Hi-Drive
Designing Automation

Deliverable D4.3 /

Experimental procedure

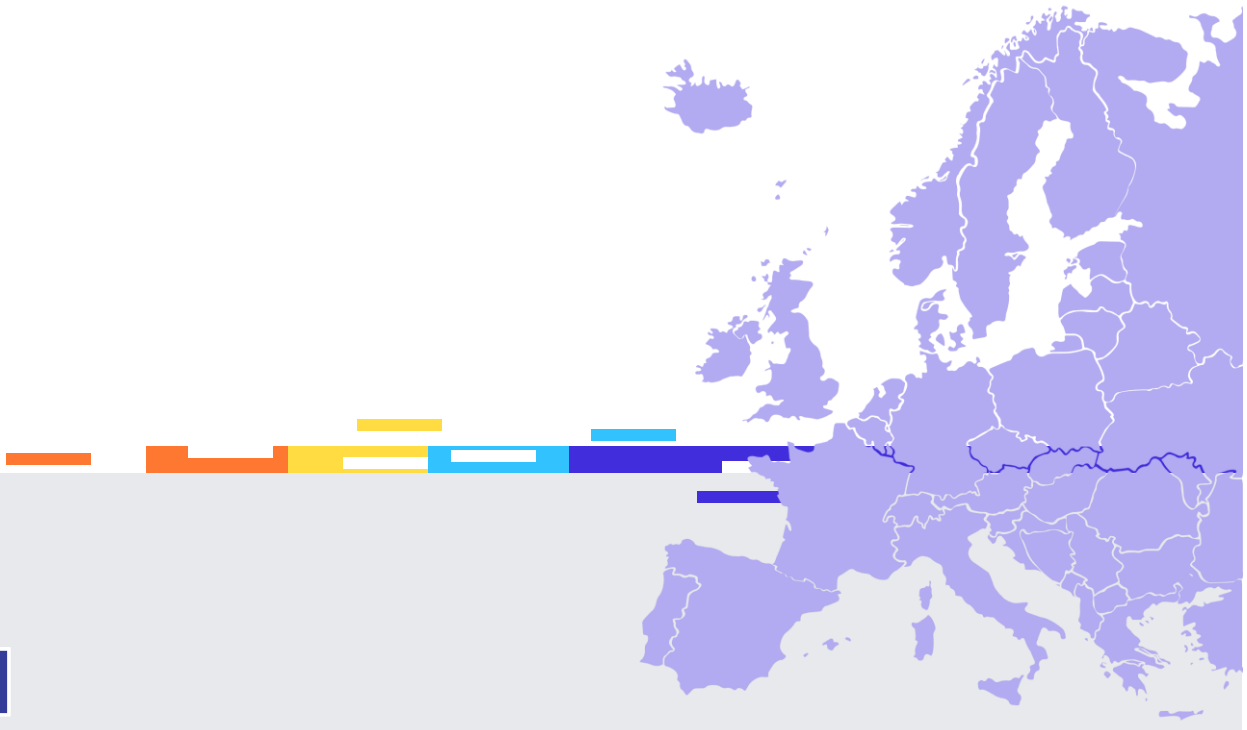
Version: 1.1

Dissemination level: PU

Lead contractor: VTT

Due date: 31.01.2023

Version date: 04.07.2023



This project has received funding from the
European Union's Horizon 2020 research and innovation
programme under grant agreement No 101006664.

Document information

Authors

Henri Sintonen – VTT
Anastasia Bolovinou – ICCS
Burak Guelsen – BMW
Jeroen Hogema – TNO
Satu Innamaa – VTT
Yee Mun Lee – University of Leeds
Jorge Lorente Mallada – TME
Barbara Metz – WIVW
Felix Reimer - fka
Fabio Tango – CRF
Lennart Vater – ika
Benjamin Altpeter - fka
Felix Fahrenkrog - BMW
Teemu Itkonen - VTT
Esko Lehtonen – VTT
Ruth Madigan – University of Leeds
Hendrik Weber – ika

Coordinator

Aria Etemad
Volkswagen AG
Berliner Ring 2
38440 Wolfsburg
Germany

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

Project funding

Horizon 2020
DT-ART-06-2020 – Large-scale, cross-border demonstration of connected and highly automated driving functions for passenger cars
Contract number 101006664
www.Hi-Drive.eu

Legal Disclaimer

Table of contents

List of figures

List of tables

Executive summary

Experimental procedure is a crucial part of the Hi-Drive methodology, as it influences data collection and subsequent evaluation. Its aim is to ensure that the required data is collected in a way that the evaluation team can answer the research questions set for the project.

The development of the experimental procedure in Hi-Drive was largely done through collaboration with relevant work packages across different subprojects. First, it was important to start with an understanding of the goals, needs, and limitations of each Hi-Drive operation and its evaluation. Then, this information was processed and the experimental procedure instructions and recommendations were formulated.

Specifically, the steps of the work were as follows:

- Gather information on the goals of the project, research interests of partners, plans of the operation owners (i.e., those executing the experiments), and recommendations from previous projects and the first drafts of the evaluation plan.
- Study the international state-of-the-art and good practices.
- Collaborate with the other work packages across all relevant subprojects to ensure that everyone involved is working in parallel towards the same goal with compatible ideas.
- Form the instructions and recommendations for the operations.
- Communicate these instructions and recommendations, and check with the operation owners whether their operation is in line with them, and search for solutions to any issues that may be encountered.

This deliverable is structured in two main parts: one providing information for the methodology and evaluation teams to better understand the operation landscape, and the other focusing on providing instructions for the operations on the experimental procedure and study design.

For the evaluation, the operations were grouped based on their experimental setup. This is important, as one of the tasks they need to solve is the puzzle of which operations the performance indicator data could be pooled before statistical tests will be performed, from which operations the results of operation-specific statistical tests could be merged, and which operations would have to be analysed and reported individually.

Additionally, the operations were summarised, analysed, and checked for how well they line up with the instructions given. After this, the feasibility of the research questions was evaluated from the viewpoint of the experimental procedure. They were deemed feasible if

the instructions were followed and if the pooling of data or 'merging of results' puzzle could be solved by the related work packages.

The aim of the operations is to push the state-of-the-art towards highly automated driving (HAD) in various ways, but most often either through extending the Operational Design Domain (ODD) of the automated driving functions (ADFs) or by improving the automated driving (AD) performance in a wide set of conditions or manoeuvres. This push is accomplished by integrating various technological enablers into the automated vehicles. In order to evaluate the ODD extension and AD performance goals, the following instructions were provided for the operations:

Regarding the test environment, they should

- not test in isolation;
- use chained use cases if possible, i.e. all use cases that can be considered by the enablers should be included in the same trip;
- introduce or balance conditions regarding time of day, weather and road surface conditions, static and dynamic road infrastructural elements, and different traffic volumes.

To perform the comparisons required to answer the research questions, the operation should collect treatment data that is *automated driving supported by the enabler technologies*. This will be compared to baseline data, for which the instructions are:

- The operations should collect the *manual driving* baseline and *AD without enablers* baseline, in both the extended ODD and nominal ODD whenever possible.
- The total data amount should be such that 50% of it is treatment data and the other 50% is equally split between baseline conditions.

