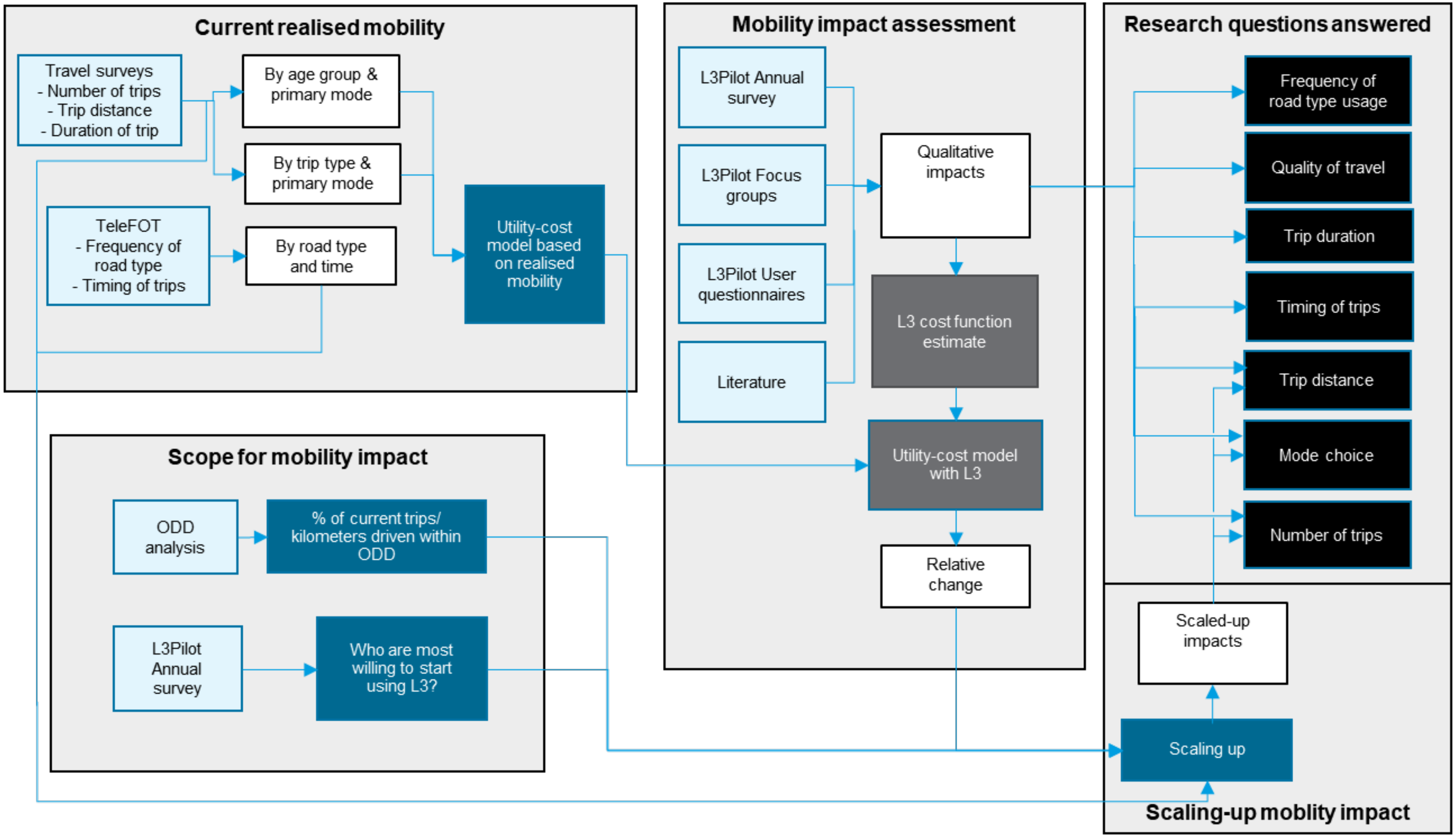


Figure 4.2: Overview of the methodology for mobility impact assessment in L3Pilot.

4.4.2 Data sources and methods

Definition of current realised mobility

The overview of the current realised mobility will be based on the latest available travel exposure datasets in Europe. The main source for the amount of travel and mode will be the outcomes of the SHANTI project (COST Action TU0804), which include travel survey data from eleven European countries (Belgium, Switzerland, Germany, Great Britain, Denmark, Spain, Finland, France, the Netherlands, Norway, and Sweden). The data includes the daily number of kilometres, trips and time use per traveller and main travel mode, and can be aggregated by, for example, age group, type of trip and type of day. Where feasible, the SHANTI data will be updated with the latest statistics and complemented with data from other countries. The current situation for the timing of trips and route choice will be based on actual travel data from previous projects such as TeleFOT (Innamaa et al., 2013; Schulze, Mäkinen, Kessel, Metzner, Stoyanov, 2014). Start and end points for a sample of trips will be extracted from the TeleFOT data. Then, the fastest current routes between the start and end points will be determined using, for example, Google Maps API. The corresponding road infrastructure will be extracted from OpenStreetMap. Based on this analysis, we can determine which proportion of the shortest routes utilise the road network within the ODD.

Scope for mobility impact

The analysis of the scope for mobility impact addresses the potential users' current trips and travel options that could be affected by using ADFs. The analysis is based on two main sources of information. The first is the ODD defined for each mature ADF (i.e. a specification of the conditions under which the ADF is assumed to work (see mature ADF definitions, Chapter 4.1.1)). These conditions include variables related to, for instance, infrastructure, weather conditions and speed. In other words, it is to be defined for what or which trips automation would be available. The mobility impact assessment will use the data developed in the efficiency scaling-up to estimate the kilometres travelled within ODD.

The second source is the international survey carried out in the project that will provide data on the general acceptance, willingness to use, and perceptions of ADFs' usefulness among European car drivers, and how these depend on user characteristics such as age, gender, and current mobility patterns. These results, based on data collected among persons who have not yet experienced ADFs, need to be carefully interpreted relative to the results gained from the pilot tests. Penetration rates for ADF in use are those determined in Chapter 4.1.3.

Qualitative mobility impact analysis

As in L3Pilot, ADFs are prototypes tested in controlled experiments and not yet on the market. The potential mobility impacts cannot be directly measured and need to be estimated. Based on the conceptual mobility framework (Kuisma, Louw & Torrao, 2019), we identified the main factors influencing mobility behaviour: a) land use, infrastructure and transport planning, b) social, cultural, and psychological factors, c) situation-specific factors such as need to transport goods, and d) habits. During the mobility impact assessment, we will assess how these factors could be affected by ADFs and how large impacts these could generate.

Qualitative analysis of the potential mobility impacts will determine the directions, influencing factors and, if possible, relative magnitudes of the impacts for each Level-3 research question. A synthesis of how ADF could change the mobility considering the current travel behaviour and users' willingness to use ADF will be also presented.

Qualitative analysis will use primarily data from L3Pilot's pilot site questionnaires, focus groups, and international survey. Pilot site questionnaires to the test participants will provide views from persons who have actually experienced an ADF and also how their views changed after the experience. Thus, the mobility impact assessment is linked with user evaluation and utilises input from it.

The user experiences are relevant for mobility impact assessment in two ways: First, the experienced usefulness and comfort of ADF presumably affect people's travel behaviour, resulting in changes in mode choice and the decision on which trips are made or not made. As the participants in L3Pilot's user tests are not necessarily representative in terms of mobility behaviour, we can use *pilot site questionnaire* items describing the demographics and travel patterns to map the participants to current realised mobility in Europe. The data which can be used in mapping includes age, typically used travel mode(s), and how often participants drive on motorways overall and on congested motorways, rural roads, or urban streets (TJM.24, see pilot site questionnaire items in L3Pilot Deliverable D3.3 by Metz et al., 2019).

The second way to utilise user experiences is to organise *focus groups* among the test participants, which will give further insights into the reasoning processes and valuations underlying travel choice. Specifically, we will investigate how mode choices are justified and categorise arguments. One way to categorise arguments is to group them into instrumental or affective (Anable & Gatersleben, 2005). Instrumental factors concern practical aspects of travelling, such as time used, monetary costs, predictability, or impact on health. Affective factors concern the feelings evoked by travelling, including stress and boredom as well as excitement and relaxation. Instrumental reasons to select a mode can be further divided, for example, into predictability, cost, convenience, and flexibility. Predictability relates to knowing how long a trip will take and is strongly influenced by familiarity with the transport mode. Cost can mean the monetary costs of using a certain transport mode, but also the value of time when travelling with different modes. Convenience refers to the personal effort involved in using the transport mode. Flexibility means how dependent the transport mode is on timetables and how well it supports changes in the travel plan. Affective reasons include freedom, feeling of control, and lack of stress.

The *international survey* data will reveal how the general public views the expected changes in mobility behaviour once ADFs become available.

The qualitative analysis will be translated into quantitative estimates of the relative changes in the mode choice and number of trips performed. This will be done using a *cost-utility model* developed during the project. The starting assumption for the model is that all the possible trips have some utility and the utility values have some distribution. On the other hand, each transport mode can be understood to incur some 'cost' if it is used on a certain trip. The cost can be monetary, but also time or something more abstract like comfort. Here, cost comprises all aspects of the mode choice: predictability, cost, convenience, and flexibility. The cost values will also be understood as a distribution. We will assume that an individual will choose the transport mode with the lowest cost available for the trip and will undertake the trip only if its utility exceeds its costs. With these simple assumptions we can understand that the current modal split reflects the relative costs of the available transport modes among the trips that were performed.

Using the data for the current modal split, we will estimate the relative cost distributions of transport modes currently in use. Then, using the scope analysis, we will determine the share of the trips where ADF would be available. Finally, based on the qualitative assessment, we will form a cost distribution for ADF.

The cost function will be selected primarily based on L3Pilot's:

- Pilot site questionnaires items regarding the change in the number of trips made (TJM.33.v, U.33.u), mode choice (U.36), intention to use the system on everyday trips (TJM.33.p, U.33.p), and willingness to use a longer trip with ADF (TJM.35).
- International survey results based on the expected mobility impacts (e.g. the expected impact on the use of public transport and active travel modes).
- Focus groups, where the aforementioned impacts can be investigated further.

After performing these steps, we can simulate the model and generate the estimated modal share for ADF and other transport modes, as well as the number of trips performed for each mode. Due to the nature of the model, it is also possible to estimate how many new trips could be travelled if the costs of using ADF were lower than with current passenger cars.

Scaling-up relative to the current realised mobility

Lastly, the quantitative estimates of the relative changes will be applied to an EU context to illustrate the overall potential mobility impacts of ADFs. For example, what it would mean if there were a change in the number of trips or in the kilometres travelled or in the mode choice. This will be done with the quantitative relative estimates and the current realised mobility data. The outcome serves as input for the evaluation areas: safety and efficiency and environment.

4.4.3 Evaluation plan

RQ-I5: What is the impact of ADF on amount of travel?

- What is the impact of ADF on number of trips made?
- What is the impact of ADF on trip distance?
- What is the impact of ADF on trip duration?

The current realised mobility will be based on available travel exposure datasets (e.g. SHANTI data) in Europe. The data includes the daily number of kilometres, trips and time use per traveller, and main mode, which are aggregated by age group, type of trip and type of day.

The scaling-up of safety and efficiency impacts will provide the share of kilometres driven within the ODD (scope of impact). These can be further translated into a percentage of trips driven (partly) within the ODD of the current realised mobility, assuming the penetration rates defined in Chapter 5.1.3. The results of the L3Pilot international survey will be used to define who will be the most potential users per ADF, age group, and traveller group (e.g. all-purpose car drivers or active mode users).

The impact on the number of trips will be based on L3Pilot's pilot site questionnaires (e.g. intention to make more trips TJM.33.v, U.33.u), international survey (expected impact on the use of travel modes), focus groups, and literature. The impact on the trip distance can be assessed with L3Pilot's pilot site questionnaires (e.g. TJM.33.w, U.33.b) and focus groups. The impact on trip duration can be assessed based on L3Pilot's pilot site questionnaires (TJM.35, TJM.36, U.35) and focus groups. Additional data from the third wave of L3Pilot's international survey can be included in the assessment if available.

The cost-utility model will be used to estimate a relative change in the number of trips per mode. For that purpose, the cost function for ADF will be determined based on the qualitative mobility impact analysis. Utilising the cost-utility mode, it will be possible to perform a quantitative scaling-up for the number of trips relative to the current realised mobility.

RQ-I6: What is the impact of ADF on travel patterns?

- What is the impact of ADF on mode choice?
- What is the impact of ADF on timing of trips?
- What is the impact of ADF on the frequency of road type usage? (urban, rural, motorway)

Travel survey data will give the current situation for mode choice and timing of trips by trip purpose and age group. The current situation for the timing of trips and route choice will be based on the trip data from the TeleFOT project (Innamaa et al., 2013).

The impact on timing of trips will be assessed with L3Pilot's pilot site questionnaires (TJM.36, U.35) and focus groups. The impact on the frequency of road type usage will be assessed based on L3Pilot's pilot site questionnaires (TJM.35) and focus groups.

The cost-utility model will give a relative estimate of the total number of trips and the modal choice once ADF is available. In scaling-up, this can be used to produce a scaled-up impact also for the mode choice data.

RQ-I7: What is the impact of ADF on quality of travel?

Quality of travel is addressed in L3Pilot from the following aspects: user stress, user uncertainty, feeling of safety, feeling of comfort, and secondary task engagement. Since all of these are related to Level-3 research questions under user & acceptance evaluation (RQ-U3, RQ-U5 & RQ-U9), the analysis of RQ-I7 will be closely linked to the user & acceptance evaluation (Chapter 5.3). As opposed to the four steps used for RQ-I5 and RQ-I6 and for user & acceptance evaluation, the analysis focuses on identifying where the quality of travel could be improved for AD compared to human driving. Thus, the aim is to identify situations and preconditions for where there could be a change in amount of travel and travel patterns.

Table 4.17 provides an overview of the evaluation plan for the mobility impact assessment linking the main data sources to the research questions.

Table 4.17: Overview of evaluation plan for mobility impact evaluation's research questions.

RQ-ID	RQ Level 3	Main data for current realised mobility	Main data for the impact
RQ-I5	What is the impact of ADF on number of trips made?	Travel surveys	Pilot site questionnaire (TJM.33.v) International survey (expected mobility impacts) Focus groups
	What is the impact of ADF on trip distance?	Travel surveys	Pilot site questionnaire (TJM33.w) Focus groups
	What is the impact of ADF on trip duration?	Travel surveys	Pilot site questionnaires (TJM.35/U.35, TJM.36)
RQ-I6	What is the impact of ADF on mode choice?	Travel surveys TJM./U.22 TJM./U.23	Pilot site questionnaire (U.36) International survey (expected mobility impacts) Focus groups
	What is the impact of ADF on timing of trips?	Literature TeleFOT TJM.23	Pilot site questionnaires (TJM.36, U.35) Focus groups
	What is the impact of ADF on the frequency of road type usage? (urban, rural, motorway)	TeleFOT TJM./U.24	Pilot site questionnaires International survey (expected mobility impacts), Focus groups TeleFOT
RQ-I7	What is the impact of ADF on quality of travel?	TJM./U.25	Pilot site questionnaire (TJM./U.32, TJM./U.33, TJM./U.34) International survey (expected mobility impacts) Focus groups

4.5 Safety impact assessment

4.5.1 Overall approach

The main objective of the safety impact assessment is to investigate the effect of ADFs on traffic safety (Level-1 research question, Table 2.3). According to the theoretical background proposed by Nilsson (2004), traffic safety consists of three dimensions: (1) exposure, (2) expected number of accidents per unit of exposure, and (3) consequences (i.e. severity of an accident). The exposure aspect is covered in the mobility impact assessment (Chapter 4.4). In addition, the effect of ADF on the frequency of driving scenarios under consideration of the penetration will be estimated as part of safety simulations. Therefore, the remaining important research questions (Level 2) to be handled during the safety impact assessment concern frequency of accidents and accident severity:

- What is the impact of ADF on the number of accidents in a certain driving scenario / for certain road users?
- What is the impact of ADF on accidents with a certain injuries level / damage in a certain driving scenario?

These research questions are further divided into more specific research questions (Level 3). This chapter provides an evaluation plan for all the research questions of all levels.

In general, the assessment of the above research questions requires the quantification of possible effects of ADFs on traffic safety, as shown in Figure 4.3. In this context, it must be considered that each technology can affect traffic scenarios in three ways. First, there are scenarios in which a technology reduces risks. These are typically scenarios, in which something goes wrong today and hence an accident occurs. Second, there are scenarios that are not directly affected by the technology and hence these scenarios do not need to be analysed further. These are for instance scenarios outside the ODD of the mature L3Pilot ADFs. Third, there are scenarios in which potentially new risks occur (e.g. due to a minimum risk manoeuvre). Typically, the occurrence of these types of risks are minimised in the development of ADFs throughout different measures (e.g. functional safety, safety in use). However, for a comprehensive assessment, it is necessary to analyse all potentially affected scenarios – positive as well as negative.

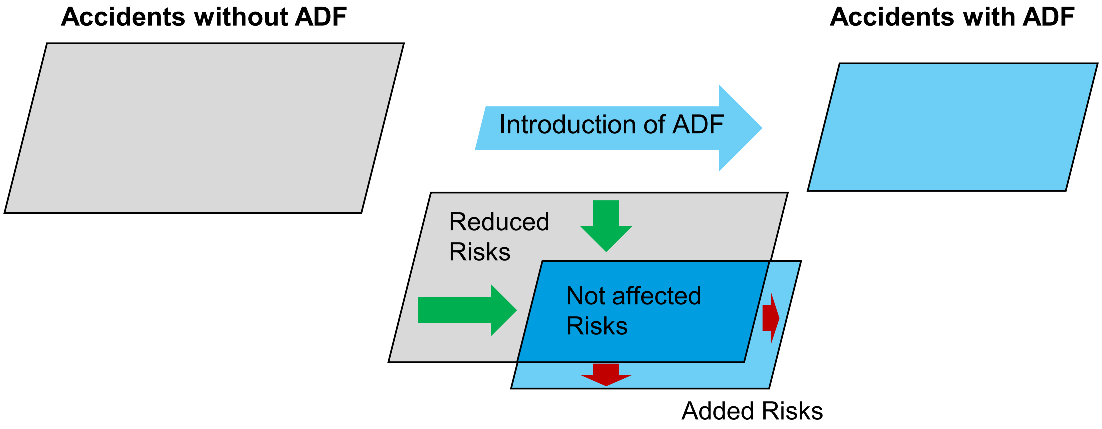


Figure 4.3: Basic principle of an impact of ADFs on traffic safety.

The adequate quantification of the effects of ADFs on traffic safety is a challenging task that requires a comprehensive approach. The approach defined for the L3Pilot safety impact assessment is presented in Figure 4.4 and described in more detail in L3Pilot Deliverable D3.3 (Metz et al., 2019). The basic principle of the approach is to investigate the effects of ADF in dedicated scenarios – which consist of driving scenarios as well as traffic scenarios. The results that are determined in these scenarios are then scaled up at a later stage at EU level.

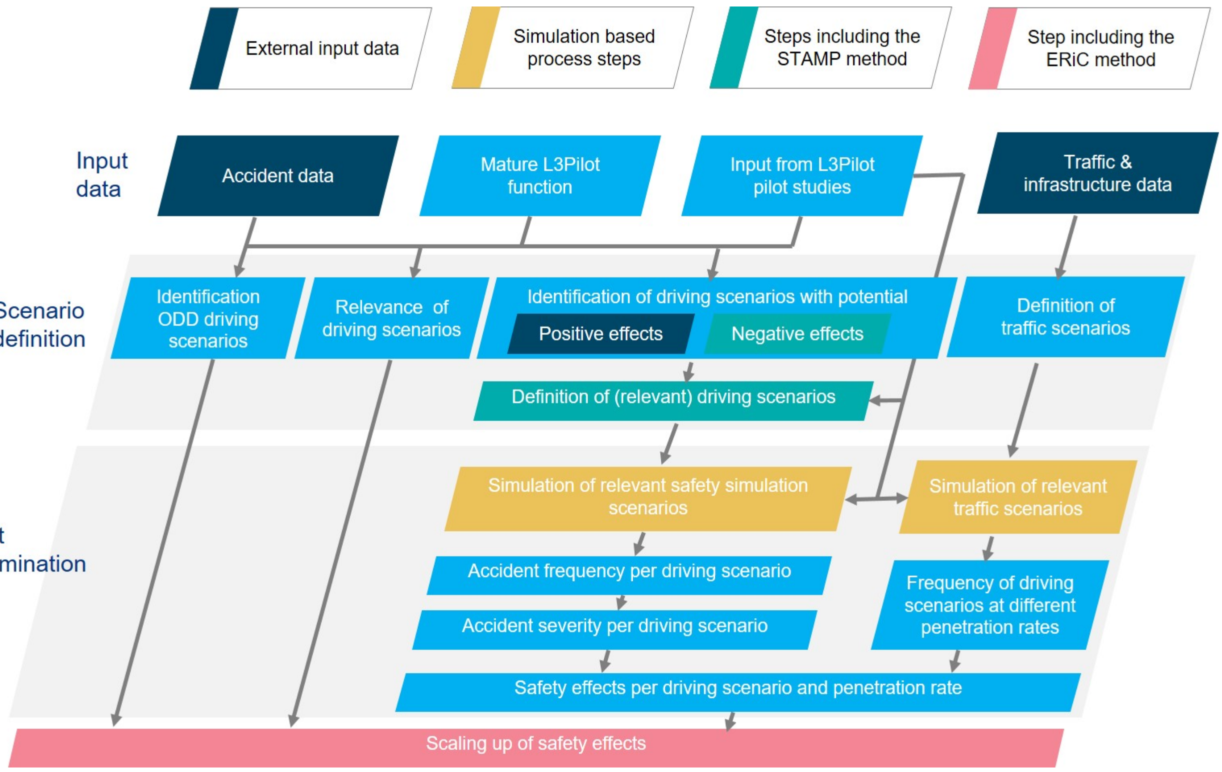


Figure 4.4: L3Pilot approach to safety impact assessment.