For the quantity of data that should be collected, the instruction is:

- As an overall order of magnitude, the operations should collect "at least several hundreds" of observations of interest. Here, an observation is any scenario or trip for which a performance indicator can be calculated for comparison between baseline and treatment.

Additional instructions are given regarding the selection of participants and administration of the pre- and post-test drive questionnaire.

All the instructions and recommendations given in this deliverable are formulated in a general way. Thus, the operations and the partners involved in the processing of the data need to plan in detail how to apply them in practice. The status of the operation was checked before publication of this deliverable, and no major issues were found regarding compliance with these instructions.

1 Introduction

1.1 The Hi-Drive project

Connected and automated driving (CAD) has become a megatrend in the digitalisation of society and in the economy. CAD has the potential to drastically change transportation and to create far reaching impacts. SAE level 3 (L3) automated functions were piloted in Europe by the L3Pilot project in 2017–2021 (L3Pilot consortium 2021). Hi-Drive builds on the L3Pilot results and advances the European state-of-the-art from SAE L3 'Conditional Automation' further up towards 'High Automation'. This is done by demonstrating in large-scale trials the robustness and reliability of CAD functions under demanding and error-prone conditions with special focus on:

- Connected and automated vehicles (CAV) travelling in challenging conditions covering variable weather and traffic scenarios and complex infrastructure
- Connected and secure automation providing vehicles/their operators with information beyond the line of sight and on-board sensor capabilities
- Complex interaction with other road users in normal traffic
- Factors influencing user preferences and reactions including comfort and trust—and eventually through a wide consumer acceptance of automated driving (AD) resulting in purchase and use, enabling viable business models for AD.

The project's ambition is to extend the CAD's operational design domain (ODD) from the state-of-the-art level 3 systems, which frequently demands taking over control of the vehicle by a human driver. As experienced in the EU flagship pilot project L3Pilot, on the way from A to B, a prototype level-3 automated vehicle (AV) encountered a number of ODD boundaries, leading to fragmented availability of the AD function. Hi-Drive addresses these key challenges which are currently hindering the progress of vehicle automation. The concept builds on reaching a widespread and continuous ODD, where automation can operate for longer periods, and the interoperability is assured across borders and brands. Hi-Drive strives to extend the ODD and reduce the frequency of take-over requests (TORs) by selecting and implementing technology enablers leading to highly capable CAD functions, operating in diverse driving scenarios including, but not limited to, urban traffic and motorways. The removal of fragmentation in the ODD is expected to give rise to a gradual transition from conditional operation towards higher levels of AD.

The work in Hi-Drive started in July 2021 with the collection and description of the different AD functions, their ODDs and limitations, and the enabler technologies that help overcome these limitations. When testable functions and use cases of driving automation were defined,

research questions were formulated, leading to specification of data needed for evaluation and recording of vehicle and driver behaviour.

The evaluation will focus on three areas: 1) users; 2) AD performance and possible extension of the ODD; and 3) assessment of impacts (on safety, efficiency, environment, mobility, transport system, and society). Furthermore, these assessments serve as input to determine whether the socioeconomic benefits of higher driving automation outweigh the costs. The project also engages in a broad dialogue with the stakeholders and the general public to promote the Hi-Drive results. Dissemination and communication are boosted by demonstration campaigns to show project achievements.

Overall, Hi-Drive strives to create a deployment ecosystem by providing a platform for strategic collaboration. Accordingly, the work includes an EU-wide user education and driver training campaign and series of Codes of Practice (CoP) for the development of automated driving functions and road-testing procedures, while also leading the outreach activities on standardisation, business innovation, extended networking with interested stakeholders, and coordinating parallel activities in Europe and overseas.

1.2 Overall implementation plan for Hi-Drive

Implementation of a large project involving a multitude of experiments and wide-ranging evaluations like Hi-Drive requires a solid implementation plan. The FESTA Handbook (FOT-Net, CARTRE & ARCADE (2021)) compiles the knowhow gained since 2007 on testing and evaluation of driver support systems and functions. The FESTA methodology was designed for field-operational tests (FOTs) with market-ready products. Therefore, it does not fully apply to studies with prototypical AD functions[1] (ADFs). Thus, some adjustment of the FESTA implementation plan, described as the "FESTA-V" structure, was needed to accommodate the testing of AD.

Figure 1.1 illustrates the FESTA implementation plan adapted for Hi-Drive. The plan is divided into three phases: (I) prepare, (II) operate, and (III) evaluate. In the beginning of the preparation phase (I), ADFs, the technology enablers, and their use cases and associated test scenarios across multiple test environments (test track, open road, simulation) are selected and described in detail. Then, an initial list of research questions is set up and organised as high-, medium-, and low-level questions. The state-of-the-art is summarised for topics covered by these research questions. The feasibility of each research question is checked next

[1] According to the Hi-Drive glossary: Automated driving function (ADF) is a common feature addressed by a group of automated driving systems, for example: Motorway ADF, Urban ADF

in terms of data availability, suitability of the experimental design and procedures, availability of research tools, methods and external data sources, and availability of resources (e.g., project duration and human) required.

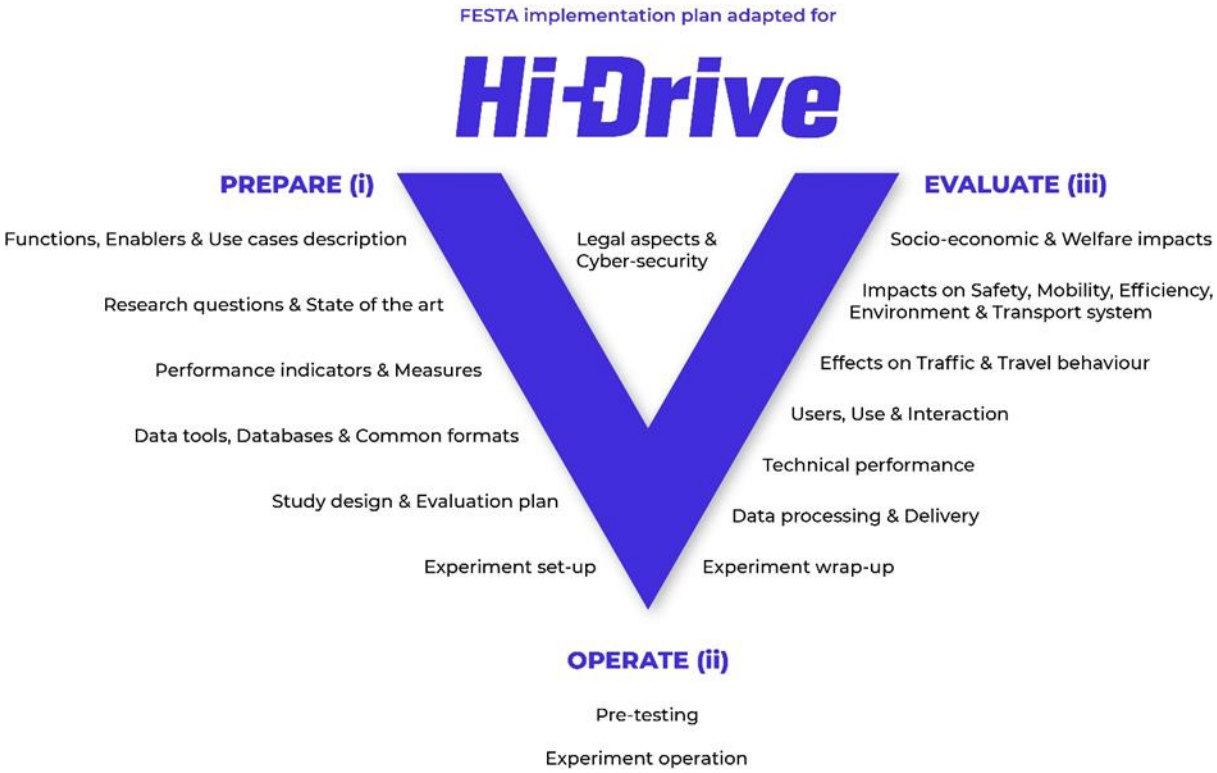


Figure 1.1: FESTA implementation plan adapted for Hi-Drive.

Next, the performance indicators and other data with which the research questions are answered, and the evaluation tools, are defined and calibrated. Based on these requirements for evaluation, five lists—one for different data categories[2]—with the required information are defined. In the following step, the five lists with the required information are merged into one signal list which specifies all the signals needed. Next to the signal list, a common data format (CDF) applicable to the project evaluation is specified for them. The data to be shared for evaluation is agreed with the data providers. Various databases and data tools are defined for data processing and storage.

The experimental design and procedures are set to test highly automated driving and its technology enablers, and to provide data on them for evaluation. The plans for all operation sites are approved between the site owners and those setting the methodology for evaluation.

[2] The data categories are closely linked to the different databases which will become the tool for making the data available for evaluation.

An evaluation plan is developed for each research question to agree on who is responsible for what, to specify the methods, tools, and data to be used, scenarios to be addressed, and to plan the dependencies, i.e., linking the inputs and outputs as well as their timeline.

The experiment setup includes preparation of test vehicles, testing of selected parts of the technology and use cases, getting permissions, selection of participants, and implementation of data logging.

The operation phase (II) starts with the pre-testing step. It involves running all the phases of the project on a small scale to ensure that all the processes and tool chains function as intended. Once everything is confirmed to work as intended, the experiment operation begins. This phase involves the actual data collection.

The evaluation phase (III) starts with the data delivery as part of the experiment wrap-up. 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. The data are converted to CDF, processed, and delivered to the evaluation team.

In the effects evaluation, technical performance of the tested technology is assessed. User evaluation focuses on the users, usage, and interaction. Effects on traffic and travel behaviour are assessed together with their societal impacts on safety, mobility, efficiency, and environment and later scaled up to European level. The final step is to assess the socioeconomic and welfare impacts.

1.3 Activity objective, scope, and structure of the deliverable

To be able to follow the described implementation plan in a structured way, the work within Hi-Drive is organised as subprojects (SP). This deliverable is part of the *Methodology* (SP4) subproject. The objectives of this subproject are to:

- Specify the Hi-Drive research questions for both *Users* (SP6) and *Effects* (SP7) evaluation, how they will be addressed, and the related data needs.
- Agree on CDF for provision of different datasets.
- Agree on experimental design and procedures for testing and evaluation of ADFs and related enablers in challenging environments.
- Reconsider the theoretical background and impact mechanisms to build a multidisciplinary evaluation methodology, covering not only the expected positive impacts on safety, comfort, and the environment, but also the unintended and possibly negative impacts on users and the transport system.

- Refine the state-of-the-art methods to address user and human-factor aspects of high-level driving automation and facilitate understanding of possible effects on the transport system level, addressing travel behaviour, safety, efficiency, and emissions.
- Provide lessons learned from the methodology point of view.

Within SP4 *Methodology*, three deliverables define the basic requirements and goals of data collection from a methodological point of view: Specifically, these are the deliverables on the research questions (D4.1 *Research questions* by Metz et al. 2023), on the data requirements (D4.2 *Data for evaluation* by Fahrenkrog et al. 2022), and on the experimental design (this deliverable D4.3 *Experimental procedure*). Based on that, detailed analysis plans for user evaluation (D4.4 *User evaluation methods*) and effects evaluation (D4.5 *Effects evaluation methods*) will be developed. All input and requirements will guide the work in SP5 *Operations*, which will collect the data needed for effects evaluation, and SP7 *Effects* which will analyse it to answer the research questions on effects. In a similar way, SP6 *Users* will collect and evaluate data to answer the user-related research questions. At the end of Hi-Drive, the project results on the research questions will be presented in the deliverables of SP6 *Users* and SP7 *Effects*. See Figure 1.2 for an overview.

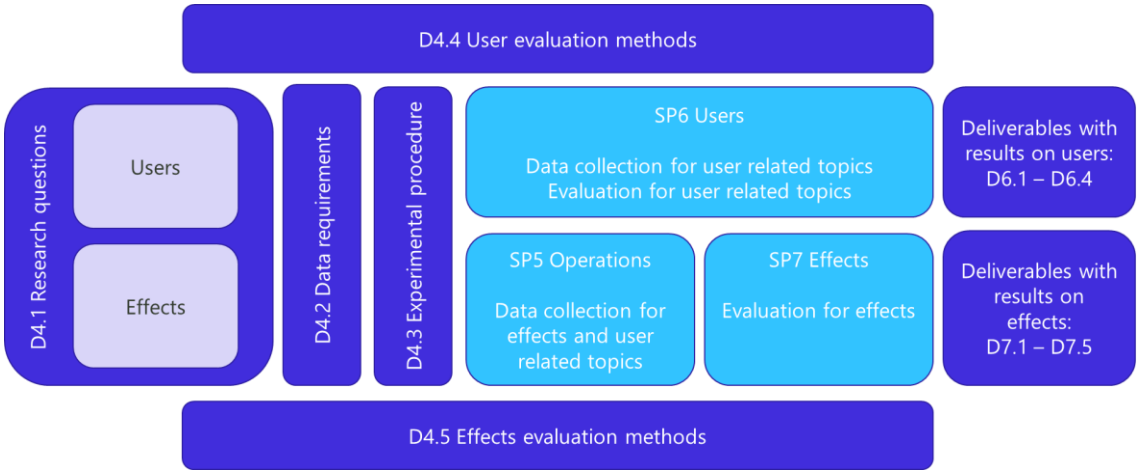


Figure 1.2: Role of different deliverables on methodological and evaluation results

This deliverable reports on the activities of Work Package (WP) 4.5 *Experimental procedure* under SP4 *Methodology*. The purpose of the WP is to ensure that the required data for the evaluation are collected in the operations with appropriate experimental design and procedures. This WP also provides input for feasibility checking of the initial list of Hi-Drive research questions (WP4.3 *Research questions*) and for setting the evaluation plans (WP4.6 *Methods for user evaluation*, WP4.7 *Methods for effects evaluation*).

This deliverable provides information for the Hi-Drive project's internal stakeholders by detailing the instructions and recommendations given for the operation owners, i.e., for those responsible for the experiments[3] (operations[4]), and by providing summaries of the operation plans, from the viewpoint of experimental design and procedures, to the methodology and evaluation partners. This deliverable also provides information to external stakeholders regarding how field operations of driving automation should be conducted and what is the motivation behind the instructions and recommendations given. Additionally, the deliverable also includes a safety manual to support the safe execution of operations both within Hi-Drive and by others.

This report is structured as follows.