The L3Pilot safety impact assessment approach combines different data sources and methods in order to achieve the required comprehensiveness. An overview of these is given in Table 4.18 (for input data) and in Table 4.19 (for methods).

Table 4.18: Description of input data used in L3Pilot safety impact assessment.

Input data	Purpose in L3Pilot safety impact assessment
Accident data (external input data)	The safety impact assessment exploits accident data both from the European-wide CARE database (European Commission, 2019) and from national in-depth (or semi in-depth) road accident databases from selected countries. In addition, data from insurance companies will be used to estimate the potential effect of parking ADF. – CARE database: Definition of the number of accidents per accident type and per traffic scenario within ODD. Scaling up the safety effects from simulated and traffic scenarios to European level. – GIDAS (Germany), The Finnish Crash Data Institute, Police reported accidents in Saxony (Germany), All police reported accidents covering EUSka + additional parameters (Germany), STRADA database (Sweden), VOIESUR database (France): Definition of relevant accident types and driving scenarios (and their distribution) within ODD per scenario (simulated, driving and traffic). AZT insurance accident database (based on claims of Allianz Germany and Allianz Netherlands): Estimation of the potential effects of ADF on parking accidents.
Mature L3Pilot ADF descriptions	The mature L3Pilot ADF descriptions (defined and described in Chapter 4.1.1) and their ODD will be used to define the target accidents (within ODD).
Input from L3Pilot's pilot studies	Results from the technical assessment will be used to determine the distributions for setting up the simulated scenarios and to define parameters for the mature L3Pilot ADFs (longitudinal and lateral control). Results from the user assessment will provide inputs on road user interaction, models for the take-over behaviour (e.g. take-over time) and produce information on usage of ADFs for the scaling-up. Specifically, – Data collected by means of the AIM method will provide insights on the interaction between automated vehicles and VRUs (D7.4, in progress). – Results from the driving simulator studies and Wizard-of-Oz experiments (D7.2, in progress) will provide insights on the direct and indirect effects of ADFs for the ERiC method.
Traffic & infrastructure data (external input data)	Traffic and infrastructure data will be used to identify the representative motorway and urban environments for simulations.

Table 4.19: Description of methods applied in L3Pilot safety impact assessment.

Method	Purpose in the L3Pilot safety impact assessment
Field of application analysis based on accident data	Identify relevant accident types in order to define the scenarios for safety simulations (i.e. scenarios in which a positive effect can be achieved by the ADF).
The STAMP approach and the STPA methodology (to support the identification of driving scenarios with potential negative effects)	The STAMP (System Theoretic Accident Model and Processes) approach and its STPA (Systems Theoretic Process Analysis) derived methodology allow adding a qualitative analysis to the quantitative analysis conducted in L3Pilot. The method enables identification of safety requirements necessary to design technology in a safe manner, at whatever level of granularity, by considering potential problems during the functioning and use of the technology, at all steps of the control structure of the whole designing system (from the top regulating bodies to the single element of a sensor). Hence, it goes beyond the analysis of defects and failures. STPA methodology could specifically explain the quantitative estimates of reduced (or increased) crashes and crash risks and could also identify potential driving situations for which ADFs could have a negative effect on safety.
Stochastic computer simulations of safety simulation scenarios (simulation-based process step)	Determine the change in the accident frequency and accident severity in the simulated scenarios from the baseline scenario to the defined treatment scenarios. Conditions are compared by means of standard statistical approaches (e.g. confidence interval, effect size). Different simulation tools (e.g. openPASS, Virtual Test Drive, MATLAB based tools) will be used depending on the simulated scenario and road type.
Stochastic computer simulations of traffic scenarios (simulation-based process step)	Determine the change in the frequency of driving scenarios per traffic scenarios from the baseline scenario to the defined treatment scenarios. Different simulation tools (e.g. openPASS, VISSIM) will be used depending on the ADF.
Injury risk function	Transfer of the accident situation determined (e.g. speed and velocity) during the safety simulations into injury level of the involved traffic participants by applying the appropriated injury risk functions to the simulation results.
ERiC method	Scaling up of safety effects of ADF from the simulation results to EU level. This process includes several main steps: – Determination of relevant direct and indirect impact mechanisms (Innamaa, Smith, Barnard, Rainville, Rakoff, Horiguchi, et al., 2018). – Identification of expected changes for each relevant mechanism (i.e. changes in vehicle, driver and road user behaviour) based on L3Pilot's pilot study results, safety simulations, literature, statistics (mobility, vehicle fleet etc.), and in-depth road accident statistics. – Mapping of simulated scenarios with relevant accident types based on the CARE database and data from in-depth accident statistics. The changes in safety will be estimated based on i) the change in the frequency of driving scenarios per traffic scenario, and ii) the change in the number of accidents per accident type and per driving scenario. – Definition of principles for sensitivity analysis. This refers to the identification of variables with high uncertainties and/or high influence on results. Calculation of quantitative estimates on safety impacts of ADFs on EU level by considering the estimated penetration rates (Chapter 4.1.3). The estimates will be calculated in terms of prevented fatalities and injury accidents (fatal, severe and slight injuries) and material damage (for parking ADF only).

4.5.2 Evaluation plan

As mentioned earlier, the Level-2 research questions regarding the safety impact assessment relate to frequency of accidents and accident severity (Table 2.3). The following text summarises the evaluation plan for them, i.e. how these research questions will be answered during the L3Pilot project.

RQ-I1: What is the impact of ADF on the number of accidents in a certain driving scenario / for certain road users?

This research question will be answered by means of the planned stochastic simulation of selected scenarios for the mature motorway and urban ADFs. For the mature parking ADF, the analysis is conducted by means of a field of application analysis.

For each scenario, the proportion of accidents related to conducted simulations is calculated both for the baseline and treatment scenarios. The relative safety effect for each scenario can then be calculated based on these accident proportions in the baseline and treatment condition with comparison of both conditions by means of standard statistical approaches (e.g. confidence interval, effect size).

The research questions related to accidents involving other road users such as pedestrians and cyclists will be answered by means of the VRU-related scenarios on motorway and urban environments.

RQ-I2: What is the impact of ADF on accidents with a certain injuries level / damage in a certain driving scenario?

This research question will be answered by further analysing the accidents detected during the safety simulations with respect to the position and velocity of the involved traffic participants. This will be done to identify the relevant accident types occurring within the safety simulation scenarios. The accident situation determined (e.g. position and velocity) during the simulations will be transferred into injury level of the involved traffic participants by applying the appropriated injury risk functions to it. This analysis addresses all the research questions regarding the impacts of ADF on accidents with different levels of severities (fatal, severe and slight injuries) and accidents with material damage only. The safety effects will be estimated in terms of change in the number of fatalities and injury accidents. However, the more detailed information on changes within different injury severities (severe and slight) are also calculated based on the injury risk functions, since they are needed for the socio-economic assessment (further described in Chapter 4.7 of this deliverable).

Improvements in the rescue chain after an accident has occurred are not in the direct focus of the L3Pilot project. Therefore, the last research question on the impact of ADF on the rescue chain in terms of preventing injuries can only be addressed by qualitative analysis that points out the potential effects that might or might not occur in combination with the market introduction of ADF.

4.5.3 Scaling up

The calculation of safety effects per ADF in terms of number of accidents will be conducted based on the formula below. The output of this formula is the total number of accidents after the introduction of ADF. The target accident definition takes into account the ODD of each ADF. The assessment of safety impacts varies between the investigated ADFs mainly due to their different nature and the input data used in the assessment (e.g. different types of road accident databases). The adaption of the formula for the different ODDs is described in the text following the formula.

$$\sum_{j} T_{i,j} \cdot \Delta f(S_j) \cdot \Delta I_i(S_j)$$

$T_{i,j}$ = Number of target accidents with severity *i* for scenario *j*

$\Delta f(S_j)$ = Change in frequency of scenario *j*

$\Delta I_i(S_j)$ = Change in accident risk with severity *i* for scenario *j*

The formula was further adapted from Rösener et al. (2018). The main difference is that the L3Pilot version includes also the number of target accidents. Note that the value for both deltas in the formula is 1.00 for 0% change, 0.95 for -5% change, 1.05 for +5% change, etc.

The driving scenarios considered in the safety simulations include a subset of the driving scenarios addressed in the technical & traffic evaluation (Chapter 4.2) focusing on safety-critical scenarios among them. In addition, some other potentially safety-critical situations such as the minimum risk manoeuvre, wrong activation of the ADF, or an obstacle in the lane are addressed. A third group of driving scenarios considered here are those triggered by a change in road infrastructure (e.g. an end of lane or a change in speed limit).

Motorway and urban ADF

The number of target accidents for motorway and urban ADF will be defined based on the accident data reported to the European-wide CARE database (European Commission 2019). However, since the data reported to the CARE database has some limitation regarding the details and coverage of the data, the in-depth road accident data from selected countries will be used to complement this analysis. The target accidents (or situations) for scenarios related to 'new' accidents and situations which are defined as challenging to motorway ADF will be mainly estimated based on the prevalence of these situations in traffic (e.g. occurrence per distance driven).

The change in the frequency of driving scenarios and the change in accident risk for each scenario for motorway and urban ADF will exploit inputs from two types of simulations:

1. Simulation of traffic scenarios to assess the effect of ADF on the frequency of driving scenarios (e.g. number of cut-in conflicts per travelled distance).

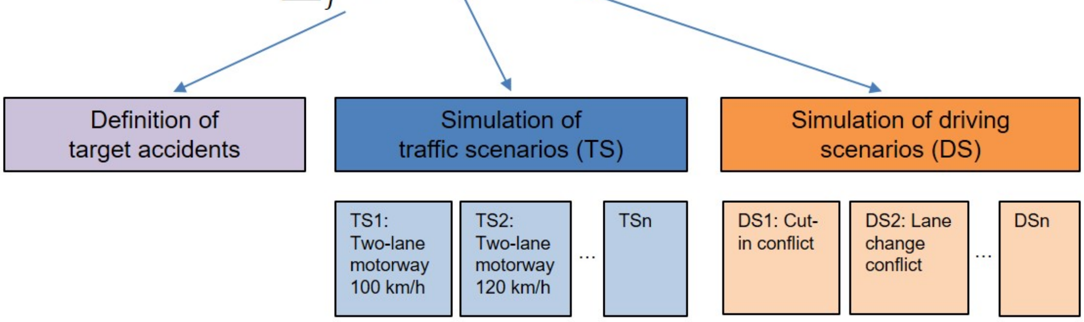
2. Simulation of driving scenarios

- Driving scenarios where the potentially positive effects of ADF are expected (e.g. lane change or conflicting paths with another vehicle at an intersection). The distribution of target accidents to single driving scenarios will be estimated based on in-depth accident data.
- Driving scenarios considered challenging for automated vehicles (e.g. end of lane, change in speed limit) to assess the potentially negative effects of ADF. The frequency of these scenarios will be estimated based on the prevalence of these situations in traffic (e.g. occurrence per distance driven).

The analysis cannot be limited to short driving scenarios only, as the interactions in traffic between different traffic participants is also relevant for ADF. Here, it must be taken into account that the ADF operates for much longer time frames than an active safety function which intervenes only close to an imminent accident. Therefore, the safety impact assessment of motorway and urban ADF covers also the impacts of ADF in a larger traffic context (traffic scenarios), as some effects are expected to change with penetration rate.

The overview of the above process and its relation to the earlier presented formula is presented in Figure 4.5.

$$\sum_j T_{i,j} \cdot \Delta f(S_j) \cdot \Delta I_i(S_j)$$



Definition of target accidents						Change in frequency of DS in TS		Change in accident risk
All fatal accidents	Accidents inside ODD	Side collisions	… in TS1	… in DS1	·	$\Delta f(DS1)$	·	$\Delta I(DS1)$
				… in DS2	·	$\Delta f(DS2)$	·	$\Delta I(DS2)$
				… in DSn	·	$\Delta f(DSn)$	·	$\Delta I(DSn)$
			… in TS2	Similar calculation as for TS1				
			… in TSn	….				
		Rear end collisions	Similar calculation as for side collisions					
		…	Similar calculation for all accident types					
	Accidents outside ODD							

Figure 4.5: Approach to the calculation of safety effects of motorway and urban ADF in terms of number of accidents.

The driving scenarios and traffic scenarios to be simulated for the safety impact assessment of motorway and urban ADFs are presented in Annex 3. The impact on single driving accidents is an important class not covered by these scenarios. It is treated differently, since these scenarios are not related to conflict due to interactions between traffic participants. These scenarios are mainly related to an issue in the vehicle control and results rather in challenges for the vehicle dynamics. The applied simulation tools focus rather on interaction-related conflict types. Therefore, the single driving conflict will not be addressed by these simulation tools. Instead these scenarios will be assessed based on expert opinion similarly to the approach for parking ADFs (described below). Here, it must be considered fairly unlikely that the automated vehicle will leave the lane as long as the vehicle is in its ODD (lane markings are required) and no functional issue occurs.

Parking ADF

For the safety impact assessment of the parking ADF, the estimation of target accidents will be conducted by identifying the theoretical maximum percentage of insurance claims in accident databases that can be addressed (and theoretically prevented) by the respective parking ADF. For this purpose, the insurance claims databases (e.g. Allianz, other insurance companies) will be filtered according to the ODD of the respective parking ADF, as the parking accidents are not well represented in the CARE database. In addition, this step includes estimation of the share of parking manoeuvres that take place within the ODD (for a vehicle that has this ADF in use). Insurance claims offer a comprehensive basis for the evaluation, especially regarding parking accidents, since parking and manoeuvring accidents (excluding those involving personal injuries) are in most cases reported only to the insurance company and not the police.

The effect of parking ADF (i.e. the change in the accident risk) will be estimated by identifying the share of (relevant) claims (i.e. target accidents) that can be prevented under real road traffic conditions (e.g. bad weather or road conditions or technical failures reduce the efficiency of a certain system). This estimation will be made based on the description of the mature parking ADF descriptions and on data from the field experiments in L3Pilot. The data from field experiments concerns, in this case, the frequency of safety incidents while parking and other technical characteristics of parking. Presumably, the main restriction limiting the effect of parking ADF is the ODD of the studied ADF. For instance, a limitation of a function to home zone parking will limit the availability of the ADF to a rather small subsample of parking manoeuvres.

See Annex 4 for a more detailed description of the safety impact assessment method for the mature parking ADF.

Overall remarks

There are some variations on how the penetration rate estimates will be considered in the above calculations. In safety simulations (for motorway, traffic jam and urban ADF), the penetration rate is one of the parameters for defining a traffic scenario to be simulated, and it will then be included in the above calculation of the safety effects. For parking ADF, the result of the calculation of safety effects will be multiplied by the penetration rate at the end. The penetration rates used in impact assessment were defined as part of L3Pilot Deliverable D3.3 (Metz et al., 2019) and are also presented in Chapter 4.1.3 of this document. These penetration rates are the same for all ADFs: 5%, 10%, 30% and 100% of cars with ADF in use within the ODD.

The results of the above calculations serve as one important input in the upscaling of the results to derive the safety impact of the analysed ADFs in terms of number of prevented fatalities and injury accidents at EU level. In addition, the scaling-up process will exploit inputs from a wide variety of sources, such as from L3Pilot supplementing studies, literature, statistics (mobility, vehicle fleet etc.), and in-depth road accident statistics to identify the effects of ADF on each relevant safety mechanism (see Innamaa et al., 2018) via expected changes in vehicle, driver, and road user behaviour.

4.6 Efficiency and environmental impact assessment

4.6.1 Overall approach

The main objective of the efficiency & environmental impact assessment is to examine the impact of ADF on traffic efficiency and the environment (Level-1 research question). These research questions are studied in terms of changes in macroscopic traffic flow (e.g. travel times, throughput) and greenhouse gas emissions (e.g. CO_2).

The overall methodology for the efficiency & environmental assessment is shown in Figure 4.6. Both direct and indirect impacts are considered. Direct impacts are those resulting from changes in individual vehicle operation, such as differences in speed or in time headway. The results of the technical & traffic evaluation will be taken into account when defining the parameters of the mature ADFs for the simulations. Indirect impacts derive from changes in other impact areas: If the ADF reduces accidents, incident-induced congestion is reduced; or if the motorway ADF makes driving on motorways more attractive and changes mobility behaviour to favour motorways over other road types, vehicle kilometres travelled (VKT) on motorways are affected.

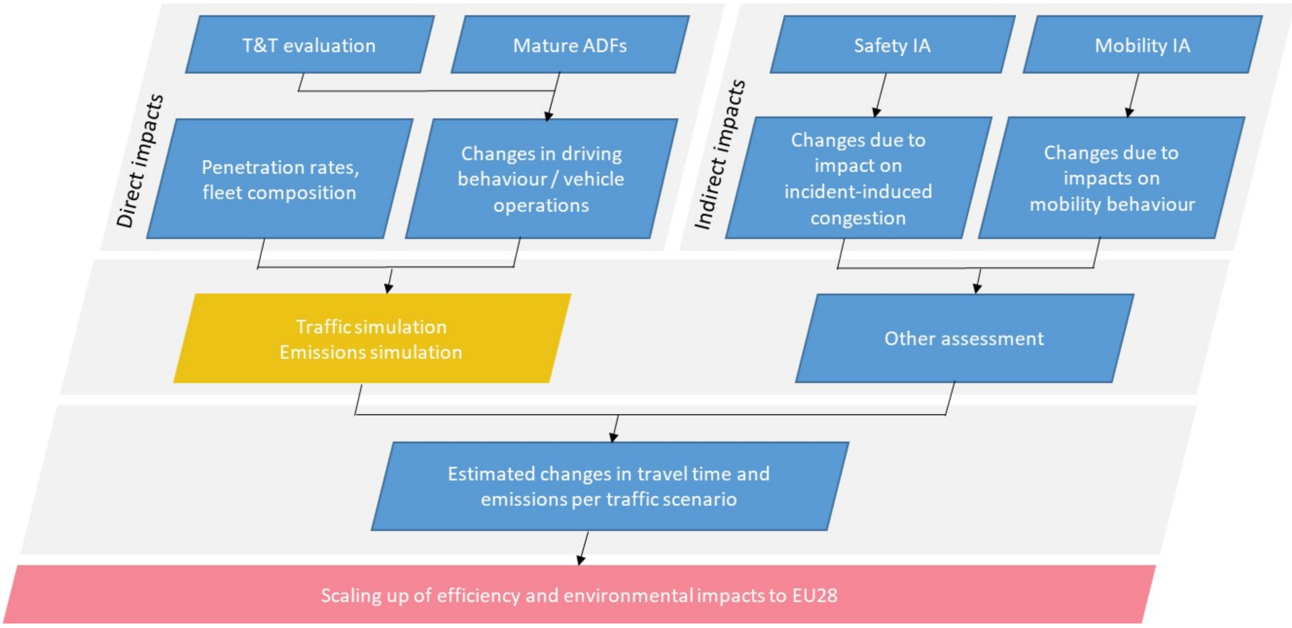


Figure 4.6: L3Pilot approach to efficiency and environmental impact assessment.

The efficiency & environmental impact assessment considers the L3Pilot motorway, traffic jam and urban ADFs. The parking ADF is excluded from the direct assessment due to its different nature – the travel time used for the parking manoeuvre, or the emissions produced during the manoeuvre, are such a small proportion of the overall travel time and emissions that we exclude them from the impact assessment.