- Chapter 2 explains the process for development of the experimental procedure and other work done by WP4.5 *Experimental procedure*
- Chapter 3 provides a summary of Hi-Drive operations from the evaluation viewpoint and the results of the feasibility check of the research questions from an experimental procedure perspective
- Chapter 4 provides instructions for the Hi-Drive operations on the experimental procedure
- Chapter 5 is the summary and describes the outlook for future work to be continued in the other work packages
- Annex 1 includes the Safety Manual and Annex 2 detailed examples of power analysis for definition of data quantity requirements.

It should be noted that the report reflects the project's status as of January 2023, i.e., month 19 of the project. In a project that runs for 4 years, changes might occur that have implications on the experiments that will take place after publication of the deliverable.

3 Hi-Drive glossary: An "experiment consists of a series of test runs/trips to investigate a common aspect (ADF, Enabler, User) and is conducted under comparable circumstances. It is made up of several test runs/trips. Experiment types include open road, test track, driving simulator, simulation models, etc."
4 In the Hi-Drive glossary: An "operation is the execution of experiment(s) in a defined place and time."

2 Process for the development of the experimental procedure

Experimental procedure is a crucial part of the methodology in any empirical investigation and influences the data collection and subsequent evaluation. The development of the experimental procedure in Hi-Drive was largely done through collaboration with relevant work packages. First, it was important to start with an understanding of the goals, needs, and limitations of each of the operations and of the evaluation. Then, this information was merged, and the experimental procedure instructions and recommendations were formulated. It was also important to keep the partners of the relevant work packages WP4.3 *Research questions*, WP4.5 *Experimental procedure*, WP4.6 *Methods for user evaluation*, WP4.7 *Methods for effects evaluation*, and WP5.3 *Operation preparation* up to date with the development done in the other work packages.

Harmonised approaches and the ability to use shared, common tools is an aim for Hi-Drive's vehicle data analysis method to ensure that all data from whichever operation is evaluated in the same way and conclusions can be drawn. Therefore, from an effects evaluation methodology viewpoint, the biggest difficulty in setting up the experimental procedure was the various approaches the operation owners had planned for tackling their chosen challenges for the extension of the ODD and improvement of the AD performance. All the approaches to tackle these challenges did fit within the main goals of the project, but still led to a variety of experiment setups. Despite of all this variety, the aim was to provide a harmonised set of instructions and recommendations for the experimental procedure and to find feasible ways to implement these in practice. The steps of the work were as follows:

1. Gather information on the goals of the project, research interests of partners, plans of the operation owners, previous projects, and plans of the evaluation team
2. Study the international state-of-the-art and good practices
3. Collaborate with the other work packages across all relevant subprojects to ensure that everyone involved is working in parallel towards the same goal with compatible ideas
4. Form the instructions and recommendations for the operations
5. Communicate these instructions and recommendations, and check with the operation owners whether their operation is in line with them and search for solutions to any issues that may be encountered.

The collaboration realised by the interactions between the work packages has been both an absolute necessity in working towards this goal, and one of the most valuable outcomes of the work done in this work package. It has allowed the different areas of the project to work

in parallel with up-to-date information towards the shared goals. Specifically, the inter-WP interaction included:

- WP5.3 *Operation preparation* is responsible for description of the planned tests and for pre-testing everything in the operation. WP4.5 participated in the creation of the operation description templates in WP5.3. These templates were filled in by the operation owners, which allowed us to understand the similarities and differences of the plans of operation (a summary of operations can be found in D5.1 *Description of "Operations"* by Sauvaget et al. 2022). The information provided with the filled templates were complemented with a series of one-on-one discussions by WP4.5 with each of the operation owners to better understand their motivation for the operations, expectations for the outcomes of the operations, and the areas where there was flexibility for changes to their plans.
- WP4.3 *Research questions* defines the research questions for users and the effects evaluation. WP4.5 provided input to the development of the research questions and to their feasibility check. Furthermore, information about the research questions and the requirements on what would be needed to answer them were compared with the operation plans and provided to the operation owners.
- WP4.6 *Methods for user evaluation* makes the evaluation plan for SP6 *Users*. WP4.5 checked that the plans for the user experiments in Hi-Drive matched the overall goals of the project. In addition, through the one-on-one discussions with the operation owners, new opportunities for collaboration to address user-related topics were identified, as some operations were open to including a user component in their trial but lacked the resources to do it themselves. Within SP6 *Users*, there were partners that were capable and willing to handle them.
- WP4.7 *Methods for effects evaluation* makes the evaluation plan for SP7 *Effects*. It was important to understand in WP4.5 how the evaluation of the effects in terms of vehicle data analyses is planned to be made. This plan forms the needs of the evaluation for the experimental procedure, which, in turn, must be reflected in the data collection. These needs were especially important for the development of the instructions and recommendations in WP4.5 on the baseline and treatment conditions, the quantity of data that should be collected, and for the test environment. WP4.5 provided WP4.7 information about the operations, used there as an input in the definition of relevant driving scenarios, the planning for whether merging of datasets or results coming from different operations can be made, and for identifying relevant performance indicators for answering the research questions on the effects.

- SP2 *Enablers* selects, adapts, and adopts the most advanced technology enablers to make CAD functions able to operate in defragmented ODDs and in various driving scenarios. WP4.5 collaborated with SP2 to understand the technology enablers addressed in Hi-Drive and the performance indicators relevant for them.
- SP3 *Vehicles* collects the variety of ADFs that are enhanced by implementation of enablers and prepares the vehicles and defines the use cases for their testing (see D3.1 *Use cases definition and description* by Bolivinou et al., 2023). In addition to the operation plans, WP4.5 used as input the descriptions provided by SP3 of the AD functions and enablers implemented to the test vehicles, and of the planned use cases for their testing.

Figure 2.1 summarises these interactions.

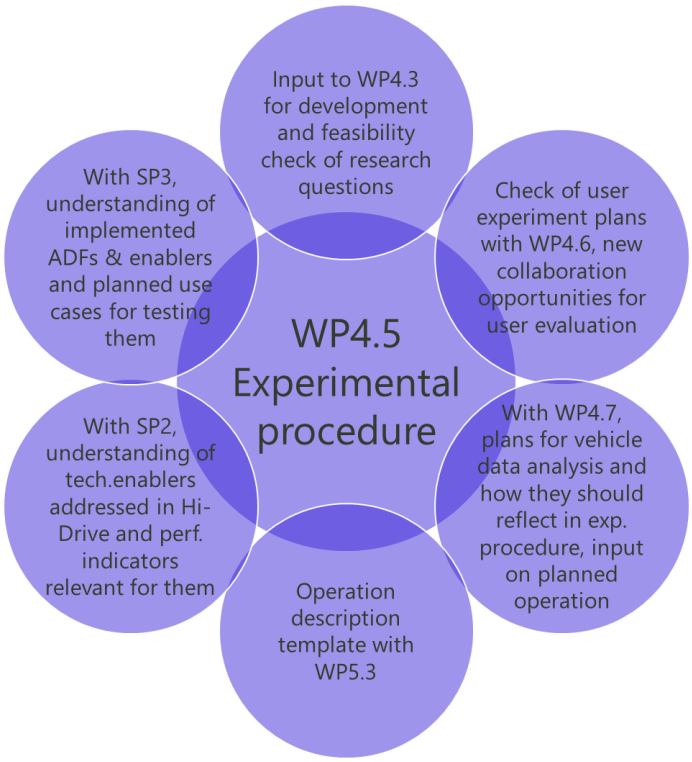


Figure 2.1: Summary of interactions of WP4.5 Experimental procedure with other subprojects and work packages.

Since the Hi-Drive project is closely linked to the L3Pilot project, the guidance for experimental procedure given in L3Pilot was reviewed (original guidelines in Penttinen et al., 2019; revised guidelines in Innamaa et al., 2020). The conclusion was that an extensive review of the possible basic approaches to testing AD functions (from traffic simulation to naturalistic driving studies) should not be repeated in Hi-Drive. The review by L3Pilot was still seen as a valid basis for experimental procedure work in Hi-Drive, to be adapted to reflect the

international landscape of CAV testing (standards by BSI PAS, NHTSA, SAE AVSC). Also, the instructions for the L3Pilot pilot sites were still deemed valid for comparing manual driving with AD in public road experiments.

The technological readiness level of the highly automated driving systems and the safety requirements of public road tests do not yet allow for FOTs with unsupervised ordinary drivers or naturalistic driving studies. Some of the operations in Hi-Drive were similar to the large-scale piloting experiments done in L3Pilot but with different focus. Yet, Hi-Drive also included targeted tests, simulation experiments, and the shadow mode ADF experiments related to the contribution of the technological enablers. Thus, the instructions relevant to ODD extension and robustness testing can be regarded as the novelty outcomes of the work package.

The method for the development of the safety manual is described in Annex 1 with the safety manual itself.

3 Hi-Drive operations from the evaluation viewpoint

3.1 Purpose of operations

Hi-Drive has operations both in SP5 *Operations* and SP6 *Users*. The operations in SP5 serve different purposes in the project by providing data to:

- SP7 *Effects*:
  - Most of the SP5 operations provide data for effects evaluation and contain an experimental setup with baseline and treatment conditions where performance indicators can be compared between them with statistical tests. These operations are the main focus of this deliverable and will be referred to as *technical operations* in the following chapters.
  - Some SP5 operations only collect data about driving scenarios and edge cases for technology development. In these operations, the enablers may not affect the AD functions of the vehicles, or the vehicles may not drive in AD mode. Thus, it is not feasible to use the collected data for comparison between baseline and treatment conditions in effects evaluation. However, the tools and databases related to driving scenario and edge case extraction and storage are under SP7. This data collection is an important part of the overall goal to work towards higher automated driving and is, thus, important for the project.
- SP6 *Users*: Some of the SP5 operations are designed to provide data for user evaluation. Additionally, SP6 makes a significant number of additional user experiments that are not part of the SP5 operation list. All these operations with user focus regardless of the SP in which they take place will be referred to as *user operations* in the following chapters. They will be expanded upon more in Chapter 3.3.
- SP2 *Enablers*: Some of the SP5 operations are intended to study the enablers in highly technical detail, and these topics will not be studied in the SP7 *Effects* evaluation. Therefore, they focus on, for example, signals that are internal to the enablers. These are not part of the common signal list produced by WP4.4 *Data requirements* in D4.2 *Data for evaluation* by Fahrenkrog et al. (2022), which the SP7 *Effects* evaluations are based on. In addition, some of the operations aim to collect data for enabler development. If the enabler is not ready and integrated into the vehicles, it will not affect the AD and consequently is not relevant for the SP7 *Effects* evaluation. Yet, these operations are just as important for the overall goals of the project. The data flow and use of these operations will be described in the deliverables of SP2 *Enablers* and are outside the scope of this deliverable.

User operations were designed directly for collection of data for a certain research question in the user evaluation. The same partners are involved in designing the operation, developing the research question, and evaluating the data. Thus, the experimental procedure was a natural part of the methodology, and the purpose of the operation in relation to the evaluation methodology is apparent. The technical operations do not share an equally direct link, and the partners involved in different parts of the process are often not the same. Therefore, a set of instructions was compiled (presented in Chapter 4) with emphasis on the technical operations, this emphasis being reflected in most of this chapter. However, the instructions should be followed also by user operations where applicable.

In order to understand the task at hand for setting up the requirements for the experimental procedure for the technical operations, it was necessary to know the reason why the operations are conducted. That is, what is the overall goal of the project, how those goals are represented in the challenges selected by the operation owners to be addressed in their operations, and how the success of tackling those challenges could be evaluated. Thus, all available material from the relevant WPs as described in the previous chapter, as well as the Description of Action (DoA), were reviewed.

It should be noted that the operation owners were free to plan their operation from their own perspective in the development of higher-level driving automation. They could select the challenges in the context of ODD extension, AD performance, or the users they wanted to concentrate on, and potential solutions to those challenges. Since these decisions already dictate certain elements of the experimental setup in each operation, a completely harmonised setup is not possible across all operations. However, since data from the technical operations must be evaluated using the common methodological processes developed in SP4 *Methodology*, certain elements do need harmonisation in these operations. Additionally, the DoA has some expectations for the operations and their evaluation. According to it, the operations should contain:

- complex interactions with other road users in normal traffic in motorway and urban environments and in the transitions between motorway and urban;
- challenging and variable weather and traffic scenarios to see the robustness, reliability, and performance of highly automated driving in them;
- testing that reveals information about user preferences and reactions;
- comparisons between highly automated vehicles and human drivers;
- comparisons that allow for evaluation of the enabler's contribution to automated driving;

- testing that allows for studying the time permitted in the ODD and changes in the number of take-overs;
- testing that would lead to chained use cases with multiple integrated enablers that address the use cases and enable long journeys in the ODD.

From the descriptions of the automated vehicles, their ODDs, and the enablers integrated into them, it was evident that no operation would contain all these features listed above. Rather, the objective should be that all features would be included in multiple operations. The main goals of the project were also clear from the DoA: improving AD availability and performance.

In the methodology development, AD availability is divided into two sub-concepts: extension of the ODD and enhancement of AD robustness.

Extension of the ODD refers to allowing the vehicle to drive in AD mode in an environment or a situation where it was not able to drive before implementation of the enabler technologies in the AD system. That is, without the enablers the vehicle would issue a TOR and the human driver would have to continue the driving activities.

In contrast, *enhancement of AD robustness* refers to the elimination of sudden and unexpected TORs that can happen in the nominal ODD, i.e., in the ODD which the AV would have without the integration of technology enablers in the AD system. For example, if the vehicle driving downstream of the AV suddenly brakes, in the Hi-Drive tests the safety driver might take back control of the vehicle. If a hypothetical enabler could inform the AV about the braking manoeuvre sooner than the vehicle could detect it without any enablers, the AV could start braking earlier and more smoothly and avoid harsh braking, issue a TOR, perform a minimum risk manoeuvre, or have the safety driver intervene during testing.

In contrast, the boundaries of ODD extension are likely known in advance. For example, if the vehicle cannot drive inside a long tunnel in full AD mode since it is unable to localise itself without a Global Navigation Satellite System (GNSS) signal, the existence of the tunnel on the road the vehicle is driving on could be assumed to be known from map information. Thus, a hypothetical enabler that allows the vehicle to localise itself inside the tunnel without GNSS would then extend the ODD to potentially include all tunnels within it.