The effect will be examined with traffic microsimulation software for a number of different road sections with respect to road layout and speed limit. The outcomes of the simulation (e.g. travel time or emissions) will be assessed over the simulation runs of baseline and treatment condition for each scenario as well as in comparison of both conditions by means of standard statistical approaches (e.g. confidence interval, effect size).

Several traffic scenarios with different traffic volumes (low, medium or high) and fleet compositions will be simulated and the network performance in terms of travel times, throughput etc. investigated. The fleet composition includes, in addition to the ADF penetration rates, the share of heavy-duty vehicles and share of electric vehicles in the traffic flow, which have a large impact on efficiency and environment. Also, the size and age of vehicles are important. In addition, parameters used in simulation of manually driven vehicles need to be decided for the baseline and for the non-AD in treatment. All parameters are set to represent the current European fleet, in line with the snapshot approach taken in the socio-economic impact assessment.

For the motorway and traffic jam ADFs, a number of sections of motorway with varying speed limits and number of lanes, as well as smaller segments around on-ramps, off-ramps and weaving sections, will be examined. The simulated 'road layout - speed limit' combinations are selected so that they together represent a majority of the European motorway network.

For the urban ADF, a variety of scenarios with varying urban speed limits will be considered. This will be combined with a selection of lane counts including one-way streets. The scenarios are selected to represent the road network layout in a selected set of European cities, including those where piloting activities are located.

4.6.2 Evaluation plan for efficiency impact assessment

An overview of all transport network efficiency-related research questions is given in Table 2.3. Efficiency impacts will be studied with traffic micro-simulations using PTV VISSIM. The traffic scenarios to be simulated are selected in such a way that it is possible to scale up the results to EU level. The motorway and traffic jam ADF will be studied in more detail (simulations covering a larger proportion of different road layouts in Europe), while the urban ADF is studied on a more general level due to the complexity of the urban networks and less data being available for scaling up.

The Level-3 research questions and preliminary indicators to be used in the assessment are described in Table 4.20 for the efficiency impact assessment. This research question RQ-I3 on traffic efficiency impact will be studied in terms of indicators, such as throughput, density, travel time, delay, and their standard deviations. In addition, network performance in terms of capacity is examined.

Table 4.20: Preliminary indicators for the efficiency impact assessment.

RQ Level 3	Motorway and traffic jam ADF	Urban ADF
RQ-I3a: What is the impact of the ADF on throughput in a road section or intersection?	Changes in throughput on line sections and areas with on-ramps, off-ramps or weaving sections in the simulated traffic scenarios.	Changes in the throughput of intersections as well as the waiting time at intersections.
RQ-I3b: What is the impact of the ADF on reliability of travel time?	Indicators related to travel times and average speeds and their standard deviation.	Indicators related to average speed and waiting time at intersections.
RQ-I3c: What is the impact of the ADF on travel times?	Change in travel time and its standard deviation; change in average delay and its standard deviation.	Change in travel time and its standard deviation; change in waiting times and crossing times at intersections as well as their standard deviation.
RQ-I3d: What is the impact of the ADF on speed differences between vehicles?	Speed differences between vehicles will be studied as part of the research questions regarding throughput and travel times. Large speed differences between vehicles cause heterogeneous traffic flow.	Speed differences between vehicles and VRUs will be studied as part of the research questions regarding throughput, travel times and interaction distances for VRUs.
RQ-I3e: What is the impact of the ADF on network capacity?	Qualitative assessment taking into account changes in average time headways, speeds and traffic homogeneity.	Qualitative assessment taking into account changes in average time headway, encroachment times, waiting and crossing times as well as speeds.

4.6.3 Evaluation plan for environmental impact assessment

An overview on all environment-related research questions is given in Table 2.3. Impacts of ADF on the environment will be studied with EnViVer, an emission modelling software calculating the emissions for the traffic simulations made with VISSIM. Thus, the environmental impact assessment takes as input VISSIM output files of efficiency assessment.

Changes in average emissions (within the fleet composition to be defined) will be estimated based on the changes in speed profiles of vehicles in the specified traffic scenarios. However, in order to capture changes in energy demand independently of the fuel type of vehicles, also the changes in physical energy demand with mature ADF will be investigated using the equation

$$\frac{E_{demand}}{s} = F_{total} = F_{air} + F_{roll} + F_{accel} + F_{grad}$$

as introduced in Chapter 4.2.1. This change in physical energy demand of automated vehicles in the simulated traffic scenarios will be studied for pre-defined personal vehicle types with an energy-flow simulation tool.

The Level-3 research questions and preliminary indicators to be used in the assessment are presented in Table 4.21 for the environmental impact assessment.

Table 4.21: Preliminary indicators for the environmental impact assessment.

RQ Level 3	Motorway and traffic jam ADF	Urban ADF
RQ-I4a: What is the effect of the ADF on fuel consumption?	Change in fuel consumption of a pre-defined fleet composition	Change in fuel consumption of a pre-defined fleet composition
RQ-I4b: What is the effect of the ADF on energy demand?	Change in physical energy demand	Change in physical energy demand
RQ-I4c: What is the impact of the ADF on CO_2 emissions?	Change in CO_2 emissions of a pre-defined fleet composition	Change in CO_2 emissions of a pre-defined fleet composition

4.6.4 Scaling up

The direct impact for the efficiency and environmental indicators specified above is calculated based on the formula below. The assessment of impacts varies between the investigated ADFs mainly due to their different nature and road environments, and the input data available for assessment (e.g. road infrastructure, traffic volumes). Impact estimates are received for different fleet compositions (ADF and heavy-duty vehicle penetration rates) within different traffic scenarios.

$$Impact_{PR,R} = \sum_t (x_t * VKT_{t,R} * \left(\Delta x_{t,PR} + 1\right) * (1 + \Delta k_{t,PR,R})) - \sum_t (x_t * VKT_{t,R})$$

or to make it simpler

$$Impact_{PR,R} = \sum_t (x_t * VKT_{t,R} * (\Delta x_{t,PR} + \Delta k_{t,PR,R} + \Delta x_{t,PR} * \Delta k_{t,PR,R}))$$

where:

- *R:* Region under investigation (NUTS3/EU)
- *t:* Traffic scenario under investigation (incl. number of lanes, speed limit, traffic volume, share of heavy-duty vehicles)
- x_t: Indicator, e.g. emissions / travel time / delay per VKT in traffic scenario *t* (baseline)
- $VKT_{t,R}$: VKT in traffic scenario *t* in region *R* (baseline)
- $\Delta x_{t,PR}$: Change (%) in emissions / travel time / delay per VKT in traffic scenario *t* (result of simulation) with ADF penetration rate *PR*
- $\Delta k_{t,PR,R}$: Change (%) in VKT in traffic scenario *t* (indirect impact) with ADF penetration rate *PR* in region *R*

The effect on an environmental or efficiency indicator (such as total CO_2 or total delay) for different penetration rates of ADF is estimated with the simulations for different types of infrastructure and traffic conditions. In the overall impact calculation, these effects are weighted according to their distribution in the region under investigation (NUTS3 or EU).

The infrastructure and traffic variables form the traffic scenarios under investigation. These differ for motorways and urban environments, as well as for basic motorway sections and ramp or weaving segments. For example, basic motorway sections without ramps or weaving sections consist of two or three lanes with speed limits from 80 km/h to unlimited (see Table 4.22), when the road types that represent more than 1% of infrastructure in the EU were selected for the simulation. In addition, typical ramp and weaving segment configurations will be simulated.

Table 4.22: Infrastructure combinations for simulated motorway traffic scenarios.

Number of lanes (per direction)	Speed limit (km/h)
2	80
	90
	100
	110
	120
	130
	140
	No limit
3	100
	110
	120
	130
	No limit

Traffic scenarios for urban areas are selected based on their occurrence in 30 selected European cities, including those with piloting activities within the project. The number of lanes as well as the speed limit is grouped (cf. Table 4.23). In contrast to motorway scenarios, the number of lanes is the total number of lanes and can be for one or two directions. This is due to one-way and restricted access roads. These segments will be connected using typical intersections and augmented using traffic lights, pedestrian crossings and other urban street network elements.

Table 4.23: Infrastructure combinations for simulated urban traffic scenarios.

Number of lanes	Speed limit (km/h)
1	30
	50
2	30
	50
3	50
4	50
4+	50

The EU level data for upscaling of traffic efficiency and environmental impacts leans on the work of the ecoDriver project (Jonkers, Wilmink, Nellthorp, Gühnemann, & Olstam, 2016), extending it to the assessment of Level 3 automation. In the ecoDriver project, different driver assistance systems for promoting eco-driving were developed and tested, and their potential impacts on efficiency and the environment assessed with field tests and simulation. Afterwards, the local impacts were scaled up to EU level, for which a new methodology was created.

The dataset from the ecoDriver project provides a general overview on efficiency and environmental indicators in Europe on a NUTS3 level, split into motorways, urban roads and rural roads (the latter being beyond the scope of L3Pilot). In L3Pilot, this data is enhanced with the categories necessary to reflect the ODD definitions for the mature ADF. For example, it is necessary to add the dimension of weather to the dataset to be able to estimate the percentage of time that the ADF can be used, as they cannot be used in heavy rain or snow. More detailed data on traffic and emissions will be used for regions for which this kind of data is available. However, the ecoDriver data serves as a basis for scaling up for those areas where more detailed data is not available – enabling the scaling up to EU level. Table 4.24 provides an overview of the data sources used in scaling up of efficiency and environmental impacts.

Table 4.24: Overview of data sources for scaling up of efficiency & environmental impacts.

Data type	Data source	Use
Weather	European Centre for Medium-Range Weather Forecasts	Share of time/VKT driven within ODD
Traffic volumes	ecoDriver, statistics, expert assessment	Covering different traffic conditions, representing European conditions
Infrastructure (number of lanes, speed limits, intersection density)	OpenStreetMap	Setting up representative traffic scenarios
Current fleet composition	Statistics	Setting up representative traffic scenarios

When scaling up impacts to EU level, the ODD both in space and time needs to be taken into account. The ODD in space is considered static. It represents the share of the network where ADF can be used, in terms of the infrastructure dimensions such as road type and layout. In contrast, the ODD in time is dynamic and represents the share of time where ADF can be used. For example, periods of time with heavy rain or snow are excluded. Both ODD dimensions – space and time – need to be suitable at the same time for the ADF to be used (see Figure 4.7). The parameters for ODD definition are defined in the mature ADF descriptions (Chapter 4.1.1).

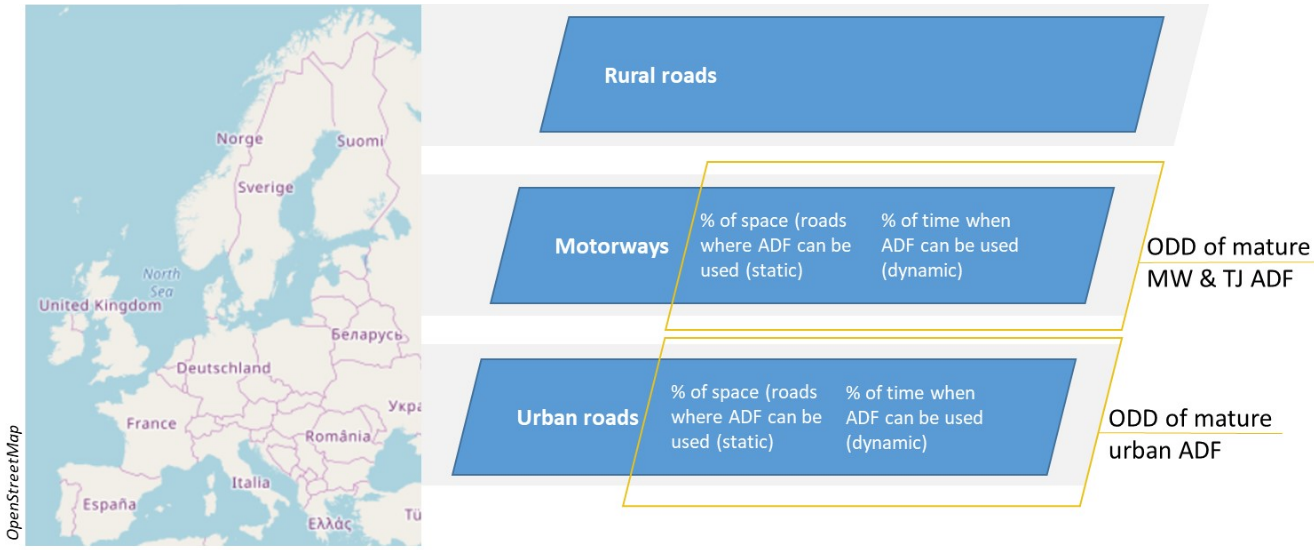


Figure 4.7: Description of ODD for scaling up.

4.7 Socio-economic impact assessment

4.7.1 Overall approach

The Level-1 research question for the socio-economic impact assessment in L3Pilot is:

> *What is the overall socio-economic impact of ADFs designed for motorway, traffic jam, urban and parking environments?*

The assessment of the socio-economic impacts is concerned with the net welfare gain for the society regarding each of the ADFs in question. Net economic impacts may also be estimated for different stakeholders. The purpose of the socio-economic impact assessment is to clarify whether the benefits of equipping vehicles with ADFs outweigh their costs. If this is the case, the implementation of the technology is considered to be beneficial to society and/or the stakeholders. The research questions are presented in Table 2.4.

Cost-benefit analysis

A commonly used method of socio-economic impact assessment is cost-benefit analysis (CBA). CBA is a systematic method for evaluating expected costs and benefits of the implementation of a project (for example, investment in transportation systems, a technology, or a service), compared to a situation without the implementation of that project. CBA is grounded in welfare economic principles and is widely used to clarify whether the implementation of a project is beneficial from the society's point of view. Therefore, CBA often serves as an important support tool for the authorities in their decision making.

Traditional CBA approach

In CBA, a reference scenario is compared to one or more alternative treatment scenarios. The reference scenario serves as a reference point, also called baseline, which reflects how the world is expected to develop in the absence of a particular project (treatment). The treatment scenarios reflect how things are expected to develop with the intervention. Then the impacts of the project are measured by the differences between each treatment scenario and the baseline scenario.

In many cases, such as for transportation-related investment projects, the benefits and costs of a project accrue over a period of time, for example 10 to 30 years. Public authorities are usually in charge of these investment projects. The main policy concern for the authorities is to clarify whether such an investment is a good way of using society's scarce resources and to decide whether to invest in the project or not. This is an ex-ante perspective. CBA analysis depends on the provision of estimates on the value of future costs and benefits due to the specific project. By assigning economic values to the impacts of a project over a specified time period, the net present value of the project can be calculated. The net present value represents the total present value of the project's benefits minus the present value of all costs over the project's life cycle.

Simplifying the traditional CBA approach: From net present value to net annual value

In L3Pilot, the experiments are based on testing of vehicles (prototypes) equipped with ADFs in real-life-traffic situations, thereby providing the basis for assessing the impacts of ADFs with regard to traffic safety, efficiency, and the environment. Since the ex-ante perspective is forward-looking, it is necessary to compare an expected development without ADFs (baseline) to a situation with these ADFs (alternative scenario/s).

To fully comply with the ideas of traditional CBA, it is necessary to forecast reliable future scenarios in order to calculate the net present value of the different ADFs. However, since the ADF technology is already in the development phase, it does not make sense to construct a future scenario without this technology. In addition, it is extremely challenging to construct reliable scenarios by forecasting how the traffic situation would develop with the ADFs in question for the next 10-30 years in Europe. Such forecasting should rely on assumptions with regard to the technological progress in the automotive industry, legislation, government policy, and consumer demand for ADFs, all of which are factors that will determine to what extent the technology in question is implemented, and when. Furthermore, another challenge is related to the parallel development in other areas, which will affect people's mobility choices such as sharing economy, new mobility concepts, electrification of vehicles, urbanisation, etc. Therefore, the impacts of these factors with and without the ADFs need to be predicted, which is an impossible task.

The arguments above imply that the most common steps in conducting a CBA are not quite applicable to the socio-economic assessment in L3Pilot. This is because regardless of how the scenarios are constructed, the reliability of their content will be highly questionable. This will draw attention away from the essence of the CBA, which is to calculate the social net value of the ADF. Therefore, rather than trying to apply a traditional CBA approach with forecasting of how the world is expected to develop over the coming decades with and without the ADFs, we have chosen to simplify the traditional approach.

The simplification framework is as follows: The baseline scenario is the world of today, which is without the technology in question. Alternative scenarios are created to investigate how safety, efficiency and the environment would change if a proportion of the vehicles in the world of today were operated with ADFs. In this way, it is possible to capture the pure impacts of implementing ADFs. We call this a *snapshot approach*.

By using the simplified approach above, it is investigated how much higher or lower the welfare would be if (a certain proportion of) today's vehicles were operated with ADFs. This means that a context is created where nothing happens except for the introduction of vehicles with the relevant ADF. Apart from the impacts due to the ADF in question, everything else such as infrastructure, available travel modes and travel behaviour remains the same as in the baseline scenario. Economists will call this approach a comparative static analysis, *ceteris paribus*. It implies that the society is in a steady state, in an otherwise fixed and stable situation, meaning that nothing happens except for the introduction of vehicles with ADF technology in the current traffic situation.

The advantage of using this snapshot approach is that official public statistics for the latest available year will form the basis for describing what the world would look like without the new technology. In L3Pilot, it will presumably be the year 2017 or 2018 depending on the availability of the relevant accident statistics and other statistics. This means that the baseline in this approach, reflecting the world of today without ADF technology implemented, can build on existing statistics regarding accident rates, accident severity, traffic flow, and so on.

Calculating the economic values of ADFs in the simplified approach will require some further modification. Since it is chosen to focus on a snapshot of the net welfare effects in one year, it implies that the traditional net present value is to be replaced with the annual net benefits. Positive annual net benefits imply that implementation of ADFs is beneficial for the society and vice versa.

<u>*Operationalising the snapshot approach*</u>

Ideally, the penetration rate for the new technology in use should be *endogenous* (i.e. determined by how potential users perceive the benefits of vehicles with the relevant ADFs compared to vehicles without and other ways of travelling than driving). This does not, however, make sense in the simpler snapshot approach applied for investigating the socio-economic impacts in L3Pilot. Different exogenous penetration rates will be experimented instead.

The snapshot approach implies that the size of the fleet of vehicles eligible for the relevant ADF is constant. In treatment scenarios, each vehicle with the ADF replaces exactly one without. This implies that apart from the fact that some drivers choose vehicles with the relevant ADF, all travellers continue to travel in the same way as before. The personal mobility impact assessment may reveal some effects in terms of route change (e.g. in favour of motorways that could be expected due to the use of motorway ADF). However, these will most likely have no major effects on travel behaviour and traffic patterns, and hence, there should be no serious objection to the exogenous penetration rates approach.

In order to conduct the analysis along the lines above, it is reasonable to assume that the new vehicles that are introduced in the context of today replace the oldest ones in the current stock of vehicles. This is in accordance with how the real-world pilot testing in L3Pilot is conducted. The pilot tests compare driving with the use of Level 3 ADF with manual driving. As lower level automation already exists in the current stock of vehicles, we need information on the current distribution of vehicles across them to examine how well the baseline for the socio-economic assessment resembles the baseline used in the real-world testing in L3Pilot.

The next step is to elaborate on the magnitude of impacts by assuming different penetration rates for the new technology:

- What does it mean if the ADF is installed and activated in 100% of the current stock of vehicles?
- What does it mean if the ADF is installed and activated in 5%, 10%, or 30% of the current stock of vehicles?

The scenario with 100% penetration rate is meant to capture the (theoretical) full-potential impacts of the ADF. In the real world, this is as if the installation of the technology, as well as its active usage, were mandatory in all vehicles.

The lower penetration rates indicated above reflect the most realistic future deployment of vehicles with ADF. In either case, the impacts are estimated within the ODD defined by the current infrastructure. It is also desirable to consider the impacts if the ODD were enlarged, which would also require information on costs with respect to infrastructure improvements.

Applicability of snapshot approach in CBA

The proposed snapshot approach outlined above differs from the traditional CBA approach. However, it is not unique to L3Pilot. The euroFOT project also used a snapshot CBA approach for investigating the impacts of in-vehicle functions on traffic safety, traffic efficiency, and the environment (euroFOT, 2012). The in-vehicle functions in the CBA included forward collision warning (FCW) and adaptive cruise control (ACC), which were grouped into one bundled function and scaled up to EU level. Assuming that 100% of the cars were equipped with the relevant FWC/ACC technology, the expected annual impacts for the whole EU, given the traffic context of 2010, were estimated with respect to traffic safety, traffic efficiency, and the environment.

The snapshot approach in euroFOT has clear similarities with the one in L3Pilot. Neither constructs a future for the next 10 to 30 years with or without the technology. Both use the current situation as the baseline. Both manipulate the current situation to illustrate what it would look like if all or a fraction of the vehicles were equipped with the technology in question. However, a snapshot is considered an unusual approach when it comes to CBA in general. That is why, in contrast to euroFOT, we have provided explicit arguments for why this approach is preferable.

4.7.2 Evaluation plan

Assessing the socio-economic impacts of ADFs

The impacts of ADF technology are assessed based on real world testing of ADF. Also the socio-economic impacts will be assessed for each type of ADF (i.e. for four different types of ADF technology separately, and not for Level 3 ADF technology in general where the equipment is bundled as one technology).

Separate analyses of socio-economic impacts for each type of ADF equipment ensure that we avoid the risk of double counting the impacts. Further, it takes into account that drivers have different preferences. Not all drivers will be interested in equipping vehicles with a bundle of ADFs. Some may be interested in one particular ADF only and hence consider the costs for purchasing a bundle of ADFs too high. On the other hand, since some sensors in an ADF (e.g. motorway ADF) also serve in an urban environment, the marginal cost of equipping a vehicle with several ADFs will not be as high as adding up the costs of each separate one. This is a disadvantage of conducting separate analyses.

Levels of aggregation in the socio-economic impact assessment

The socio-economic impact assessment is carried out at EU level. It is challenging to identify all costs and benefits of ADF and to upscale these values to depict socio-economic impacts at the EU level. Sensitivity analyses are a good way to illustrate the effects of uncertainties and lack of information in CBA.

Another way to deal with such issues is to conduct CBAs at lower levels of aggregation than the whole EU. In-depth studies of one or two countries with the most reliable information foundation could be useful in this respect. Furthermore, in-depth studies of representative travellers in specific travel situations, such as commuting from suburban areas to city centres, may provide important additional information which is more detailed and easier to interpret. Here the results of the mobility impact assessment can be used.

Impacts of ADF technology for different stakeholders

The welfare effect for a society is the difference between benefits and costs to all stakeholders in the society (i.e. individuals/households, producers and the government). In the context of L3Pilot, individuals are not only those who travel with ADF-equipped vehicles, but also those who travel using conventional vehicles or other modes of transportation (e.g. bus, train, bicycle, etc.). These are travelling individuals. In a society, there are also individuals who do not use the above travel modes and hence are considered non-travelling individuals. Producers are all manufacturers of vehicles and suppliers of ADF technology, while the government in this setting consists of the authorities in charge of road infrastructure and providing healthcare for the society.

In the CBA, it is convenient to distinguish between the following types of stakeholders: users of ADF-equipped vehicles, other travellers, producers, the government, and the rest of the society.

Travellers

When choosing how to travel from one place to another, travellers consider the cost of travelling, which includes direct travel costs, travel comfort, and travel time. Travellers may also consider the risk of possible accidents. Travellers' expected welfare will increase if the expected travel or accident cost decreases and/or its variance (or standard deviation) is reduced.

Generalised travel cost is often used in transport economics for measuring travellers' costs (see e.g. Bruzelius, 1981). It consists of the sum of monetary and non-monetary costs of transport/trip, for example costs of fuel, parking, tolls, value of travel time, etc., for travelling individuals. A decrease in expected generalised travel cost or its variance (or standard deviation) will imply higher welfare for travellers.

In L3Pilot, benefits and costs related to travelling are mostly determined by how the ADF technology affects the frequency and severity of traffic accidents, as well as travel time. Drivers of vehicles with ADF technology may also experience that driving becomes more comfortable, and that time spent on the road may add some productive or recreational value because the ADF technology assists the driver during the trip. Such effects are taken into account as a reduction of generalised travel cost. Net welfare gains for these drivers depend on the extra monetary costs of ADFs (market price of purchasing ADFs, maintenance costs, etc.).

Furthermore, other travellers may also be affected, as they are likely to experience that expected accident costs or travel costs will change. To the extent accident rates decrease for vehicles equipped with ADF, it will also reduce the risk of conventional vehicles being involved. Changes in accident rates will in turn also affect traffic flows and travel time for travellers in all vehicles on the road. For travellers in conventional vehicles, these impacts occur with no extra monetary costs such as buying ADF equipment. Such indirect effects represent *externalities*. By externalities, it is meant that other travellers – whether those who drive conventional cars, those who use other travel modes, or pedestrians – may gain from the benefits of ADF technology without having to pay for it directly. This is, in fact, a positive externality. In such situations, government intervention may be welfare improving. Intervention by the government can be in the form of subsidies provided for the drivers of ADF vehicles (e.g. reduction in or exemption from road charges) and to have plans for making the equipment mandatory.

Producers

Producers are manufacturers of vehicles with ADF. If the ADF technology allows producers to increase their profit margins, the net gain of producer surplus should be included in the CBA. However, we do not expect that competition, and hence producers' profit margins, will be affected by the introduction of ADF technology, which means that this possible impact may be neglected. Therefore, producers' perspective is not part of the CBA in L3Pilot.

Government

In socio-economic impact assessment, it is necessary to keep track of the effect on the government's budget (called *budget effect*), not because the government in itself counts in the welfare calculations, but because a positive budget effect gives rise to extra public services and/or tax relief for taxpayers. A negative effect on the government budget has to be financed through reduced service production and/or tax increases. Since taxation has negative incentive effects, it is usually assumed that an (extra) euro into/out of a public purse is worth more than an (extra) euro into/out of a private purse.

In L3Pilot, relevant budget effects may be direct effects related to using vehicles with ADFs, but more importantly, budgetary consequences are also associated with necessary infrastructure investments. If infrastructure investments or adjustments to the existing road infrastructure are not required for the operation of vehicles with ADFs, then there will be no budgetary consequences for the government. In the snapshot approach, impacts of ADF are detected within the current infrastructure context, which means that infrastructure-related costs do not need to be included in the CBA. However, when considering impacts, if the functionality of ADF technology is extended by improving road infrastructure, these costs should be taken into account in the analysis.

The government is also the main provider of healthcare services. A reduction in the number and severity of traffic accidents is likely to reduce public expenses for the handling of accidents and injuries such as costs related to the involvement of police and rescue teams along with healthcare costs for the injured. The CBA should, therefore, include these budgetary effects.

Rest of society

Socio-economic analyses of costs and benefits of a project for the society at large generally include a consideration of how the project is expected to impact the environment with regard to noise, pollution and the like. In L3Pilot, the focus is on CO_2 emissions derived from information on how the ADF technology affects fuel consumption in driving. We do not expect any other environmental impacts.

Within economics, it is also well established that significant improvements in travel efficiency may enlarge labour markets and enhance productivity (Venables, 2007). This is referred to as wider impacts. There is, however, no reason to expect that vehicles with Level 3 ADF will affect traffic efficiency in a way that such wider impacts can be expected. Thus, the potential for wider impacts regarding changes in the labour markets and productivity may be negligible. Realistically, notable impacts on traffic efficiency, and hence effects on the labour markets and on productivity as such, may be expected to arise from higher automation levels than Level 3 ADF. Therefore, the topics of wider impacts will most probably deserve to be addressed when investigating the impacts of Level 4 and 5 ADFs, where the vehicles handle all driving tasks or are equipped with fully automated driving with no need for a driver.

<u>*Assessment of costs and benefits*</u>

The core element of the socio-economic impact assessment relies on the outcome of impact assessments, revealing the potential benefits of ADFs in terms of traffic safety, efficiency, and environmental effects. All potential impacts of ADFs should in principle be captured in the socio-economic impact assessment as benefits and costs. Figure 4.8 illustrates an overview of the overall socio-economic impact assessment in L3Pilot.

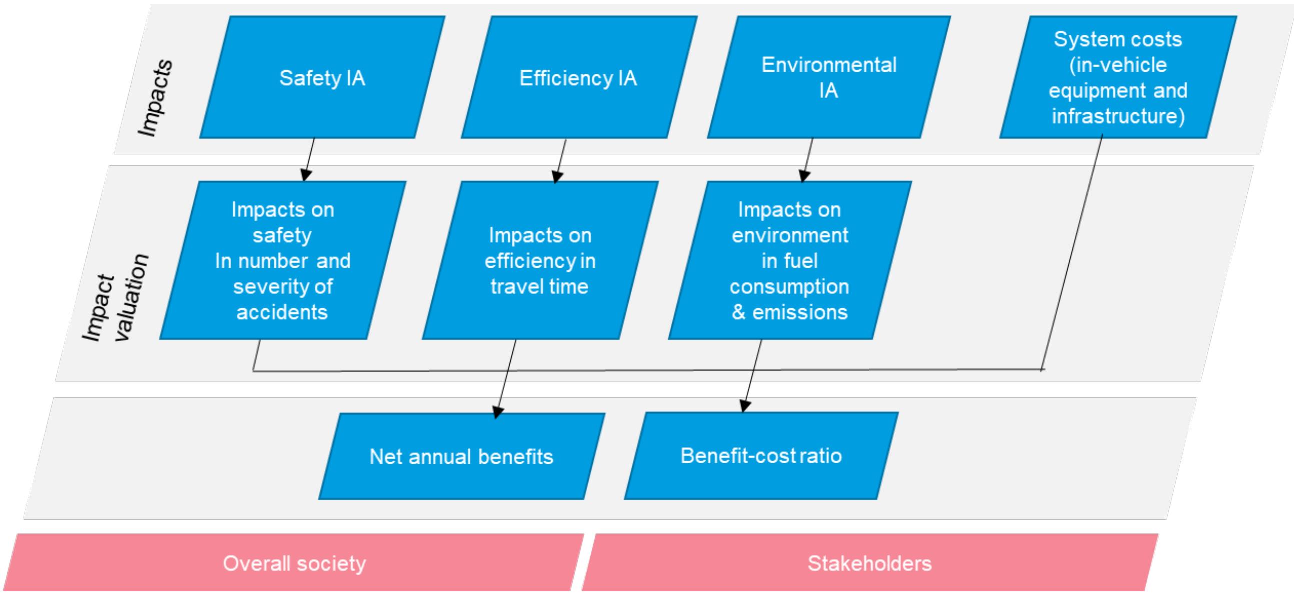


Figure 4.8: Socio-economic impact assessment.

System costs

System costs involve the costs of investing in ADF research and advanced engineering, installing ADF systems, but also their operating and maintenance costs. Basically, the ADF technology in question has already been developed, but some adjustments are still necessary.

In addition, system costs are related to road infrastructure costs and road infrastructure-related costs (i.e. operating and maintenance costs). However, the assessment on safety, efficiency, and the environmental impact assume no change in the existing infrastructure (i.e. the effects are estimated within the current ODD). Therefore, the CBA does not need to include other costs than those related to installing and maintaining in-vehicle ADF systems. However, the functionality of the ADF on the roads may depend on adjustments to the existing road infrastructure such as traffic signs and road markings. In that case, the CBA may also investigate the economic value of ADF impacts if the road infrastructure were improved to enlarge the ODD for ADF technology. Then the additional costs related to investments and maintenance of the road infrastructure would have to be included as well.

Safety impacts

Potential safety impacts of ADFs are related to changes in the number of accidents, injuries (minor and severe), and material damage to vehicles. A further impact of the ADF is the change in perception of safety due to the expected change in the number of accidents. Vehicles with ADF are expected to provide safer driving by reducing human error, and thereby reducing the frequency/likelihood and severity of accidents compared to conventional vehicles.

The parking ADF handles actual parking manoeuvres and the final stages of driving to the parking spaces at private homes, dedicated parking areas, or public parking. The potential benefits of automated parking are expected to arise mostly from increased safety, decreased driver stress, and increased comfort.

Efficiency impacts

Efficiency impacts of ADF consist of direct and indirect effects.

- Direct effects result from the impact on vehicle operations (e.g. car following, speed selection, etc.) and changes in traffic flow (e.g. changes in vehicle operations aggregated to traffic flow level).
- Indirect efficiency effects are expected to result from changes in the number of accidents and resulting delays, and from changes in travel behaviour.

Traffic efficiency impacts of ADFs can also be expected in terms of changes in travel time due to changes in traffic flow and congestion costs. For instance, a reduction in accidents may indirectly lead to reduced travel time due to less accident-induced congestion. However, if drivers without ADF tend to drive at speeds clearly higher than the permitted speed limits, travel times may increase, as ADFs do not drive faster than their programmed speed limit. Thus, whether the overall impact on travel time is positive or negative has yet to be determined. If it turns out that travel time is expected to increase because ADF ensures that driving takes place in accordance with the traffic rules, it is not obvious that this could be regarded as a cost to society.

Time spent while driving manually cannot be used to perform other activities. ADF can facilitate the driving task, which makes it possible to spend parts of travel engaging in other activities. Travel time has a positive monetary value if it is possible to work while travelling. This is called *productive value of travel time*. Travel time also has a positive monetary value if it becomes possible to relax in the car while travelling, which is referred to as *recreational value of travel time*. Performing other activities while driving can be included in the CBA by monetising the value of time per hour depending on the type of activity.

Environmental impacts

Environmental impacts of ADF are expected from changes in fuel consumption and CO_2 emissions. Potential benefits result in terms of reduced fuel consumption and CO_2 emissions. However, if drivers of vehicles with ADF tend to drive longer distances, or if other travellers and non-travellers find it attractive to buy automated vehicles, this could lead to increased vehicle kilometres travelled and thereby increased emissions.

Summary of impacts

Summarised, we will assess the impacts of ADF on:

- Traffic safety in terms of the number of accidents and their severity.
- Traffic efficiency due to changes in traffic flow and congestion.
- Environment due to changes in fuel consumption and CO_2 emissions.

In the socio-economic impact assessment, the value of benefits has to be compared to system costs to equip vehicles with the Level 3 ADF technology and to ensure that the technology will work. If the benefits outweigh the system costs, the technology is considered beneficial to the society.

Monetary valuation of impacts

The socio-economic impact assessment will include all perceivable costs and benefits related to safety, efficiency, and environmental effects of the new technology. This is done by assigning monetary values to these effects. We acknowledge that not all relevant costs and benefits are reflected in market prices, such as the monetary value of avoided fatalities, CO_2 emission costs, and so on. In order to calculate the overall economic values of the ADF impacts, the estimated changes in individual fatalities and injuries, travel time, fuel consumption, and CO_2 emissions are to be multiplied with generally accepted cost unit rates.

Safety impacts of ADFs are related to changes in accidents. The economic value of changes in accidents can be derived from costs of personal injuries and property damage. Costs of personal injuries include healthcare costs for treating the injured and administrative costs related to injuries such as rescue services, police reports, and even judiciary costs. Productivity loss because of injuries and reduction in quality of life (either temporarily or permanently) are also injury-related costs. For the property damage, costs to transport authorities (police, rescue teams, etc.) come in addition to repair costs or vehicle replacement costs. However, obtaining costs for all of these areas is practically not possible. Therefore, in practice, only the major components of costs are usually included in the CBA (e.g. healthcare costs and costs of property damage).

Further, the socio-economic impact of changes in fatality rates due to changes in accidents can also be obtained by considering the value of human life. Value of statistical life is a common measure for the valuation of a fatality avoided (Ashenfelter, 2006). This value is based on estimating how much people are willing to pay to reduce the risk of death. Value of statistical life varies across countries, as countries differ with respect to gross domestic product per capita. The literature provides some statistics for values for some EU countries (e.g. Bickel, Friedrich, Burgess, Fagiani, Hunt, De Jong, et al., 2006).

Traffic efficiency impacts of ADFs are changes in travel time due to changes in traffic flow and the amount of congestion. Travel time costs refer to the amount of time spent on the road and its best alternative use. For example, any reduced travel time can be spent on performing activities such as work/business, personal, recreational, or leisure activities. Travel time is usually valued as percentage of average wage based on the estimates of what people would be willing to pay to reduce travel time (Henscher, 1977). The monetary values of time savings vary depending on the type of activity. For instance, time saving spent on business has a higher monetary value than time savings spent on personal travel (Wardman, Batley, Laird, Mackie, Fowkes, Lyons, et al., 2013). To give another example, the unit prices for travel time and delays are higher for freight transport than for passenger transport. The monetary value of time savings in uncongested and congested traffic may also differ. The different unit values for time are reported for some countries in the literature. Environmental impacts are associated with changes in fuel consumption and hence changes in CO_2 emissions, which also have monetary values.

4.7.3 Data sources

The main task of the CBA is to find out whether the ADF is beneficial from the society's point of view. To that end, data are needed for the baseline and treatment scenarios. The calculations will be done for each ADF (motorway, traffic jam, urban, and parking) within the current road infrastructure, based on different (exogenous) penetration rates of the activated ADF (usage rate). Input data requirements may vary between the different ADFs. The net benefits of enlarging the ODD for ADF technology will also be attempted to be calculated, which means that also additional costs related to investments and maintenance of road infrastructure will have to be included.

The data required for the CBA will originate from the assessments on traffic safety, efficiency, and environmental impacts, scaled up to EU level. Also, some of the results from pilot site questionnaires and international survey will be used in the analysis, such as answers to questions on willingness to use and willingness to pay. Finally, other data sources such as official statistics are required to provide further data for the analysis, in particular concerning standard unit costs.

The analyses can be carried out at different levels of aggregation. At the top is the EU, treated as one entity. As standard unit costs are not well established at this level, it might be interesting to supplement these analyses with more in-depth studies of one or two countries with good data coverage and perhaps also for representative travellers within specific traffic situations. Accordingly, input data are needed for different levels of aggregation, which also should guide the up-scaling of results from experiments and simulations.

Input from impact assessment

The impact assessment reveals the potential effects of ADF in terms of safety, efficiency, and environment. These data will be integrated into the CBA model.

The CBA analyses will use input at the EU level for each of the four ADFs and for in-depth analysis of specific countries, and for the representative traveller level in specific traffic situations. Countries and traffic situations will be specified later in collaboration with those simulating the impacts on accidents, efficiency, and environment. It means that information is needed on how baseline variables should be expected to change if different fractions of vehicles had been operated with the tested ADFs.

Specific input data from the different areas of assessments, which describe the expected difference between baseline and treatment conditions, are crucial for the assessment of socio-economic impacts:

- Safety impact assessment will provide information consistent with the snapshot approach on the frequency and severity of traffic accidents, measured as the change in the number of accidents with fatal outcome, with major and minor injuries of persons, and with property damage only. This information is needed for the baseline and treatment scenarios covering the whole EU and for a couple of countries with generally good data coverage.

- Efficiency impact assessment will provide information on impacts on travel time for motorways and urban roads, which directly follow from the use of ADF equipment, and which indirectly are expected to occur because of changes in the numbers and types of traffic accidents. This information is needed for the whole EU, for a couple of countries with generally good data coverage, and is measured as average travel time per standard distance for the different types of travelling.
- Environmental impact assessment will provide information on fuel consumption for different types of driving, as for the input from the efficiency assessment. These data should also cover the whole EU, as well as the countries selected for in-depth studies.
- In mobility impact assessment, information on additional effects on driving and driving behaviour will be derived and included in qualitative discussions regarding socio-economic impacts of ADF.
- Expected effects of ADF on comfort and performing secondary tasks will also be included in the qualitative discussions of socio-economic impacts.
- Efficiency and safety impact assessments will also provide information on uncertainty related to travel time (reliability, punctuality) and user & acceptance evaluation for perceived risk of accidents.

Input for costs

As mentioned above, the impacts of ADF are derived from the changes in accidents, travel time, vehicle operating costs (fuel consumption), and CO_2 emissions. Given the results from the impact assessment, a major task in the CBA analysis is to assign monetary values to the most important impacts reported in some technical units. CBA should, however, include all safety, efficiency, and environmental benefits (impacts) and costs of ADF, both in monetary terms and those which will be addressed qualitatively.

Monetary costs of accidents involve property/material costs and personal costs such as hospital costs due to accident injuries. The cost data for accidents should be based on available hospital treatment costs and insurance claims for property damage. Non-monetary costs of safety impacts are related to personal injuries, both fatal and injuries.

Changes in travel time also have a monetary value. The monetary value of time spent on productive or recreational activities, or depending on the trip purpose (work or leisure), can be found for some countries in the literature.