Additionally, a second research area for the vehicle data analysis is the effect of AD on driving behaviour. Driving behaviour can be affected within the nominal ODD, but also in the extended ODD and during take-over situations. For example, the vehicle could drive through the aforementioned tunnel with the support of the enabler—in the sense that the vehicle was in AD mode when it entered the tunnel, throughout the time it was inside the tunnel, and when it exited the tunnel—but at the same time, the AD performance should be evaluated in

terms of, for example, how safely or comfortably the tunnel driving was performed. This evaluation could be done to see what the performance is in terms of comparison with human drivers and whether the AD performance is the same inside the ODD extension as in the same driving scenarios in the nominal ODD by comparing how the vehicle performs in AD mode without the enablers. The latter comparison could be done by comparing how similar the AD performance is in the extended ODD and in the nominal ODD.

These concepts can be seen in the research questions of the effects evaluation in D4.1 *Research questions* by Metz et al. (2023). From the research questions, it is clear that the technical operations should contain data collection both in nominal and extended ODDs, allow comparisons between automated driving supported by the enablers and humans, AD with and without enablers, various challenging environmental and infrastructure conditions, and various driving scenarios within them, and interactions with other road users.

3.2 Summary of technical operations

3.2.1 Method for creating operation summaries

Every experiment has a goal and a set of practical limitations. In Hi-Drive, these individual goals need to be aligned with the overall goals of the project and the limitations of the experiments need to be understood by the methodology team that is setting up the evaluation plan. For many of the operations, the people performing the experiments, setting up the evaluation plan, and later performing the evaluation itself are all separate from each other. Therefore, it was important for the experimental procedure team not only to read all the written information provided by the operation owners, but also to have a series of one-on-one discussions with each of them. These discussions aimed to understand the goals and limitations, as well as their initial plans to reach the goals given their limitations, and to understand the areas where there was flexibility in their plans, in more detail.

In the first round of one-on-one discussions, conducted in July 2022 (project month 13), the operation owners were asked about the following aspects of their operations:

- Description and motivation for the operation regarding the timeline, the use cases or events they saw as the challenges they wanted to solve, the environment or test route, and whether they were planning for large-scale public road tests or more targeted ones
- Initial ideas and possibilities regarding the baseline and treatment conditions
- Initial estimate about how much data could be collected
- Expected size of the effect the operation owners think their enablers could have, as input to a power analysis to determine the amount of data required to show the effect

- The plan for the type and number of participants
- Any open topics the operation owners wanted guidance on.

The answers to these topics were combined with the information provided in the Operation Description tables collected by WP5.3 *Operation preparation*. These formed the basic understanding of the topics, possibilities, and limitations the operation owners had. The conclusions are presented in the following Chapters, 3.2.2–3.2.5.

A second round of discussions was held in December 2022 – January 2023 (project months 18–19) to discuss the instructions given in this report and whether the operations are able to follow them. The outcome of these discussions is presented in Chapter 3.2.6.

3.2.2 Grouping of operations

One of the bigger puzzles the methodology team will be engaged in once the work of this WP is done, is to plan for:

- from which operations the performance indicator data could be pooled[5] before statistical tests are performed,
- from which operations the results of operation-specific statistical tests could be merged, and
- which operations would have to be analysed and reported individually.

One of the keys to this puzzle is the experimental setup in the operations. To help the evaluation methodology team in their work, the operations were grouped based on the experimental setup as follows (Figure 3.1) for the operations suitable for effects evaluation:

[5] By *pooling,* we mean combining datasets to use the combined data to answer certain research questions.

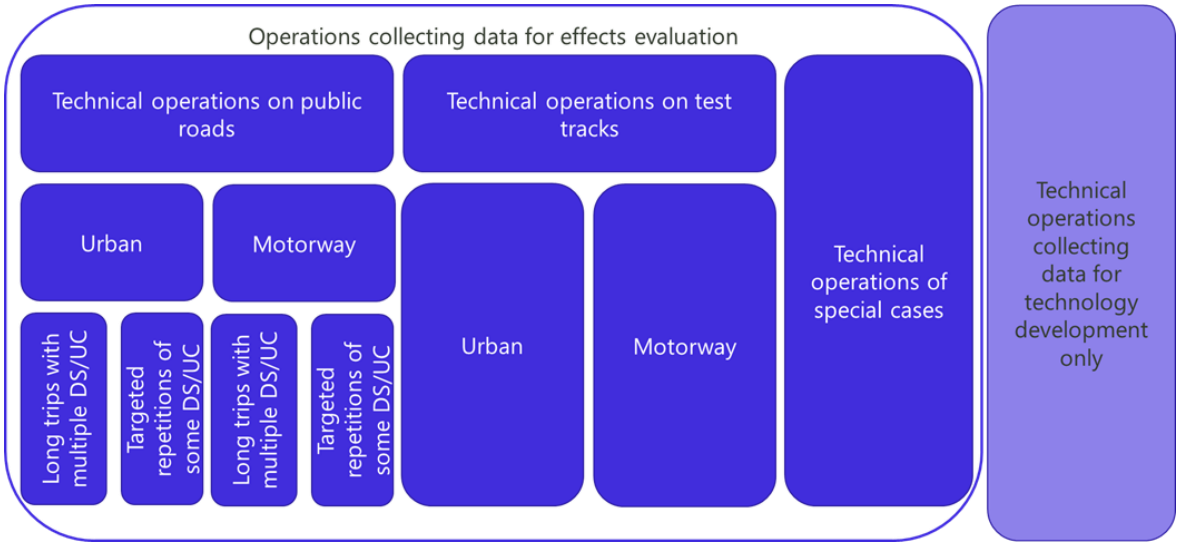


Figure 3.1: Categories in which the operations were grouped , and which ones contribute to the effect evaluation. DS stands for driving scenario and UC for use case.

The operations on public roads were grouped into operations in urban environments and on motorways, and both these groups were divided again into those consisting of longer trips with multiple driving scenarios or use cases (piloting-like) and those consisting of shorter trips with targeted repetitions of some driving scenarios or use cases, with maybe less variability in the naturally occurring conditions. The operations on test tracks were divided into those representing urban roads and those representing motorways.

In the public road testing and test track testing groups, it was more common for the operations to differ substantially than be similar in the challenges they aimed to solve or the enabler technologies they had to solve them. Given the large number of groups and the variety within the groups themselves, it is clear that the instructions and recommendations given would need to be partly agnostic to the use case or enabler present in the operations. The detailed implementation of the instructions and recommendations would be left to the operation owners and the partners dedicated to processing of the data but supported and checked by the experimental procedure work package.

A separate group was made of the so-called *special cases*, which do contribute to the effects evaluation but focus on something so different from the other operations that the dataset or result cannot be merged with the others. These operations are not treated separately from the instructions and recommendations point of view, but they require special consideration by the evaluation team. One of them is the only operation that specifically performs the experiment in snowy conditions, and the rest have baseline and treatment setups where the

desired outcome could mask the effects of the other operations if the data or the results were merged.

For example, in one operation the aim is not to improve performance but to match the AD performance at baseline, with a cheaper sensor setup, to the treatment condition with more expensive sensors. The proposition is that the benefits of delivering highly automated driving to the public come from more accessible enabler technologies in the treatment case. In another operation that is concerned with increasing the robustness of the system to cyber-threats, the desired effect of the system is the generation of a TOR or a minimum risk manoeuvre, both of which would be regarded as ODD fragmentation in other situations. Thus, it is important for the evaluation team to understand where the improvement sought out by the operations could exist and how that interacts with the performance indicators they will design for the evaluation. The designation of some operation into the special case group was seen to highlight the special considerations required.