There are some studies in the literature which provide estimates of the different unit values for some European countries (e.g. Bickel et al., 2006). These values can be applied in CBA for calculating unit values at EU level. Input is needed at EU level (averages), country level, and representative traveller level for the traffic situations to be considered. The following provides an overview of the input needed:

- Safety: Unit accident cost for different types of accidents (same classification as above).
- Travel time: Value of time per unit (same classification as above).
- Emissions: Cost of CO_2 emissions per unit and fuel costs per unit (e.g. litre per 100 km driven).
- Uncertainty: Willingness to pay for reduced uncertainty (travel time and accidents).
- System costs: These costs consist of costs of equipping the vehicle with ADF, operating costs, and maintenance costs of the ADF. Benefits and costs of ADF should be calculated so that they cover the same time period when it comes to system costs. This means that the cost of equipping the vehicles with ADF should be recalculated to an annual cost by taking into account the number of years that the equipment is expected to last. This will also be the case when considering additional infrastructure costs of expanding the ODD for the different types of ADF equipment in question.
- Other potential cost items: Insurance premiums with and without ADF, cost to automobile repair shops, and user charges if applied (subsidies for ADF-equipped vehicles for tolls/road charges).

Input from the L3Pilot pilot site questionnaire and survey

In L3Pilot, two main types of surveys are conducted. One is the pilot site questionnaire addressed to the users of ADF-equipped vehicles. The other is the international survey. The pilot site questionnaire is conducted once per participant and is to be answered by participating drivers (users) and accompanying passengers. The international survey is to be conducted among the general public. Both the pilot site questionnaire and international survey contain questions regarding travel behaviour and willingness to use the ADF, among others. They also include questions regarding willingness to pay for each ADF. In L3Pilot, the questions on willingness to pay aim at capturing the individuals' demand for equipping their vehicles with ADF based on their preferences (i.e. according to what they are willing to pay for each ADF). In CBA, the benefits and costs of ADF technology can be assessed without asking potential users and others about their perceptions of this technology. However, a willingness to pay study can provide supplementary information for the socio-economic impact assessment by revealing individuals' preferences for ADFs.

4.7.4 Remarks on chosen approach

Chapter 4.7 provides the evaluation plan for the socio-economic impact assessment of ADFs related to traffic safety, efficiency, and environment. We concluded that the traditional and widely used CBA approach is not applicable in L3Pilot and argued that it would be better to choose a snapshot approach to quantifying the socio-economic impacts. The socio-economic impact assessment will also include qualitative considerations of unquantifiable impacts, also when interpreting the empirical findings in the snapshot approach.

The main focus in the socio-economic impact assessment of ADFs is on the net welfare effects for the society at EU level, but country-specific or representative individuals' welfare may also be relevant in a CBA. Data input for the analysis is discussed in this chapter. The major part of input data comes from impact assessments on traffic safety, efficiency, and environment, along with some subjective data from surveys. Monetary valuation of all costs and benefits of ADFs are further discussed. Since data may not be available for all EU countries, it is necessary to find appropriate methods for aggregating monetary values from national levels to the EU level.

Accurate identification of all costs and benefits of ADF to the society is challenging. Uncertainty related to the identification of model parameters in the CBA may lead to overestimation or underestimation of the net welfare impacts for the society or its stakeholders. Conducting sensitivity analyses is essential for taking account of changes in model assumptions and uncertainties related to different parameters (e.g. uncertainty about unit costs). In this way, sensitivity analyses can contribute to improving the robustness of the main analysis.

5 Summary of SP3 recommendations for Code of Practice

This chapter gives an overview of contributions to the Code of Practice (CoP) from the viewpoint of methodology. The draft and results of the pilot application of the draft CoP can be found in L3Pilot Deliverable D2.2 Draft and results from Pilot application of draft CoP by Fahrenkrog, Schneider, Naujoks, Tango, Knapp, Wolter, et al. (2020).

One of the objectives of L3Pilot is to generate a Code of Practice for AD (CoP-AD), which is to provide a comprehensive guideline with best practices for the development of ADFs at automation levels 3 and 4. The guideline is being developed by the Code-of-Practice subproject. The methodology subproject supported the development of the guideline by specifying requirements and providing input regarding testing, evaluation and impact assessment from a methodology point of view. This chapter gives an overview on how the methodology subproject contributed to the development of the CoP-AD.

5.1 Procedure to support development of a Code of Practice for AD

First, a scope and structure of the new CoP-AD was defined by the Code-of-Practice subproject (SP2). The methodology subproject (SP3) participated in several workshops contributing to the refinement of the CoP-AD structure. This structure served as a basis for the iterative process, which was established by the methodology subproject to provide input to the content of the CoP-AD (see Figure 5.1).

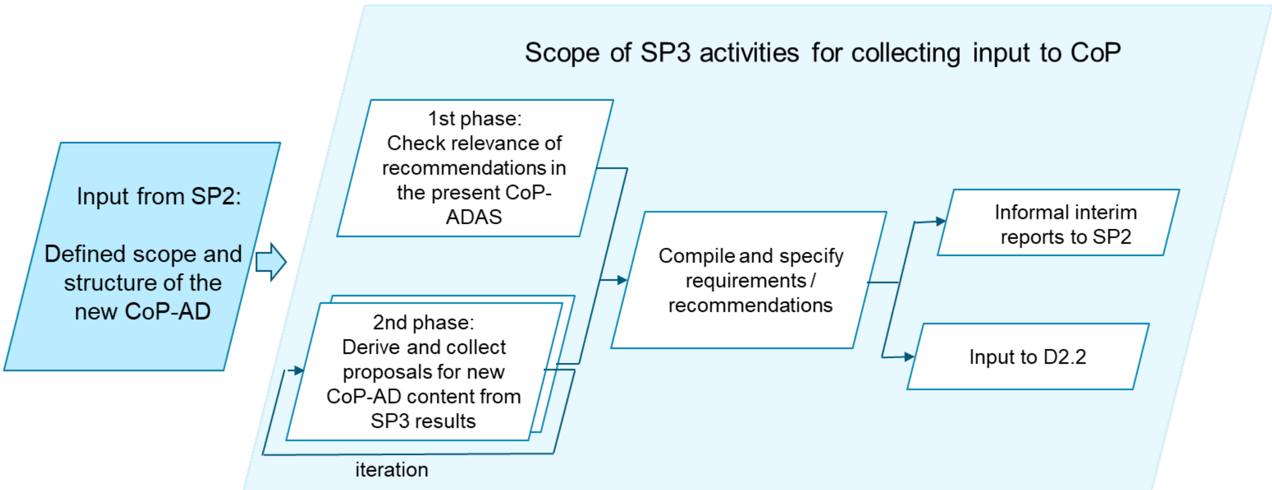


Figure 5.1: Process used by the methodology subproject (SP3) to support development of the Code of Practice for AD (CoP-AD).

In a first phase, recommendations covered within the already existing Code of Practice for ADAS (CoP-ADAS) were checked for their relevance and applicability to ADFs. The synopsis of the results was forwarded to the Code-of-Practice subproject as an informal report.

The second phase aimed at collecting proposals for piloting, testing, and evaluating ADFs. The proposed methods were derived from the ongoing work of the methodology work packages. As the work of the methodology subproject had to be performed concurrently with the development of the CoP-AD, input was provided step by step. In the first step of this iterative process, rough recommendations from early results of the methodology work packages were provided. In the second step, further recommendations according to the progress of the work were collected and selected after discussion with the Code-of-Practice subproject. Finally, detailed recommendations on a selected topic (i.e. application of study approaches) were compiled and provided as an additional input to the CoP-AD in the beginning of 2019.

5.2 Recommendations resulting from a CoP-ADAS review

The first recommendations by SP3 reported to the Code-of-practice subproject resulted from the review of the present CoP-ADAS. Relevant chapters in the CoP for ADAS were identified and checked for whether their content can be considered useful also for testing and evaluation of ADFs.

- *Terms and definitions*

  It is recommended by SP3 that the new CoP-AD provides an overview on the defined levels of automation and the characteristics of different types of ADF.

  Relevant terms and abbreviations related to methods and HMI aspects may be defined. Definitions may be taken from the glossary of the methodology subproject and a recently published ISO technical report that introduces basic concepts related to driver performance and state in the context of AD (ISO, 2019). The concepts are applicable to human factors evaluations using driving simulators, tests on restricted test areas, or tests on public roads.

  With regard to AD, adaptations of the terms and methods used for impact assessment are recommended by SP3. Details may be derived from expected results of the methodology subproject and could be included in the annex of the new CoP-AD.

- *Development process*

  From the SP3 perspective, an overview of the development process of ADF would be useful for structuring the CoP-AD.

  An early chapter of the new CoP-AD may give an overview on HMI issues related to AD. Such an overview would make it easier for the function developer to identify where action is needed in the design and evaluation phase of the development process. HMI issues to be included in the description may be willingness to use, driver status with regard to vigilance and fatigue, mode awareness and situation awareness, transitions of control, failure of use, misuse, communication between automated vehicles and non-automated road users, etc.

  However, in contrast to CoP-ADAS, the demands of AD place new priorities on HMI concepts. Therefore, items have to be revised regarding these new challenges. For example, considerations regarding the driver’s situation awareness and perception should take into account that the driver is allowed to perform non-driving related tasks when driving in automated mode.

From the SP3 perspective, impact assessment should be given more importance in the development process than it carried in the present CoP-ADAS. The CoP-ADAS links the necessity of performing an impact assessment to changes in the considered ADAS function. However, in L3Pilot, impact assessment is also regarded as being important in the early phases of the development process and before market introduction of new ADFs. It is recommended to show what kind of impact assessment is applicable in which phase of the development process.

- *Controllability*

Aspects of controllability cover a major part of the CoP-ADAS and may also be addressed in the CoP-AD. The definition of controllability has to be adjusted to meet the demands of AD. As the operational control of the driving task switches between the ADF and the human driver, requirements regarding controllability are more complex for the different kinds of use of the ADF than for use of an ADAS. HMI aspects may affect controllability of the ADF in different ways depending on the kind of use.

In higher automation levels (SAE Level 3 and above), the driving task is completely performed by the ADF during periods when the function is active. However, HMI-related issues of ADFs used within system limits do not only concern manually switching on/off the ADF. Despite the driver being temporarily out of the control loop, controllability during ADF use depends both on functional as well HMI issues (e.g. driver's mental model of ADF's behaviour and driver status). Furthermore, it is important, from an SP3 perspective, to keep in mind that HMI issues may carry over on driver performance during manual driving (e.g. mode awareness and vigilance).

The ADF is not able to continue performing the driving task when system limits are reached. The ADF will request the human driver to take over control. Thus, the re-establishment of the driver's control of the vehicle and driving task is one of the key requirements regarding controllability at and during exceeding system limits from the SP3 perspective.

During and after system failures, the technical aspects of functional safety (e.g. reliability of the ADF, single safety-related components, ability of the ADF to monitor its own safety status and that of other safety-related components) as well as HMI aspects (e.g. driver's awareness of the necessity to intervene in case of silent failures) have an impact on controllability.

- *Application concept: Main part and corresponding annexes*

A structure of the CoP-ADAS – (a) main part giving a quick overview on recommendations, (b) referenced annexes providing context-specific details – appears to be useful for the CoP-AD.

Despite the fact that some of the CoP-ADAS content is also relevant for the CoP-AD, it is recommended by SP3 to arrange the CoP-AD as a standalone document where all needed information on AD is available, even if this means that some content of the Cop-AD and CoP-ADAS may overlap. Such a CoP-AD is easier to handle than a document needing many links to the CoP-ADAS in order to be fully comprehensive.

5.3 Recommendations collected from the methodology work

The following list gives an overview of proposed topics for CoP-AD input that were collected by the methodology work packages when developing the methodology for L3Pilot. The proposed topics were derived from preliminary results, allocated to the draft dimensions of the new CoP-AD structure provided by the CoP subproject, then forwarded as an intermediate input to the CoP-AD work. Besides other aspects of the methodology, the proposed topics highlight the results relevant for setting the experimental framework in terms of research questions and appropriate measures, experimental procedures and how to perform data collection, and evaluation methods for analysing data and estimating the potential impacts of ADFs.

- Research questions and appropriate indicators and measures, which are defined for technical and traffic evaluation, user & acceptance evaluation, impact assessment and socio-economic evaluation.
- Overview of approaches for data collection, which can be used for comparing different types of experimental studies, survey methods, and analytic computer simulations with regard to how they can be used for different research questions and measures.
- Criteria of how to appoint specific groups of test participants (ordinary drivers, professional drivers) to tests in different phases of the development process and for different kinds of research questions regarding human-machine interaction.
- Definition of study design and baseline in order to enable researchers to conduct pre-post- / baseline-treatment comparisons (with/without ADF) when analysing data and estimating impacts; recommendations also include considerations, if baseline data cannot be collected during internal tests drives (i.e. how to make external baseline data available).
- Recording of all relevant aspects of the experimental environment (road type, traffic characteristics, weather conditions, geographical location etc.) in which the data were collected. This is especially important for evaluations based on data collected during on-road tests in real traffic, as environmental conditions vary a lot and invalid conclusions may be drawn in case of disregarded environmental conditions.
- Methods for user & acceptance evaluation, which focus on data to be collected via questionnaires, video or vehicle-based sensors in order to answer user related research questions and test hypotheses; recommendations also include concepts on how to connect results of an on-road pilot/FOT with results from supplementary studies (e.g. driving simulator study, Wizard-of-Oz study) and surveys in general public.
- Principles for scaling up, which allow for incorporating test drive results into the scaling-up process and support the estimation of safety impacts of ADFs at EU level.
- Scenario-based assessment of AD, which aims at controlling the variation of moderating factors in on-road tests and proposes to use both driving scenarios and traffic scenarios for impact assessment.

- Common methodology for impact assessment, which provides recommendations on how to form a coherent comprehensive methodology by combining different assessment methods (e.g. simulation-based assessment of impacts on safety, traffic efficiency and environment).

After discussion of the above-mentioned topics with the CoP subproject, the proposed overview of approaches for data collection was selected for further particularisation by the methodology subproject. The detailed overview describes the pros and cons of potential approaches and how the approaches fit different categories of research questions related to AD. It consists of two tables: one compares approaches which are applicable for objective data collection (e.g. driving simulator, test track, Wizard-of-Oz, micro-simulation of traffic). The other compares approaches suited to subjective data collection (e.g. observation, focus group, open-ended / closed-ended interview questions). The tables are available as an annex to the draft of CoP-AD (see L3Pilot Deliverable D2.2 by Fahrenkrog et al., 2020). Details on the methodological background of the approaches can be found in L3Pilot Deliverable D3.2 (Penttinen et al., 2019).

6 Lessons learned

This chapter is modified from the conference papers published in ITS World Congress 2019 by Innamaa, Merat, Louw, Metz, Streubel & Rösener, 2019 and Transport Research Arena 2020 (Conference cancelled) by Louw, Merat, Metz, Wörle, Torrao, Innamaa, 2020.

6.1 General remarks

When designing the experimental procedures for an on-road pilot study, one must understand the difference between a FOT of (nearly) market-ready products and the piloting of systems that are at an earlier technology readiness level, such as those under investigation in the L3Pilot project. Therefore, for safety and ethical reasons, for this type of AD pilot it is only realistic to conduct tests first with designated and trained safety drivers, rather than with members of the public. This procedure is very different to FOTs where ordinary drivers use the system as part of their daily lives. Thus, such a pilot study produces partly indicative estimates of impacts, and one must make assumptions about the eventual use of market-ready versions, whereas a FOT provides more direct proof of impacts from the field measurements directly.

The goal of L3Pilot is to demonstrate and assess the functionality and operation of Level-3 ADFs of passenger cars in real or close-to-real use contexts and environments. The project provides a great opportunity for large-scale on-road testing of automation, which is not yet available on the mass market. The engagement of a large number of different OEMs, and the implementation of various ADFs in different environments and different parts of Europe, enable a broader view of the potential impacts of automation than an evaluation based on a single trial. However, the pilot nature of these tests will bring some practical limitations to the possibilities regarding use and conclusions related to real world implementations or expected impact. To generate valid results on the impacts of the ADFs, the principles used to collect the evaluation data and any ensuing conclusions need to be considered carefully.

6.2 Set-up of the experiments

The field experiments are designed to provide the data needed to answer the research questions set for the project. A list of relevant research questions is defined for a pilot project, just as they would be formulated for a FOT. They are shaped around the focus of the project, description of the tested systems and theories of the related impact areas. However, the prioritisation and selection of research questions cannot be based on these factors alone but must also consider the practical possibilities of each pilot site and vehicles used for each study. In L3Pilot, the feasibility of research questions was also checked against the possibilities for data provision (sensors, logging, features of test rides, experimental design), as well as the role and type of participants (drivers).

Since ADFs are still prototypes, special measures regarding safety are necessary for open-road pilot tests. This is why, for example, it is not possible to study the interactions and behaviour of ordinary drivers using the vehicles during their daily routines. Instead, testing requires the inclusion of safety drivers and additional observers to monitor system and driver performance. However, to ensure a more representative sample of drivers and to overcome this limitation, L3Pilot uses a number of special safety concepts, including:

- Equipping some vehicles with driving-school-style pedals, allowing interjection by a trained observer in the passenger seat, where necessary.
- Placing an ordinary driver in the passenger seat to observe the safety driver's interactions with the system and vehicle, providing an impression of the system.
- Following the drives with the safety driver, some pilot studies consider progressing to using ordinary drivers, with increasing maturity of ADFs.

6.3 Methodology for assessing the impacts on driving behaviour

When assessing the impact of ADFs on driving behaviour or dynamics (e.g. observed differences between the ADF and human driver regarding car-following behaviour, or speed and headway distributions, see our plan for technical & traffic evaluation in Chapter 4.2), it is important to consider the maturity of the system, and whether it offers a representative and realistic driving scenario. Also, in order to obtain permission for testing ADF on an open road, public authorities require that it is safe to do so. The evaluation of the details of driving dynamics of automated vehicles can also be sensitive to car manufacturers. Therefore, it is also important to ensure that vehicle telemetry data is kept confidential; also, any information that is shared within the project or with the general public cannot allow manufacturers to be ranked or compared, compromising the competitiveness of OEMs at this crucial development stage. A further challenge for developing a broad view of driving behaviour impacts of ADF is that each result may only be limited to certain test routes, within certain speed ranges, and in certain weather conditions, as specified by the ADF's ODD. Such pilot studies are also obliged to adhere to the OEM rules and national regulations drawn up for testing of AD on open roads and in controlled environments, such as the mandatory use of safety drivers in some situations. Thus, the above specifics limit the possibility of addressing all of the research questions being defined based on scientific interest.

To address these limitations, the following solutions have been drawn up by L3Pilot:

1. To ensure that data collected at the different pilot sites can be analysed with a common process, the project partners have agreed on a common data format (Hiller et al. 2019) and produced a common methodology and analysis toolkit, to be used by different analysts, at different pilot sites, with data from different ADFs.
2. The most detailed data from each pilot site is only handled by one or two research partners, ensuring that the distribution of commercially sensitive data is controlled.
3. A sophisticated data-sharing process designed to pseudonymise the data at a level of aggregation suitable for analysis, coming from each pilot site, is used. All analysis done on a project level will be based on pseudonymised data (e.g. consisting of performance indicators for the different driving scenarios). To obtain these, the data from individual pilot sites are merged in the consolidated data base and analysed together.

4. The results from several pilot sites will be merged in such a way that sensitive information is protected. In practice, the public outcomes for a particular ADF can be presented only if it was piloted at more than one site. A fundamental requirement for the merging is that, in addition to protection of the privacy of manufacturers, the outcome is meaningful for those who utilise the results, either as such or in the following steps of evaluation. Thus, the results must give insights of the impacts of AD without compromising this privacy. The same principles are also applied for presenting the results of user experience and acceptance evaluation. This merging across different sites also ensures that the results of L3Pilot do not show the impact of single (OEM)-specific ADFs, but the averaged impact that can be expected if such systems are introduced to the road.

6.4 Methodology for evaluation of user experience and acceptance

6.4.1 Extrapolation of the user experience and acceptance to the real world

There are a number of aspects to consider when aiming to extrapolate the user experience and acceptance in an AD pilot to the real world. These can be divided into three main areas, namely, the maturity of the system, the test environment used, and the type of driver.

1. Drivers will be exposed to prototype HMI and AD control systems, which are still under development, potentially resulting in occasionally unpleasant driving and interaction experiences. This will affect the user experience and thus acceptance. For example, a development system that is prone to errors is likely to elicit different acceptance ratings compared to a market-ready system.
2. The nature of the pilot test means that participants will not be able to use the systems in their daily lives. Moreover, the test environments will likely range from free driving on a designated route to performing specific manoeuvres in a controlled environment. Therefore, the fidelity of the test environment may elicit different behaviours, uses, and perceptions than those in a daily life environment.
3. While ordinary drivers should be used as participants for user & acceptance evaluation in every instance possible, the prototypical nature of the systems means that, for safety and legal reasons, in many instances there will be the requisite of participants either being trained as safety drivers or being recruited from the OEM workforce. When using OEM employees who may not be familiar with the system, they are often required to be accompanied by safety drivers. Therefore, the perceptions and behaviour of the safety driver are likely to be influenced by their special training and their task to focus on ensuring the safety of the test, but also their presence in the vehicle may influence the behaviour and perceptions of an ordinary driver. Careful consideration must therefore be given regarding instructions to the safety driver, but ethical matters must also be well thought out (e.g. in terms of collision avoidance strategies).