The operations that are focused only on technology development were grouped separately. The data collected in these operations will be used, for example, as input to simulations supporting technology development, as training data for the development of a machine learning enabler or for identification of driving scenarios and edge cases. While these do not provide both baseline and treatment data for the evaluation of effects, they do play an important role in the development of higher driving automation and, thus, work toward the overall goal of the project.

It was checked that these operations are in line with the overall goals of the project, that there are partners dedicated to the analysis of these data and the related work package (outside evaluation teams in SP6 *Users* and SP7 *Effects*) and the deliverable where their work can be reported. It was deemed that if the operation is purely for technology development, no further recommendations from the experimental procedure team are required, and if the operation only contributes driving scenario or edge-case data, the only recommendation is to follow the data logging requirements of WP4.4 *Data requirements* and the common data format developed by WP5.5 *Data engineering tools and databases*.

3.2.3 Plans for baseline and treatment conditions

The operation owners were asked about their initial plans for baseline and treatment conditions in their operations. All the operation owners that collect data for effects evaluation were planning on collecting both baseline and treatment data. The treatment condition in all these operations was AD with enablers. Thus, the only matter to solve was the kind of baseline(s) data the operation should collect. The initial ideas of the operations were quite divided. Roughly half of the operations were only planning on collecting either manual

baseline data or AD-without-enablers baseline data, and around half were either planning on collecting both or were open to collecting both if asked.

Given this variability in the plans, and since the research questions were set for studying the AD performance also in the extended ODD, there was a strong need for a clarifying recommendation for the baseline conditions.

3.2.4 Input for estimation of the required data quantity

One of the recommendations that needed to be devised was the quantity of data that should be collected by the operations. Since larger effects require less data to show a statistically significant result and smaller effects require a larger number of observations, the operation owners were asked about the expected size of the effect the enablers could have in their operation with regard to some relevant performance indicator. Another way this question was asked was what the operation owners would themselves consider a meaningful result. However, both questions proved to be very difficult for the operation owners at the stage of the project when the discussion was held, as the experiments in Hi-Drive would be the first experiments where the vehicle, enabler, and environment in question would be tested.

One trend did become clear from the answers of the operation owners that did feel comfortable providing an expectation: for the ODD extension the effect would be “quite large” and for AD performance “quite small”. For example, if the vehicle cannot drive inside tunnels in AD mode without enablers, with the enablers the vehicle could manage that successfully most of the time. In contrast, if the improvement provided by the enabler was more about AD performance within the nominal ODD, the improvements predicted were expected to be much smaller. This could be because the vehicles, especially if they are going to be tested on public roads, already have high AD performance. Thus, any potential improvements left within the nominal ODD and with the manoeuvres the vehicle can perform on public roads are quite marginal. This is especially the case on motorways, where there is less environmental complexity.

To know beforehand if there was going to be any mismatch between the data quantity recommendation and the possibilities the operation owners had, they were also asked about how much they foresee they will collect. Apart from a couple of exceptions, the vehicle owners felt that they could collect hundreds of observations. That is, scenarios or trips where a relevant performance indicator could be calculated for both the baseline and the treatment data so that they could be compared.

3.2.5 Plans for participants

The last question posed for the operation owners in the first discussion round was about the participants in the experiment. It was clear that the great majority of operations performing

experiments related to improvements in AD performance or availability were unable to include in their experiment naïve participants, i.e., ordinary drivers without prior experience of AD. The reasons given were related to company policies on permission to drive a prototype AD vehicle on public roads, resources, focus of the operation (e.g., the stage of technological development was ensuring the safe and proper functioning of the solution, not yet studying the user aspects of it) and, therefore, splitting the activities into technical and user operations. In contrast, the great majority of user operations did have naïve participants planned.

A few technical operations were open to include studying the users involved in their experiment but were unable to perform it themselves. These operations were connected with partners from SP6 *Users*, so that their resources could be used to implement and perform the user study part of the operation.

As most of the technical operations did not have naïve participants and only had professional safety drivers, who are part of the safety protocol of the operations and not the subject of experiments, the decision was made not to compile a set of recommendations for the participants of the technical operations. Instead, it was checked that all the operations agreed, in principle, to hand out pre- and post-drive questionnaires to the safety drivers, if concerns related to personally identifiable information were considered.

While most of this chapter consists of reporting the plans for participants in the technical operations, the plans were also reviewed for the user operations detailed in SP5 *Operations*. It was noticed that the plans for the selection of participants and the instructions given to them in these operations were sound, and no further recommendations were required.

3.2.6 Compliance with instructions given

A second round of discussions was held in December 2022 and January 2023 with all 15 technical operation owners to check how well they comply with the instructions presented in Chapter 4 below.

The answers below are given per operation owner and not per operation, as the division of tests among the operations varies greatly. This gives a fairer impression of the activities, as one operation owner could have one large operation with multiple use cases and another operation owner could have multiple smaller operations.

Answers to questions regarding the test environment are shown in Table 3.1. Many of the operation owners include test tracks in their testing activities, but almost all of them were planning on testing on public roads. A couple of operation owners had only test track operations. The reasons given for these were either lack of public communicating infrastructure nearby, that their prototype vehicles did not have the approval to drive on

public roads yet, and that the use case was a cooperative manoeuvre, which is very difficult to perform when surrounded by non-automated and non-connected vehicles.

Table 3.1: Compliance with instructions and recommendations given to the technical operations (as number of operation owners) regarding the test environment (status: December 2022 – January 2023). If the total number of responses in a row does not sum up to the total number of interviewed operation owners, it indicates that some of them were still unsure, or the question was not applicable to their operations.

Question	Yes	No
Will you test on public roads?	12	3
Will you chain multiple use cases?	5	7
If your use case concerns an infrastructural element, are you able to have more than one examples of them?	7	7
Does your test route have interactions with other traffic participants?	15	0
Do your tests have variability w.r.t. time of day, weather, road surface conditions, static/dynamic road infrastructure, etc.?	15	0

Regarding chained use cases, the simplest reason not to chain them was that the operation simply only included one use case. Operation owners that did have multiple use cases were usually planning on trying to chain them. Reasons for not chaining the use cases when multiple ones are targeted were related to either very different environments where the use cases occur or to locations of infrastructure elements or weather conditions. For example, it is difficult to predict construction zones or bad weather beforehand, so planning a route that combines them or other infrastructural elements is not something that they felt they could promise, even if it was a desirable target for them. One operation owner simply had so many use cases that they want to target that they could not promise chaining before first testing all of them independently.

The reasons for not including multiple infrastructural elements were:

- Not applicable to the operation
- The operation owner had only one test track available to them and that test track includes only one example of the element. Or, in the case of one operation owner performing cross-border tests, for testing purposes there was only one suitable border section available.

- Lack of resources, as the specific use case requires temporary sensor setup at each infrastructural element in question.