The above factors limit the extent to which one can make firm and long-lasting conclusions about driver behaviour during interactions with, and perceptions of, ADF in the real world. However, while controlled test rides in prototypical systems with safety drivers on board are not the same as driving on your own during your daily commute, the behaviour and opinions of those who have participated in the test rides arguably have some value, and may be more valid than those who have had no such physical experience, or indeed than those of designers of the systems who are traditionally used for such evaluations. Therefore, it is possible to answer many of the user- and acceptance-related research questions, so long as caveats are included when results and recommendations are presented. However, the nature of the system, environment and driver type used in an on-road pilot study means that, ultimately, there are some aspects that cannot be addressed. Safety/professional drivers' perceptions of, and behaviour while using, ADF are likely to be influenced by their specialised training and more in-depth understanding of the systems' functioning. Therefore, responses from each participant type will be treated separately in the evaluation.

6.4.2 Variations in ADF maturity

The vehicles that ordinary drivers use during the pilots contain prototype HMI and AD control systems. These systems are still under development, and their maturity may inevitably vary between pilot sites, and potentially within pilot sites, should any updates be required during the prolonged testing schedule at some pilot sites. The use of "imperfect" prototypes and any unexpected behaviour of the systems may occasionally result in unpleasant driving or interaction experiences for users. These encounters may well affect user experience, and thus acceptance, since a development system that is prone to errors is likely to elicit different acceptance ratings compared to a market-ready system.

The likelihood of such shortcomings was taken into account in L3Pilot, when we defined the scope of the user & acceptance evaluation, and as a result there are little or no direct evaluations of the behaviour of the ADFs or their HMI within the project. Instead, in the pilots we sought to evaluate the indirect impacts of ADF use on, for example, acceptance of AD. While this does not remove the response bias altogether, it at least directs users' attention to aspects that are less related to the evaluation of the systems' HMI and/or behaviour. The system maturity will nevertheless need to be taken into account when drawing conclusions about users' views and behaviours while using these systems.

6.4.3 Variations between test environments

Notwithstanding the differences between ADFs across pilot sites, there is inherent variation between, and within, test environments. Test environments will vary in terms of their geography, infrastructure, drive lengths, test drive routes, and traffic conditions, and there will be seasonal, weather and lighting differences which may cause the ADFs to behave differently. From an experimental control point of view this is a concern, as variations between and within test environments may affect users' experience while using the ADFs.

The approach adopted within L3Pilot to deal with the above issues is two-fold. First, experimental guidelines have been developed to align experimental approaches across pilot sites, with an attempt to control for the study design, instructions for the selection of test participants, experimental protocol, and participant instructions and information. These are described in detail in

L3Pilot Deliverable D3.2 (see Penttinen et al., 2019). Second, where possible, information is collected about the variations between and within testing environments (i.e. confounding variables), which will be considered by the project-wide user & acceptance evaluation.

6.4.4 Type of driver

Careful consideration must be given to the instructions that the safety driver receives, in terms of how they interact with the participant. However, ethical matters must also be considered carefully, because while it may be preferable for the technical & traffic evaluation if the tested system was allowed to operate to the edge of its limits (or ODD), it is more important that safety drivers' actions prioritise the safety of the occupants inside the vehicle as well as that of other road users.

At some pilot sites, while ordinary drivers are not permitted to operate the ADF, they are included as passenger participants and driven around the experimental route by the safety driver. The views of passengers may be slightly different from those who are operating the vehicle; therefore, their responses will be analysed separately. Nevertheless, their views are still valuable by virtue of them having experienced the ADF in person.

The lack of regulatory alignment across the different European countries, such as differences in permission granted for the tests, affects the possibilities for achieving consistent results on user testing across pilot sites. For example, there may be variations across pilot sites in terms of the type of drivers who are permitted to be participants, which provides challenges for comparison across the project. Furthermore, pilot sites may vary in terms of participants' specific or unrestricted roles and permitted activities when operating prototype vehicles. For example, should drivers at a particular site not be permitted to engage in non-driving related tasks during the pilot, this may affect our ability to answer some questions at project level.

6.5 Standardised application of methods

When conducting studies across multiple sites, it is essential that any cross-pilot methods are administered using the same tools and protocols. For example, the primary data source for the user & acceptance evaluation at the pilot sites is the pilot site questionnaire. These are administered across all pilot sites, which vary in many respects, but most relevant here is the inter-experimenter variability. To minimise the effect of this variability on the quality of the data in L3Pilot, the questionnaire was implemented using the online tool LimeSurvey (www.limesurvey.org). The base format of the questionnaire was deployed to all pilot sites (as per the recommendation of Lai, Ströbitzer, De Goede, Krishnakumar, Val, Mahmod, et al., 2013), where the only task for pilot site staff is to transfer the translated versions of the questionnaire into LimeSurvey. This approach ensures that the questionnaire administration and output (i.e. coding of questionnaire items and answers) is consistent, not only between pilot sites, but also between experimenters. This approach also ensures that the data output can be integrated seamlessly into a common data format and transferred to a consortium-wide consolidated database, which can be used to analyse the combined results from all the pilot sites, per ADF and per participant type.

For the technical & traffic evaluation, the common definitions of performance indicators and driving scenarios are implemented in a common toolchain. The basis for this is the utilisation of the common data format for the hand-over of data from pilot sites to research partners (Hiller et al. 2019). Using a shared code repository accessible to all project partners, all sites for evaluation are using the same algorithms and parameters. Issues and updates are tracked and discussed among partners on a regular basis. This ensures that performance indicators to be merged across pilot sites are derived in the same way at any pilot site.

6.6 Methodology for assessing societal impacts

For FOTs, the impact analysts can assume that the field measurements represent true changes in driving behaviour. The implications of these changes on factors such as driving dynamics and travel behaviour, and their effect on traffic safety, transport network efficiency, environment and mobility could then be scaled up to make estimations about the entire traffic flow and vehicle fleet levels, and up further to understand expected impacts on national or EU levels. In AD pilots, a challenge (and source of uncertainty) is that there is one additional task in the assessment, involving the assessment of any differences between piloted versions of prototypes and their market-ready versions. These differences are relevant when assessing higher penetration rates for investigating societal impacts.

Another methodological challenge is the selection of a meaningful baseline condition for assessing the impacts of ADFs. While it would be ideal, it is unrealistic to expect a baseline that does not include any automation, as current vehicles already include low-level AD and ADAS, such as ACC and lane-keeping systems, with (low) penetration rates varying between regions and fleets. Having a mix of automation levels as baseline sets high requirements for the evaluation, as one should then understand the differences between all the levels of automation included, and the implications of their interaction. If a future dimension is added to the assessment, the influence of the other trends affecting mobility and transport (such as electrification, urbanisation, (other) new mobility services and concepts, etc.) should be included. Therefore, not only are the baseline datasets we have rather redundant, but, given the nature of the changing mobility ecosystem (e.g. electric cars, mobility as a service, 5G, shared AVs) along with automation, it is difficult to extrapolate impacts of ADFs too far into the future.

The evaluation plan in L3Pilot includes different best-practice solutions for different evaluation areas. For example, safety impact assessment is planned to be made based on different driving and traffic scenarios, assessing changes in severity and frequencies of accidents, and related impacts, using accident statistics for the scaling up. Other assessments will be based on generalised higher-level impacts, combined with national- and EU-level statistics on transport, especially for areas and networks where detailed data is not available. The impact analysis is not done for the tested prototype versions. Instead, (future) mature ADFs were defined, and the extent to which their expected impact can be directly based on the results from the field data analysis is thoroughly checked, establishing where and to what extent supplementing assumptions are needed. The descriptions of mature systems also set the evaluation on a more general level, not linking it with single solutions and locations.

7 Conclusion

The overall objective of the subproject on methodology (SP3) in L3Pilot was to develop a methodology for piloting, testing and evaluating Level-3 ADFs. This entailed reconsidering the theoretical background behind a multidisciplinary evaluation methodology, in order to gather understanding of the variety of possible impacts of ADFs on the transport system. In addition to expected positive impacts, it was important to be able to recognise also unintended, and possibly negative, impacts of ADFs. This Deliverable *D3.4 Evaluation plan* is the fourth and last deliverable of SP3 Methodology. It presents the final overall evaluation plan for all evaluation areas and research questions defined for L3Pilot, as well as lessons learned and good practices for evaluation of AD studies with real-world pilot data in general.

Although the overall methodology defined for FOTs (i.e. the FESTA V) has been developed for driver support systems, our efforts in the L3Pilot project show that the evaluation process can be adapted to suit the needs of AD pilot projects, as long as some caveats related to the pilot nature of AD studies are acknowledged. Also in AD pilot projects, it is essential to plan the testing and evaluation phases beforehand to ensure smooth operations and valid conclusions. Compared to ADAS FOTs, piloting of ADFs requires more flexibility and iteration between the different phases to deal with the ongoing development of ADFs as well as legal and other restrictions.

The methodology described in this deliverable relates to the full chain of assessment, starting from technical assessment of ADF and user evaluation, to adaptation of modelling approaches for impact assessment and scaling up of the impacts to EU level, and ending with choosing the best set-up for socio-economic evaluation of the expected impact. The work started by setting the research questions. They were organised at three levels of detail, with 11 main questions (Level-1) on four evaluation areas: technical & traffic evaluation, user & acceptance evaluation, impact assessment, and socio-economic impact assessment. More detailed questions were developed related to specific components of the questions where appropriate. In all, there were 37 Level-2 and 68 detailed Level-3 research questions. An overview of the interrelations between these evaluation topics was defined, and an evaluation plan including methods and data sources was specified for all the research questions.

Experimental procedures were developed for the pilots to provide a solid base for the evaluation methodology, and to ensure that the results from tests across all pilot sites could lead to an L3Pilot-wide evaluation, taking into account the practical limitations of their implementation. Furthermore, the aim was to harmonise the evaluation criteria by providing detailed recommendations for the pilots with the intention of creating holistic evaluation results of the L3Pilot project. In all, the pilots appeared to confirm that the given recommendations and guidelines were good goals, even if some of them were not (yet) fully applicable to all pilot sites. The most challenging recommendation for practical implementation in the pilots concerned the test participants. Everyone agreed on the preference of using ordinary drivers from a scientific perspective, but since the ADFs are still prototypes and their safety must be ensured, for most pilot sites the safety concept foresees the use of professional safety drivers. In this case, it is acknowledged that the demographic factors of this group do not necessarily reflect the driver population of the future customer or user base. It is highly appreciated that some pilot sites were allowed to recruit also ordinary drivers as test participants. Another methodological challenge concerns the representativeness of the pilot tests with respect to the traffic environment. As there are only a small number of urban pilots in L3Pilot, and urban environments are so diverse, it is also acknowledged that the representativeness of urban ADF results may be limited. For the motorway environment, with more pilots and less variation in the road layout, the representativeness of the results is expected to be better.

The work on setting the methodology used the best practices of previous projects on ADAS by further developing and adapting them to SAE Level 3 automation. For example, due to the nature of ADFs, the prototype systems tested in the project, and the need to ensure the safety of the driver and all surrounding traffic throughout the tests, adaptation and further development of existing methods was needed and implemented. This deliverable presented the final overall evaluation plan for all evaluation areas, i.e. details on how research questions covered by the project will be addressed, including data sources and methods with possible updates to the topics presented in the previous three deliverables (L3Pilot Deliverables D3.1–D3.3).

Technical & traffic evaluation and user & acceptance evaluation is based on the 20 different ADFs used in the field tests at the 16 pilot sites for driving on a motorway, in an urban environment or for parking. To be able to analyse the overall impact of these functions, that is, independent of the individual implementations of specific vehicle brands, the ADFs will be grouped in the analysis and indicators will be presented such that they are merged across pilot sites. Impact assessment and socio-economic assessment do not address the tested prototype implementations as they target the potential impacts of ADFs in perspective, when they are in use on a larger scale, and it is expected that the ADFs will be developed further from the ones tested in L3Pilot. Therefore, so-called mature functions were defined to represent such future ADFs.

The objective of the technical & traffic evaluation in L3Pilot is to answer research questions related to the technical performance and the effects resulting from an automated system's behaviour within traffic. Different effects are evaluated by comparing performance indicators. For user & acceptance evaluation, a multifaceted approach is used to form a holistic view of users' behaviours with, and acceptance of, the ADFs. The design of this approach was underpinned by a list of priority research questions. They address user acceptance and trust of the systems, willingness to use and pay for the functionalities, measures of driver state, user risk perception, driver engagement in non-driving related tasks, user behaviour during and after take-over situations, and user motion sickness.

Impact assessment addresses personal mobility, traffic safety, traffic efficiency and the environment. For mobility impact assessment, L3Pilot uses a multidisciplinary approach and framework to define the potential mobility impacts of ADF, and further to answer questions about potential impacts on actual travel exposure. In the mobility impact assessment, the focus is on the three points of view: amount of travel, travel patterns, and journey quality. Safety impact assessment quantifies possible effects of ADFs on traffic safety. In this context, it must be considered that each technology can affect traffic scenarios in three ways: scenarios in which a technology reduces risks, scenarios that are not directly affected by the technology, and scenarios in which potentially new risks occur. Typically, the occurrence of these types of risks are minimised in the development of ADFs. However, for a comprehensive assessment, it is necessary to analyse all potentially affected scenarios – positive as well as negative. Efficiency & environmental impact assessment examines the impact of ADF in terms of changes in macroscopic traffic flow (e.g. travel times, throughput) and greenhouse gas emissions (e.g. CO_2). Planned methods include e.g. simulations, accident, traffic, road environment and mobility data collection and analysis. The evaluation plan includes also the methods for scaling up of all impacts to EU level to ensure a versatile overview of the implications of SAE Level 3 automation in passenger cars.

The purpose of the socio-economic impact assessment is to clarify whether the benefits of equipping vehicles with ADFs outweigh their costs. The traditional and widely used CBA approach was not considered applicable in L3Pilot and, therefore, a snapshot approach was chosen for quantifying the socio-economic impacts. The main focus in the socio-economic impact assessment of ADFs is on the net welfare effects for the society at EU level, but country-specific or representative individuals' welfare may also be relevant in a CBA. As accurate identification of all costs and benefits of ADF to the society is challenging, uncertainty related to the identification of model parameters in the CBA may lead to overestimation or underestimation of the net welfare impacts for the society or its stakeholders. Therefore, sensitivity analyses are essential for taking account of changes in model assumptions and uncertainties related to different parameters. Overall, the pilot results should be taken as indicative, and the findings should be confirmed with a larger field operational test once the technology is more mature.

Finally, this deliverable reported lessons learned, and good practices found, for the evaluation methodology for AD studies with real-world pilot data and for supplementary studies. The deliverable also gave an overview of contributions to the Code of Practice (CoP) from the viewpoint of methodology. The methodology subproject supported the development of the guidelines by specifying requirements and providing input regarding testing, evaluation and impact assessment from a methodology point of view. An additional set of lessons learned were also documented for future automated driving pilots and evaluations.

In L3Pilot, the final details for the methodology will be set during the actual evaluation work in SP7 Evaluation, reflecting the data and results from the pilot experiments. This deliverable concludes the work of SP3 Methodology.

References

Anable, J., & Gatersleben, B. (2005). All work and no play? The role of instrumental and affective factors in work and leisure journeys by different travel modes. *Transportation Research Part A: Policy and Practice*, 39(2-3 SPEC. ISS.), 163–181. https://doi.org/10.1016/j.tra.2004.09.008

Ashenfelter, O. (2006). Measuring the Value of a Statistical Life: Problems and Prospects, The Economic Journal, 116: C10-C23

Beggiato, M., Hartwich, F., & Krems, J. (2019). Physiological correlates of discomfort in automated driving. *Transportation Research Part F: Traffic Psychology and Behaviour*, 66, 445-458.

Bickel, P., Friedrich, R., Burgess, A., Fagiani, P., Hunt, A., De Jong, G., Laird, J., Lieb, C., Lindberg, G., Mackie, P., Navrud, S., Odgaard, T., Ricci, A., Shires, J., & Tavasszy, L. (2006). HEATCO – Deliverable 5, Proposal for Harmonised Guidelines.

Borrack, M., Gwehenberger, J., Schatz, J., & Feig, P. (2017). Aktuelle Forschungsergebnisse zur Wirksamkeit von Fahrerassistenzsystemen mit zunehmendem Automatisierungsgrad. In: *VKU – Verkehrsunfall und Fahrzeugtechnik*, p. 378-385, November 2017.

Bruzelius, N.A. (1981). Microeconomic theory and generalized cost. *Transportation*, 1o: 233-245.

Carsten, O., Hibberd, D., Bärgman, J., Kovaceva, J., Pereira Cocron, M. S., Dotzauer, M., Utesch, F., Zhang, M., Stemmler, E., Guyonvarch, L., Sagberg, F., & Forcolin, F. (2017). UDRIVE Deliverable 43.1. Driver Distraction and Inattention, of the EU FP7 Project UDRIVE (www. udrive. eu).

Coe, R. (2002). It's the Effect Size, Stupid – What effect size is and why it is important. Paper presented at the *Annual Conference of the British Educational Research Association*, University of Exeter, England, 12-14 September 2002.

euroFOT (2012). *Overall Cost-Benefit Study, Deliverable D6.7*. Deliverable D6.7 of euroFOT project.

European Commission. (2019). CARE: Community Road Accident Database. In: Observatory. Available in: https://ec.europa.eu/transport/road_safety/specialist/observatory_en (Accessed 27th June, 2019).

Eurostat (2018a). *New registrations of passenger cars, motor coaches, buses and trolley buses, by type of vehicle and alternative motor energy*. [Datafile]. Retrieved from: http://appsso.eurostat.ec.europa.eu/nui/show.do?dataset=road_eqr_carmot&lang=en

Eurostat (2018b). *Passenger cars, by age*. [Datafile]. Retrieved from: http://appsso.eurostat.ec.europa.eu/nui/show.do?dataset=road_eqs_carage&lang=en

Fahrenkrog, F., Schneider, M., Naujoks, F., Tango, F., Knapp, A., Wolter, S., Cao, Y., Griffon, T., Demirtzis, E., Lorente Mallada, J., Caccia Dominioni, G., Fabello, S., Brunnegard, O., Kucewicz, A., Whitehouse, S., Hiller, J., Bonarens, F., Schindhelm, R., & Shi, E. (2020). *Deliverable D2.2 Draft and results from pilot application of draft CoP.* Deliverable 2.2 of L3Pilot project.

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

FOT-Net & CARTRE (2018). *FESTA Handbook. Version 7*. December 2018. Retrieved from https://connectedautomateddriving.eu/wp-content/uploads/2019/01/FESTA-Handbook-Version-7.pdf (accessed 30 April 2020)

Fridman, L., Langhans, P., Lee, J., & Reimer, B. (2016). Driver gaze region estimation without use of eye movement. *IEEE Intelligent Systems*, 31(3), 49-56.

Frost & Sullivan (2018). European and North American ADAS L1 and L2 Market, 2018. *Industry Research Analysis, K25D, Automotive & Transportation*, 28 Dec 2018

German Insurance Association (2017). *GDV - Report Automated Driving - Impact on claims payments up to 2035*.

Gold, C., Berisha, I., & Bengler, K. (2015, September). Utilization of drivetime–performing non-driving related tasks while driving highly automated. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 59, No. 1, pp. 1666-1670). Sage CA: Los Angeles, CA: SAGE Publications.

Griffon, T., Sauvaget, J., Geronimi, S., Bolovinou, A., & Brouwer, R. (2019). *Deliverable D4.1 Description of taxonomy of automated driving functions*. Deliverable D4.1 of L3Pilot project.

Gwehenberger, J., & Borrack, M. (2015). Influence of Driver Assistance Systems on Insurance Claims, In: *ATZ – Automobiltechnische Zeitschrift*, p. 60–65, October 2015.

Hensher, D.A. (1977). *Value of business travel time*. Permagon Press, London.

Hibberd, D., Louw, T., Aittoniemi, E., Brouwer, R., Dotzauer, M., Fahrenkrog, F., Innamaa, S., Kuisma, A., Merat, N., Metz, B., Neila, N., Penttinen, M., Puente Guillen, P., Rösener, C., Silla, A., Streubel, T., Tango, F., van den Boom, B., Weber, H., Woerle, J., & Zerbe, A. (2018). *Deliverable D3.1 From research questions to logging requirements*. Deliverable D3.1 of L3Pilot project.

Hiller J., Svanberg E., Koskinen, S., Bellotti F., & Osman, N. (2019) The L3Pilot Common Data Format – Enabling efficient automated driving data analysis, Paper presented at the *26th International Technical Conference on the Enhanced Safety of Vehicles*.

Innamaa, S., Axelson-Fisk, M., Borgarello, L., Brignolo, R., Guidotti, L., Martin Perez, O., Morris, A., Paglé, K., Rämä, A., Wallgren, P. & Will, D. (2013). Impacts on Mobility – Results and Implications. Large Scale Collaborative Project TeleFOT, 7th Framework Programme, INFSO-ICT 224067. European Commission 2013 No: Deliverable D4.4.3.

Innamaa, S., Merat, N., Louw, T., Metz, B., Streubel, T., & Rösener, C. (2019). *Methodological challenges related to real-world automated driving pilots*. ITS World Congress, 21-25 October 2019, Singapore. 10 p.

Innamaa, S., Merat, N., Louw, T., Torrao, G., & Aittoniemi, E. (2020). Applying the FESTA methodology to automated driving pilots. *8th Transport Research Arena* TRA 2020, April 27-30, 2020, Helsinki, Finland (Conference cancelled). 9 p.

Innamaa, S., Smith, S., Barnard, Y., Rainville, L., Rakoff, H., Horiguchi, R., & Gellerman, H. (2018). *Trilateral Impact Assessment Framework for Automation in Road Transportation: Version 2.0*. https://connectedautomateddriving.eu/wp-content/uploads/2018/03/Trilateral_IA_Framework_April2018.pdf

ISO (2019). Road Vehicles: Human Performance and State in the Context of Automated Driving. Part 1 – Common underlying concepts. ISO/TR 21959-1

Jonkers E., Wilmink I., Nellthorp J., Gühnemann A., & Olstam J. (2016). *Costs and benefits of green driving support systems*. EcoDriver deliverable D54.1.

Kuisma, S., Louw, T., & Torrao, G. (2019). Assessing mobility impacts of automated driving in L3Pilot. *26th ITS World Congress*, 3(October), 21–25.

Kyriakidis, M., Happee, R., & de Winter, J. C. (2015). Public opinion on automated driving: Results of an international questionnaire among 5000 respondents. *Transportation Research Part F: Traffic Psychology and Behaviour*, 32, 127-140.

Lai, F., Ströbitzer, E., De Goede, M., Krishnakumar, R., Val, C., Mahmod, M., Malasek, J., Martin, O., Welsh, R., & Carsten, O. (2013). *Operation Sites Description and Planning*. UDRIVE Deliverable 31.1. EU FP7 Project UDRIVE Consortium. https://doi.org/10.26323/UDRIVE_D31.1

Louw, T., Markkula, G., Boer, E., Madigan, R., Carsten, O., & Merat, N. (2017). Coming back into the loop: Drivers' perceptual-motor performance in critical events after automated driving. *Accident Analysis & Prevention*, 108, 9-18.

Louw, T., Merat, N., Metz, B., Wörle, J., Torrao, G., & Innamaa, S. (2020). Assessing user behaviour and acceptance in real-world automated driving: the L3Pilot project approach. *8th Transport Research Arena* TRA 2020, April 27-30, 2020, Helsinki, Finland (Conference cancelled). 10 p.

Madigan, R., Louw, T., & Merat, N. (2018). The effect of varying levels of vehicle automation on drivers' lane changing behaviour. *PloS one*, 13(2), e0192190

Metz, B., Rösener, C., Louw, T., Aittoniemi, E., Bjorvatn, A., Wörle, J., Weber, H., Torrao, G., Silla, A., Innamaa, S., Fahrenkrog, F., Heum, P., Pedersen, K., Merat, N., Nordhoff, S., Beuster, A., Dotzauer, M., & Streubel, T. (2019). *Deliverable D3.3 Evaluation methods*. Deliverable D3.3 of L3Pilot project.

Metz, B., Wörle, J., Zerbe, A., Schindhelm, R. & Bonarens, F. (in preparation). *Deliverable D7.2 L3/L4 long-term study about user experiences.* Deliverable D7.2 of L3Pilot Project.

Mok, B. K. J., Sirkin, D., Sibi, S., Miller, D. B., & Ju, W. (2015). Understanding driver-automated vehicle interactions through Wizard of Oz design improvisation. In Proceedings of the *Eighth International Driving Symposium on Human Factors in Driver Assessment, Training and Vehicle Design*, pp. 380-386.

Muir, B. M. (1994). Trust in automation: Part I. Theoretical issues in the study of trust and human intervention in automated systems. *Ergonomics*, 37(11), 1905-1922.

Nagy, B., Hiller, J., Svanberg, E., Kremer, M., Luxen, M., Christen, F., Koskinen, S., Bellotti, F., & Osman, N. (2019). Deliverable D5.1 Pilot tools and data. Deliverable D5.1 of L3Pilot project.

Naujoks, F., Wiedemann, K., Schömig, N., Jarosch, O., & Gold, C. (2018). Expert-based controllability assessment of control transitions from automated to manual driving. *MethodsX* 5: 579-592.

Neukum, A., Lübbeke, T., Krüger, H.-P., Mayser, C. & Steinle, J. (2008). ACC-Stop&Go: Fahrerverhalten an funktionalen Systemgrenzen. In M. Maurer & C. Stiller (Eds.), *5. Workshop Fahrerassistenzsysteme - FAS 2008*. pp. 141-150. Karlsruhe: fmrt.

Nilsson, G. (2004). Traffic Safety Dimensions and the Power Model to describe the effect of speed and safety. *Bulletin 221*. Department of Technology and Society. Lund University. Sweden.

Nordhoff, S., Beuster, A., Kessel, T., Bjorvatn, A., Ertl, D., Innamaa, S., Lehtonen, E., Malin, F., Louw, T., Merat, N., & Torrao, G. (in preparation). *Deliverable D7.1 Annual quantitative survey about user acceptance towards ADAS and vehicle automation*. Deliverable D7.1 of L3Pilot project.

Ostermaier I. (2018). *Analyse von Versicherungsschäden zur Ermittlung des Schadenvermeidungspotentials durch L3+ Funktionen*. TU München, Master's thesis.

Ostermaier I., Gwehenberger J., Feldhütter A., & Pschenitza M. (2019). *Analyse von Unfallschäden zur Ermittlung des Unfallvermeidungspotentials durch automatisierte Fahrfunktionen auf SAE-Level 3 und 4.* Zeitschrift Verkehrsunfall und Fahrzeugtechnik, Heft 4, S. 144-150, April 2019.

Öörni (2016). *iMobility Support D3.1b Report on the deployment status of iMobility Priority systems and update of iMobility Effects Database*.

Penttinen, M., Rämä, P., Dotzauer, D., Hibberd, D., Innamaa, S., Louw, T., Streubel, T., Metz, B., Wörle, W., Brouwer, R., Rösener, C. & Weber, H. (2019). *Deliverable D3.2 Experimental Procedure*. Deliverable D3.2 of L3Pilot project.

Price, J. H., Daek, J. A., Murnan, J., Dimmig, J., & Akpanudo, S. (2005). Power analysis in survey research: Importance and use for health educators. *American Journal of Health Education*, 36(4), 202-209.

Regan, M., Cunningham, M., Dixit, V., Horberry, T., Bender, A., Weeratunga, K., Cratchley, S., Dalwood, L., Muzorewa, D., & Hassan, A. (2017). Preliminary findings from the first Australian National Survey of Public Opinion about automated and driverless vehicles. http://advi.org.au/wp-content/uploads/2017/07/ADVI-Public-Opinion-Survey-7June2016.pdf

Rösener, C., Sauerbier, J., Zlocki, A., Eckstein, L., Hennecke, F., Kemper, D., & Oeser, M. (2019). *Potenzieller gesellschaftlicher Nutzen durch zunehmende Fahrzeugautomatisierung.* Bundesanstalt für Straßenwesen, Bundesanstalt für Straßenwesen

Schoettle, B., & Sivak, M. (2014). A survey of public opinion about autonomous and self- driving vehicles in the U.S., the U.K., and Australia. https://deepblue.lib.umich.edu/bitstream/handle/2027.42/108384/103024.pdf?sequence=1&isAllow ed=y

Schulze M., Mäkinen T., Kessel T., Metzner S., & Stoyanov, H. (2014). *DRIVE C2X Final Report (DRIVing implementation and Evaluation of C2X technology in Europe).* Deliverable 11.6. EC 7th Framework Program.

Swiss Re & HERE (2015). *The Future of Motor Insurance*. https://www.swissre.com/dam/jcr:1d18ad1f-bf6c-4d59-9fe1-a4dfcf977616/HERE-Swiss%20Re-white%20paper-final.pdf

van Nes, N., Bärgman, J., Christoph, M., & van Schagen, I. (2019). The potential of naturalistic driving for indepth understanding of driver behavior: UDRIVE results and beyond. *Safety Science*.

Venables, A.J. (2007). Evaluating Urban Transport Improvements – Cost-Benefit Analysis in the Presence of Agglomeration and Income Taxation. *Journal of Transport Economic and Policy*, 41(2): 173-188.

Venkatesh, V., Morris, M. G., Davis, G. B., & Davis, F. D. (2003). User acceptance of information technology: Toward a unified view. *MIS quarterly*, 425-478.

Wardman, M., Batley, R., Laird, J., Mackie, P., Fowkes, T., Lyons, G., Bates, J., & Eliasson, J. (2013). Valuation of Travel Time Savings for Business Travellers: Main Report, Institute for Transport Studies.

Zhang, T., Tao, D., Qu, X., Zhang, X., Lin, R., & Zhang, W. (2019). The roles of initial trust and perceived risk in public's acceptance of automated vehicles. Transportation Research Part C: Emerging Technologies, 98, 207–220.

List of abbreviations and acronyms

Abbreviation	Meaning
ACC	Adaptive cruise control
AD	Automated driving
ADAS	Advanced driver assistance system
ADF	Automated driving function
AEB	Autonomous emergency braking
AV	Automated vehicle
CBA	Cost-benefit analysis
CoP-AD	Code of practice for automated driving
EV	Electric vehicle
FCW	Forward collision warning
FOT	Field operational test
HD	High definition [map]
HMI	Human-Machine Interaction
L3	SAE Level 3 [of automation]
ODD	Operational design domain
OEM	Original equipment manufacturer
TOR	Take over request
THW	Time headway
TOC	Take-over controllability
TTC	Time-to-collision
VKT	Vehicle kilometres travelled
VRU	Vulnerable road user

Annex 1: User and acceptance evaluation questionnaires

The main source of subjective data from the pilot sites are user & acceptance questionnaires, and in particular ADF-specific questionnaires. The pilot site questionnaires were designed to gather subjective data from the pilot-site participants who will experience the different ADFs across pilot sites. Because there are differences between ADFs, three different pilot site questionnaires were designed with ADF-specific questions for parking, traffic jam/motorway and urban ADFs.

Aim of this guide

Pilot site questionnaires are the method decided by L3Pilot to gather information about test participants' views and acceptance of pilot-tested ADFs, guided by the L3Pilot research questions. This document has several uses. First, it provides information to support the selected partner responsible for supporting the vehicle owner to set up the implementation of the questionnaire, administer the questionnaires, and collect and store subjective data for the user and acceptance evaluation. Secondly, it aims to provide practical guidance to minimise the impacts of differences between pilot site studies, including how the questionnaires are administered. Thirdly, it aims to ensure that subjective data will be gathered using a common data format to ensure that a consistent approach can be taken in the user and acceptance evaluation.

More detailed and general documentation on user & acceptance evaluation methods can be found in L3Pilot Deliverable D3.3 (Metz et al. 2019).

L3Pilot user and acceptance subjective questionnaire workflow

The official L3Pilot questionnaires were published in English, which is the reference version. Selected partners or pilot-sites responsible can download the questionnaires from Confluence. Then, they can implement surveys in digital or paper format using the LimeSurvey tool, or other software can be used so long as the common data format is adhered to. The imported surveys may then be customised and translated versions can be implemented accordingly. Pilot sites are then able to export their new generated surveys and/or to export their results to CSV or SPSS format. Note: the SPSS output of the reference version of the questionnaire is in line with the common data format requirements for the consolidated database. Although selected partners/pilot sites responsible can create, edit or view a survey, it is extremely important that the questionnaire item codes are not changed, as this allows us to track responses from different pilot sites.

Pilot site questionnaire administration and implementation

The questionnaires are structured as follows:

- *Part 1: Before study/screening questions*, which covers test participants' socio-demographic information.
- *Part 2: Pilot site questionnaire (post drive)*, which covers test participants' impression of the system's performance, including acceptance, safety and comfort, among others.

- *Part 3: Willingness to pay*, which covers how much extra the participants would be willing to pay to have the function installed in their new vehicle.
- Lastly, there are some option questions on take-over requests that can be asked during the drive, which evaluate test participants' behaviour during take-over situations while using the traffic jam / motorway and urban ADFs.

There are specific recommendations that the selected partners/pilot sites responsible are requested to follow to ensure a consistent methodological approach during data collection. This section provides guidance and recommendations for each phase: pre-piloting for the questionnaire, and questionnaire implementation for before-drive and post-drive questions.

Test participant selection

Information on how to select test participants for filling in the pilot site questionnaire can be found in L3Pilot Deliverable D3.2 by Penttinen et al., 2019. Specifically, please see sections 2.2.5 Selection of driver sample and 4.2.1 Recommendations for pilot site test participants.

The participants must be informed that at no time will any information they provide be published that allows them as an individual to be identified. Moreover, all the responses will only be used for research purposes and their details will not be linked to their answers. In other words, all responses will be anonymised.

GUIDELINES FOR SECTION 2.1 Test participant selection	
1.A	All test participants should drive regularly (at least three times a week).
1.B	Demographic factors should reflect the driver population of interest, for example, the future customer population (depending on evaluation scope).
1.C	In addition to the above, there should be a balance between female and male participants. Where possible, all age groups should be represented, including young (<25) and old (60+).

Pre-test pilot site questionnaire administration

When planning the implementation of the pilot site questionnaire, please carefully consider the following recommendations and amend plans where and when required:

GUIDELINES FOR SECTION 2.2 Questionnaires pre-piloting and validation	
2.A	Ensure that the official and final version of the pilot site questionnaires is used (see links in Section 1.4).
2.B	If the questionnaire needs to be translated, please ensure that the translated version is verified by an independent reviewer (native speaker). This is to ensure that the translations have been done correctly.
2.C	While translating, do not change the questions or terminology. It is essential that the questions are presented in the same way at all pilot sites.
2.D	While partners are strongly encouraged to use the full version of the questionnaire, if any items are excluded, please list these on Confluence
2.E	If new questionnaire items are added, please add these to the end of the relevant section in the questionnaire and list the new additions on Confluence.
2.F	Please use the token code process developed by SP5 to manage the list of test participants who will fill in the pilot site questionnaires. When adding a participant to the list, fill in the participant's name, surname, email, etc. and token. In the token field, the 8-digit L3Pilot anonymized user ID should be inserted.
2.G	Ensure that information on the test participant type is collected, for example, whether they are a safety driver/professional driver, ordinary driver, or passenger.
2.H	Generate a test trial by gathering a small set of completed questionnaires data for pre-piloting of the analyses.

Before study / screening questions administration

Part 1 of the questionnaire covers the screening questions, which can be administered before the test participant experiences the drive(s) at the pilot site. Guidelines are presented next.

GUIDELINES FOR SECTION 2.3 Before study / screening questions administration	
3.A	Screening Questions (TJM.1-TJM.27, U.1-U.27, and P.1-P.24) can be completed in the recruitment phase.
3.B	The same questionnaire should be used irrespective of whether the participant is a driver or passenger of the test vehicle.
3.C	For each test participant, screening questions should only be administrated once.
3.D	You should record and save responses using the anonymised ID token.

Post-drive questions administration

The number of times that drivers complete an experimental drive, and therefore the questionnaire, will vary across pilot sites and will depend on several factors. The intermittent administration of the user post-drive questions must be agreed between the OEM and selected partner. For example, the pilot responsible may be interested in administering the post-drive questions more than once. In that case, the selected partner could perform some longitudinal analysis. However, the pilot sites have different testing schedules and different frequency of drives, which results in different amounts of exposure to the ADF. Therefore, when it comes to gathering the data across sites for the L3Pilot consolidated database and user and acceptance evaluation, it is fundamental to ensure a standard practice across all sites. The guidance is that, at the very least, the post-drive questions should be administered to the participants after their very final exposure of the ADF. Therefore, if the same test participant has completed several drives, the responses to the post-drive questions should at least be recorded after the final exposure to the ADF. The same applies to the willingness to pay questions.

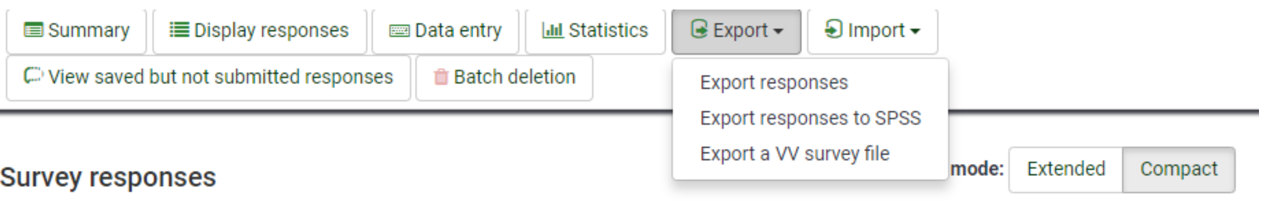
It is difficult to advise on the maximum number of times participants should be asked to complete the post-drive questionnaire. However, they should be separated by some substantial and quantifiable change in either 1) the participant's experience with the ADF, 2) the time since last using the ADF, or 3) the behaviour or ability of the ADF. These changes should be recorded so that their effects can be used in the evaluations.

GUIDELINES FOR SECTION 2.4 **Post-drive questions administration**	
4.A	Post-drive questions must, at the very least, be filled in right after the last test ride. This is valid for all the test drivers, including safety driver/professional driver, naïve drivers or passengers, and irrespective of whether a driver has multiple drives.
4.B	If you collect this survey data (post-drive questions) multiple times, make sure there is a substantial and quantifiable change in either 1) the participant's experience with the ADF, 2) the time since last using the ADF, or 3) the behaviour or ability of the ADF. This change data should be recorded.

Exporting questionnaire responses

Following the completion of user and acceptance questionnaires, test participants' responses results must be exported as recommended in this section to ensure a common data format in the consolidated database for user and acceptance evaluation area.

If LimeSurvey has been used for implementation of the questionnaire, please follow the LimeSurvey Subjective questionnaire user guide to export questionnaires. To export questionnaire responses, please click on the **Export** menu and then on **Export responses to SPSS** as illustrated below.



Once the response data export process has been completed, questionnaire responses output will be generated in SPSS, which is the abbreviation for Statistical Package for the Social Sciences. SPSS is used worldwide by researchers for statistical data analysis. As an example, the figure below shows export responses and responses output generated in SPSS, respectively.

Export response data to SPSS

Data selection: All responses | Complete only | Incomplete only

SPSS version: Prior to 16 | 16 or up

Limit:

Offset:

No answer:

Step 1: Export syntax

Step 2: Export data

Instructions for the impatient :

1. Download the data and the syntax file.
2. Open the syntax file in SPSS in Unicode mode.
3. Edit the 'FILE=' line and complete the filename with a full path to the downloaded data file.
4. Choose 'Run/All' from the menu to run the import.

Your data should be imported now.

EXAMPLE_TJM Questionnaire_SPSS output_Untitled2.sav [] - IBM SPSS Statistics Data Editor

File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Add-ons Window Help

	Name	Type	Width	D...	Label	Values	Missing	Columns	Align	Measure
1	id	Numeric	7	0	id	None	None	8	Right	Nominal
2	submitdate	Date	23	2	submitdate	None	None	23	Right	Scale
3	lastpage	Numeric	7	0	lastpage	None	None	8	Right	Scale
4	startlanguage	String	20	0	startlanguage	None	None	20	Left	Nominal
5	q_	String	31	0	Seed	None	None	31	Left	Nominal
6	startdate	Date	23	2	startdate	None	None	23	Right	Scale
7	datestamp	Date	23	2	datestamp	None	None	23	Right	Scale
8	TJM1	Numeric	4	0	1. What year were you born?	{3, 1900}...	None	4	Left	Nominal
9	TJM2	Numeric	2	0	2. What is your gender?	{1, Male}...	None	6	Left	Nominal
10	TJM3	Numeric	2	0	3. Country of residency	{1, Belgium}...	None	7	Left	Nominal
11	TJM3_other	String	1	0	[Other] 3. Country of residency	None	None	9	Left	Nominal
12	TJM4	Numeric	2	0	4. What is the highest level of education that you have completed (including ong...	{1, trade/technical/vocational training}...	None	2	Left	Nominal
13	TJM5	Numeric	2	0	5. What is your employment status?	{1, Employed full-time}...	None	2	Left	Nominal
14	TJM6	Numeric	2	0	6. Could you do part of your job whilst on transportation e.g. travelling on a bus, t...	{1, Yes}...	None	2	Left	Nominal
15	TJM7	Numeric	2	0	7. Do you have a car available for your use?	{1, Yes, (nearly) always}...	None	2	Left	Nominal
16	TJM8_A1	Numeric	1	0	[I am an employee of a vehicle manufacturer or supplier] 8. Please tick all of tho...	{0, Not selected}...	None	8	Right	Nominal
17	TJM8_A2	Numeric	1	0	[I work in the development of automated vehicle functions] 8. Please tick all of th...	{0, Not selected}...	None	8	Right	Nominal
18	TJM8_A3	Numeric	1	0	[I test automated vehicle functions] 8. Please tick all of those that apply to you i...	{0, Not selected}...	None	8	Right	Nominal
19	TJM8_A4	Numeric	1	0	[I have a professional driving qualification] 8. Please tick all of those that apply to...	{0, Not selected}...	None	8	Right	Nominal
20	TJM8_A5	Numeric	1	0	[I am a qualified safety/test driver] 8. Please tick all of those that apply to you in ...	{0, Not selected}...	None	8	Right	Nominal

Common Data Format

To ensure that data can be transferred to the consolidated database, it is extremely important to use a common data format across pilot sites. Although different export options are available in LimeSurvey, SPSS is the preferred one for transferral and adhering to a common data format.

GUIDELINES FOR SECTION 3.2 **Exporting results**	
5.A	Export responses to SPSS.
5.B	Export responses as a CSV file format in Excel.
5.C	If questionnaires are not administered in English, answers to open questions (e.g. TMJ.28, TMJ.29, TMJ.30b, TMJ.37; P.25, P.26, P.30; U.28, U.29, U.30b, U.37) will need to be translated into English before being committed to the consolidated database.

Annex 2: Guidelines for information & instructions for participants

This annex contains guidance on the instructions and information that participants should be given *before piloting*. The aim is to ensure that relevant information is not omitted and that the information that is given does not overtly bias the participants' perceptions, behaviours, or responses.

The guidance primarily focuses on information and instructions for ordinary participants (drivers and passengers). In other words, participants who are not trained/professional/safety drivers, or who are otherwise unfamiliar with the ADF. However, there are some aspects that are still relevant to ALL drivers. This is made clear under the 'driver type' column.

At the bottom of this document, there are additional instructions to the experimenter/test leader. Please take note of these.

Advice/ Guideline	**Driver type**		**Study phase**		**Information**	
	All driver types	Ordinary drivers/ passengers	Baseline	Experimental	Guideline keyword	Guideline
A	X	X	X	X	*Project introduction*	1–2 standardised sentences on the background of L3Pilot. *Example:* L3Pilot project is piloting, testing and evaluating automated driving functions in passenger cars. Our tests will expose automated functions to different users and road environments in Europe.
B	X	X	X	X	*Experiment introduction*	1–2 sentences on the background of the particular experiment you are running. *Example:* During this study, you will sit in XXXX and observe and experience automated driving in a highway and motorway(/urban) environment (/while parking the car). The test route is __ km, and it takes approximately __ minutes/hours.
C	X	X		X	*ADF introduction*	Give a short description of what the ADF does and where it can operate (ODD). You may mention that the vehicle they will be using in the experiment is a prototype and that how it looks (internally and externally) and how it behaves may not necessarily represent a future market-ready vehicle.
D	O	X		X	*ADF limitations*	Provide information (in layman's terms) about the ADF's technical limitations. This should not be a comprehensive list of ADF limitations, but rather aspects that are crucial to ensure the participants' safety or are assessed to be otherwise critical (if any). This is to avoid negatively biasing participants' perception of the system and, therefore, their behaviour or responses.
E	X	X		X	*ADF operation*	Give detailed information about how to fully operate the ADF in order to perform the task they have been given, for example observe how the driver acts. This includes how to engage/accept the automated driving mode and how to resume control (turn off automation) when the system prompts the driver to take over, as well as what the user interface looks like and how it communicates various information (e.g. engagement/disengagement, warnings, etc.). For example, if you are running a motorway pilot, you should consider explaining how to change lane using the system.

Advice/ Guideline	Driver type: All driver types	Driver type: Ordinary drivers/ passengers	Study phase: Baseline	Study phase: Experimental	Information: Guideline keyword	Information: Guideline
F	X	X	X	X	*Experiment details*	Give a detailed description of the task that participants will be expected to perform. This includes: • Per/post drive questionnaires • Experimental environment (e.g. open road, test track), • Task (e.g. route they will be driving and/or manoeuvres they will be performing), • Engagement in secondary tasks if allowed during automated driving (e.g. browsing the internet), • Timing (e.g. how long each aspect of the experiment will take) • Payment/remuneration (e.g. whether and how participants will be compensated for taking part)
G	X	X	X	X	*Safety instructions*	Include safety-related instructions. This pertains to what participants should and should not do during the experiment to avoid unsafe situations, as well as what to do in emergencies.
H		X		X	*Interaction with safety driver*	Give instructions about how participants should interact/communicate with the safety driver in the vehicle (if they are allowed to do so).
I	X			X	*Interaction with ordinary passenger*	Give instructions on: • How the safety driver should interact/communicate with the participant in the vehicle. (Communication should be limited to safety-related and operational aspects of the experiment. • What non driving related tasks the ordinary driver/ passenger will be allowed to do when the ADF is activated. • Whether they may answer questions posed by the ordinary driver/passenger.
J	O	X		X	*Answer questions*	Allow participants to ask questions related to the experiment.
K	O	X		X	*Ethics, safety and confidentiality*	• Inform participants that it is important that they understand that you are **not looking at their individual performance, abilities or perceptions, but that you are interested in the behaviour of a group to draw collective conclusions.** • Mention that the study is subject to the General Data Protection Regulation (GDPR) and that the project has ethical approval. • Mention that at no time now or in the future will any of the information they provide be published that allows them as an individual to be identified, and all video data (if

Advice/ Guideline	Driver type		Study phase		Information	
	All driver types	Ordinary drivers/ passengers	Baseline	Experimental	Guideline keyword	Guideline
						collected) are used for research purposes only, with any presentations protecting their identity. • Inform them that they are free to withdraw from the study at any time, without needing to give a reason for doing so.
L	X	X	X	X	*Contact*	Provide contact details of the person responsible for the study, or whom participants can contact should they have any further enquiries.
	NA	X	X	X	*Informed consent*	Participants should be asked to sign an informed consent form, which serves to confirm that they have understood the information that they were provided and are happy to participate in the experiment.

X = Yes, O = Optional, NA = Not applicable

Instructions to the experimenter/test leader

- The information sheet should be presented in printed format to avoid deviations from the agreed content.
- Participants should read the information sheet themselves.
- The instructions should be kept simple and clear.
- Tests should be conducted by as few experimenters as possible to avoid inter-experimenter variance. Where there is more than one experimenter, it is important that the same protocol is followed.

Annex 3: Scenarios for simulation of the safety impacts

ODD	ID	Driving scenario	Covered sub-scenario	Description	Simulation tool
Motorway & Traffic jam	P1	**Rear-end conflict**	Slower front vehicle, braking front vehicle, (standstill front vehicle)	In this scenario, rear-end conflicts are simulated in which both vehicles are driving.	MATLAB based tool
	P2	**End of traffic jam conflict**	Standstill front vehicle, very slow front vehicle	In this scenario, rear-end conflicts are simulated in which the front vehicle is driving at slow speed or standing still	MATLAB based tool
	P3	**Cut-in conflict**	Cut-in from left (braking) / right (slower)	In this scenario, a vehicle enters the ego vehicle's lane from an adjacent lane (left or right) at a relatively short distance. The speed of both vehicles must vary but the ego vehicle should be at least as fast as the other vehicle.	MATLAB based tool
	P4	**Lane change conflict**	Other vehicle approaching in blind spot (both left and right)	In this scenario, the ego vehicle performs a lane change to an adjacent lane (mainly left) in which there is other moving traffic. The velocity and position of the vehicles in the new lane must vary.	openPASS
	P5	**VRU conflict**	Standing still (lying), moving from left to right and right to left	In this scenario, a pedestrian is either crossing or standing on the road while the ego vehicle is approaching.	openPASS
	N1	**Minimum risk manoeuvre**	Different reaction time for take-over	The ego vehicle is driving automated in the lane (randomly chosen) surrounded by traffic, which is defined by different traffic densities. After a certain time, the minimum risk manoeuvre is initiated and slows down the vehicle (profile to be decided) until the driver takes over. The manoeuvre does not consider a lane change. Take-over time of the driver is to be defined based on distribution.	openPASS
	N2	**Wrong activation**	Different times for regaining control	The ego vehicle is being driven manually in lane, surrounded by different traffic densities. After a certain time (→ wrong activation by the driver), the ego vehicle starts to roll without lateral control (deceleration defined by driving resistance). After a certain time (defined by a distribution) the driver recognises the mistake and regains control.	openPASS

ODD	ID	Driving scenario	Covered sub-scenario	Description	Simulation tool
	N3	**End of lane**		The ego vehicle is driving in a lane that is about to end. Surrounding traffic is defined by traffic density. The ego vehicle has to perform a lane change into the adjacent lane.	openPASS
	N4	**Obstacle in the Lane**	Different positions of the object, different number of lanes blocked	The ego vehicle is approaching an obstacle in the lane and has to react to the situation (brake or change lane if possible). Surrounding traffic is defined by traffic density.	openPASS
	N5	**Lower speed limit**	Different speed limits, series of lower speed limits	The ego vehicle is approaching a new, lower, speed limit. The eco vehicle decelerates earlier than the manually driven vehicles to comply with the traffic rules, causing other traffic to react. Traffic is defined by traffic density.	openPASS
	N6	**Motorway entrance (passing)**		The ego-vehicle is driving in the right-hand lane approaching a motorway entrance. Another vehicle is planning to enter the motorway at the entrance (whether there is this other vehicle (or not) and time/distance between the vehicles is chosen randomly). Both vehicles have to react.	openPASS
Urban	CVL	**Crossing**	Vehicle from left (occluded by lead vehicle)	The ego vehicle passes an intersection with another vehicle crossing its path from the left.	VTD Toolchain
	CVR		Vehicle from right	The ego vehicle passes an intersection with another vehicle crossing its path from the right.	VTD Toolchain
	CPL1		Pedestrian from left before nodal point	The ego vehicle passes an intersection with a pedestrian crossing from the left before the nodal point.	VTD Toolchain
	CPL2		Pedestrian from left after nodal point	The ego vehicle passes an intersection with a pedestrian crossing from the left after the nodal point.	VTD Toolchain
	CPL1		Pedestrian from right before nodal point	The ego vehicle passes an intersection with a pedestrian crossing from the right before the nodal point.	VTD Toolchain
	CPL2		Pedestrian from right after nodal point	The ego vehicle passes an intersection with a pedestrian crossing from the right after the nodal point.	VTD Toolchain
	RVL	**Right turn**	Vehicle from left	The ego vehicle executes a right turn at an intersection with another vehicle crossing its path from the left.	VTD Toolchain
	RPL1		Pedestrian from left before nodal point	The ego vehicle executes a right turn at an intersection with a pedestrian	VTD Toolchain

ODD	ID	Driving scenario	Covered sub-scenario	Description	Simulation tool
				crossing from the left before the nodal point.	
	RPL2		Pedestrian from left after nodal point	The ego vehicle executes a right turn at an intersection with a pedestrian crossing from the left after the nodal point.	VTD Toolchain
	RPR1		Pedestrian from right before nodal point	The ego vehicle executes a right turn at an intersection with a pedestrian crossing from the right before the nodal point.	VTD Toolchain
	RPR2		Pedestrian from right after nodal point	The ego vehicle executes a right turn at an intersection with a pedestrian crossing from the right after the nodal point.	VTD Toolchain
	LVL1	**Left turn**	Vehicle from left	The ego vehicle executes a left turn at an intersection with another vehicle crossing its path from the left.	VTD Toolchain
	LVR1		Vehicle from right	The ego vehicle executes a left turn at an intersection with another vehicle crossing its path from the right.	VTD Toolchain
	LVL2		Oncoming Traffic	The ego vehicle executes a left turn at an intersection with vehicle approaching from the oncoming direction.	VTD Toolchain
	LPL1		Pedestrian from left before nodal point	The ego vehicle executes a left turn at an intersection with a pedestrian crossing from the left before the nodal point.	VTD Toolchain
	LPL2		Pedestrian from left after nodal point	The ego vehicle executes a left turn at an intersection with a pedestrian crossing from the left after the nodal point.	VTD Toolchain
	LPR1		Pedestrian from right before nodal point	The ego vehicle executes a left turn at an intersection with a pedestrian crossing from the right before the nodal point.	VTD Toolchain
	PLR2		Pedestrian from right after nodal point	The ego vehicle executes a left turn at an intersection with a pedestrian crossing from the right after the nodal point.	VTD Toolchain
	PL	**Going straight**	Pedestrian left	The ego vehicle follows a straight road with a pedestrian crossing from the left.	VTD Toolchain
	PR		Pedestrian right	The ego vehicle follows a straight road with a pedestrian crossing from the right.	VTD Toolchain
	A		Approaching Lead Object	The ego vehicle follows a straight road behind a lead vehicle which decelerates suddenly.	VTD Toolchain

ODD	ID	Driving scenario	Covered sub-scenario	Description	Simulation tool
	A		Approaching static object	The ego vehicle follows a straight road with a static object in its path.	VTD Toolchain
	RA	**Right turn**	Lead object	The ego vehicle is in a right turn behind a lead vehicle which decelerates suddenly.	VTD Toolchain
	LA	**Left turn**	Lead object	The ego vehicle is in a left turn behind a lead vehicle which decelerates suddenly.	VTD Toolchain
	OV	**Oncoming vehicle**		An oncoming vehicle changes to the lane of the ego vehicle to overtake an object.	VTD Toolchain
	CI	**Cut-in**	Left / right	A slower or decelerating vehicle cuts into the lane of the ego vehicle	VTD Toolchain
	LC	**Lane change**	Left / right - merge	The ego vehicle executes a lane change or merge either from left or right	VTD Toolchain

Annex 4: Method for assessing the safety impact of parking ADF

Research into suitable methods

The method for the safety impact assessment of parking ADF was defined based on an international literature review. The review revealed that a former European project, AdaptIVe, aimed to assess parking characteristics of vehicles (i.e., parking space needed for one vehicle) rather than their safety potential (Fahrenkrog et al., 2017); hence that methodology cannot be transposed and applied to the L3Pilot safety impact assessment.

The German Insurance Association (GDV) and Allianz Center for Technology (AZT) have conducted several retrospective and prospective effectiveness analyses of driver assistance systems and ADFs on the basis of insurance claims databases (GDV report, 2017; Gwehenberger & Borrack, 2015; Borrack et al., 2017; Ostermaier, 2018; Ostermaier, et al., 2019). Furthermore, Swiss Re estimated the impact that the growth of car connectivity and ADAS take-up rate could have on the market and distribution of motor insurance (Swiss Re & HERE, 2015).

In the following we assume that the methodology outlined in such works could be suitable for evaluating the safety potential of Parking ADF and for answering the current research questions (see Table 2.3). The additional data from the field experiments in L3Pilot will support the assessment of the robustness of the methodology and refine it to the data received.

Proposed Methodology

Accident databases containing claims of vehicles without ADFs (baseline) are used for the estimation of the expected safety benefits. Insurance claims offer a comprehensive basis for the evaluation, especially regarding parking accidents: parking and manoeuvring accidents, excluding those involving personal injuries, are in most cases reported only to the insurance company, not to the police.

For a holistic assessment of the impact of ADFs on the occurrence of damage, the following four key elements will be taken into account:

- Target accidents: The theoretical maximum percentage of claims that can be addressed by the respective Parking ADF.
- Effect: The proportion of target accidents that can be prevented under real road traffic conditions (e.g. bad weather or road conditions or technical failures reduce the efficiency of a certain system).
- Usage: The extent to which the Parking ADF is used by drivers (switched on/off by drivers).
- Market penetration: The percentage of cars on the road that is equipped with Parking ADF.

In L3Pilot, the socio-economic impact assessment has set scenarios for penetration rate in use (5%, 10%, 30% and 100%) within the ODD. Therefore, the last two items on the list are combined in these scenarios.

The figure below shows the scheme of the method utilised to evaluate the number of avoidable accidents through Parking ADF. Based on all accidents contained in the claims databases, an ex-ante analysis is used to determine the maximum number of accidents that could be avoided theoretically with the respective Parking ADF (*target accidents*). For this purpose, the insurance claims databases (e.g. Allianz, other insurance companies) will be filtered according to the operational design domain of the respective Parking ADF.

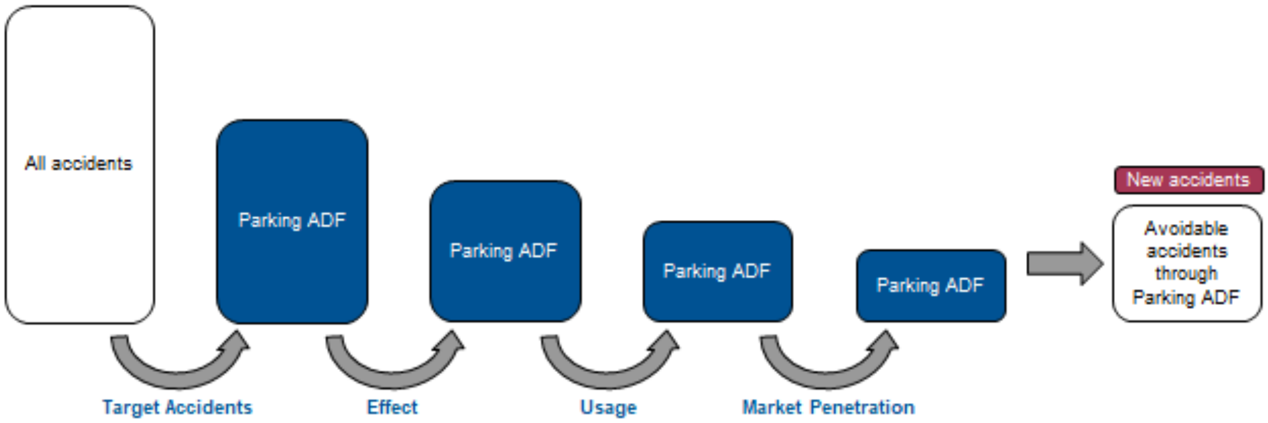


Figure: Method to evaluate the expected safety effects of parking ADF (schematic).

The next step of the methodology accounts for the impact of the parameter *effects* on the number of target accidents. The effect explicates the share of maximum avoidable (relevant) claims under real conditions in road traffic. For example, missing road markings or weather influences can lead to restrictions in this respect. The factor *effect* is based on the description of the mature ADF and on data from the field experiments in L3Pilot.

Presumably, the main restriction limiting the effect will be due to the operating domain of the studied ADFs. For instance, a limitation of an ADF to home zone parking will limit the availability of the ADF to a rather small subsample of parking manoeuvres and, therefore, it could effectively impact only a subset of all parking accidents. There are research questions dealing with the actual safety impact of parking ADFs already at the stage of technical & traffic evaluation. For instance, the frequency of safety incidents while parking and other technical characteristics of the parking performance will be analysed. Based on those results and the description of the mature ADF, the efficiency of the ADF will be assessed.

The parameter *degree of usage* represents the actual frequency of use of the Parking ADF by the driver. The experimental procedure of the parking ADF tests in L3Pilot do not allow estimation of the true degree of usage. Therefore, the use was combined with the penetration rate – penetration rate in use (within the ODD).

Multiplying these ratios for the respective Parking ADF will yield the potential for reducing claims for the period under review after the market launch. The table below illustrates the procedure exemplarily by using fictitious values. The step-by-step reduction of the accident avoidance potential can also be clearly seen from the sample calculation. While 20% of all losses considered are theoretically within the operating area of a fictitious parking ADF, the share of avoidable claims is reduced to 2.4% if the dummy values for effect, usage and market penetration are taken into account.

Table: Procedure for determining the safety benefit of parking ADF (Values are only exemplary).

Target accidents		Effect		Penetration in use		Expected benefit
20%	x	40%	x	30%	x	2.4%

In the safety impact assessment of the parking ADF, the figure *effect* will be based on results from the field experiments in L3Pilot together with the analyses of insurance claims databases.