Answers to questions regarding baseline and treatment conditions are presented in Table 3.2. Manual baseline collection in both nominal and extended ODDs, as well as AD-without-enablers baseline collection in the nominal ODD, are going to be performed by almost all operation owners. One operation owner was not planning on collecting any manual baseline at all, as it was not deemed beneficial for their needs, but promised to check if it could be included for the benefit of the project. One operation owner was not planning on collecting manual baseline in the extended ODD, but this was because their operation is not targeting ODD extension, but rather improving AD performance in the nominal ODD. For each of the questions there were one or two operation owners that still required internal verification regarding baseline data collection, but the need for it was understood.

Table 3.2: Compliance with instructions and recommendations given to the technical operations (as number of operation owners) regarding baseline and treatment conditions (status: December 2022 – January 2023). If the total number of responses in a row does not sum up to the total number of interviewed operation owners, it indicates that some of them were still unsure, or the question was not applicable to their operations.

Question	Yes	No
Can you collect manual driving baseline data in the nominal ODD?	13	1
Can you collect manual driving baseline data in the extended ODD?	12	2
Can you collect AD-without-enablers baseline data in the nominal ODD?	13	0
Can you collect AD-without-enablers baseline data in the extended ODD? (Note that "no" was the expected answer for this question, see explanation below)	6	7

It was expected that the operation owners would respond no to collecting AD-without-enablers baseline data in the extended ODD, since it is expected that the vehicles cannot drive in automated mode in the extended ODD without enablers. Thus, from an evaluation point of view, if the vehicle can drive in the extended ODD without enablers, the effect the enablers have would be more related to improving AD performance. However, four operation owners said they would collect it. This is likely due to differences in the understanding of ODD extension. Nevertheless, it is an indication that baseline data will be collected in relevant situations and is, thus, a positive sign.

Answers to the questions regarding the amount of data collection are given in Table 3.3. Most of the operation owners could collect hundreds of observations in their operations. Four operation owners mentioned that they could not. One of them felt that a smaller data amount was sufficient for the objectives of the operation owner, but they mentioned that they will try to increase the data amount to fulfil the needs of the effects evaluation, even if they could not guarantee it at this point. The other operation owners mentioned limited time on the test track, the resources being split between so many use cases, the additional travel required between the multiple different kinds of infrastructural elements they are targeting in various conditions, and how that will impact the resources they have available for the operation. Two operation owners were still unsure how much data they could collect but are aware of the instruction.

Table 3.3: Compliance with instructions and recommendations given to the technical operations (as number of operation owners) regarding amount of data collection (status: December 2022 – January 2023). If the total number of responses in a row does not sum up to the total number of interviewed operation owners, it indicates that some of them were still unsure, or the question was not applicable to their operations.

Question	Yes	No
Are you able to collect data for hundreds of observations?	9	4
Are you able to collect data so that 50% of the observations are in treatment and 50% are equally split between the baseline conditions?	14	1

Only one operation owner did not see it as a realistic goal to collect 50% in treatment and 50% split between the baselines. The reason given was a lack of resources for data collection. Thus, they will emphasise treatment data collection.

Additionally, to support WP4.7 *Methods for effects evaluation* in the planning of data pooling or merging of results, the operation owners were asked about their initial attitudes towards this subject. The answers are presented in Table 3.4.

Table 3.4: Compliance with instructions and recommendations given to the technical operations (as number of operation owners) regarding amount of data collection (status: December 2022 – January 2023). If the total number of responses in a row does not sum up to the total number of interviewed operation owners, it indicates that some of them were still unsure, or the question was not applicable to their operations.

Question	Yes	No
Are you OK with merging of results?	14	1
Are you OK if your operation is analysed and reported without merging of data or results?	8	4

Merging of performance indicator data, as was done in L3Pilot, was assumed to be an acceptable method and none of the operation owners objected to it. All the operation owners, except for one due to benchmarking concerns, were also fine with merging of results, although one agreed only if anonymity could be guaranteed during the integration process and another only if no other options are possible. However, the question was difficult to answer, as all the implications of agreeing were not yet clear due to the work in progress in other WPs. Thus, the answers should be taken as initial attitudes and should be further defined and checked by the WPs at a later stage. Additionally, this question was less relevant for operations that were planning on joint testing with multiple operation owners, as the likelihood of joint analysis and reporting is already very high.

Three operation owners were not agreeable to individual analysis and reporting of their operation, and three were unsure. The reasons given were concerns related to benchmarking and decreased anonymity, even if the name of the operation owner would not be part of the reporting.

Additionally, the operation owners were in general asked about any updates to their plans since the previous discussion round, and about project internal reporting and partner cooperation. All the answers will be provided to WP4.7 *Methods for effects evaluation* so that they can plan the evaluation.

All in all, the operation owners were strongly aligned with the instructions given. Most deviations were either for understandable reasons, given the aims of the operation in question, or related to topics the operation owners will try to aim towards even if they cannot guarantee it at this stage. The detailed answers and observations that should be taken into account when developing the evaluation plan will be provided to WP4.7.

It should be noted that a few answers from some of the operation owners were still "unsure", as they might, for example, require internal checking. They will, however, be provided as is to

WP4.7 so they can be returned to at a later stage. In addition, the operation owners will be contacted again once this deliverable is published so they can review the full set of instructions along with the reasoning behind them, as well as the safety manual.

3.3 Summary of user operations

Those conducting the technical operations (in SP5 *Operations*) are instructed to collect data using a common questionnaire. Some SP5 operations are user-centred and aim at collecting data for understanding driver comfort, investigating pedestrian-CAD interactions, and driver monitoring.

Studies in SP6 *Users* will use various methods such as driving/pedestrian simulators, test tracks, Wizard of Oz vehicles, virtual reality headsets, interviews, and global surveys, providing Hi-Drive with qualitative and quantitative data to understand users' expectations, behaviours, and the limitations of the systems. Metrics that will be used in the user evaluation include behavioural/performance-based metrics (i.e., take-over time, percentage of crossings, steering-wheel angle, head-rotation movement), physiological measures (i.e., skin conductance, heart-rate monitoring, eye tracking), and subjective measures (i.e., perceived safety, acceptance, trust, misery scale, NASA-TLX). These studies will answer project-level research questions (see Deliverable D4.1 *Research questions* by Metz et al. 2023), which have been carefully linked to scenarios and use cases in which SP5 operations will be conducted. This allows us to draw conclusions and interpretations from various perspectives, providing a thorough understanding of each research question.

User studies address acceptance and awareness, human-like driving and comfort, user monitoring and related HMI, and interaction with other road users. The experimental procedures and design planned for evaluation of these topics are briefly presented below.

Common questionnaires are developed by WP6.3 *User acceptance and awareness* and instructed to be distributed to participants of all Hi-Drive operations when applicable. Common pre-drive and post-drive questionnaires have been created to assess user acceptance of automated driving systems and related factors. In addition, WP6.3 has also designed a set of a global annual surveys which aim to collect data from 16,000 ordinary drivers across eight countries. Each survey aims to understand drivers' willingness to use the system, their expectations, and acceptance.

- Approach: Survey
- Baseline & Treatment: Not applicable
- Participant type: Ordinary drivers

- Data amount: 16,000 participants across eight countries

In addition, WP6.4 *Human-like driving and user comfort*, which looks at users' comfort and motion sickness, will investigate how doing a non-driving related task (NDRT) affects motion sickness (i.e., comparing different tasks), the effect of motion sickness on take-over performances and acceptance (i.e., with no motion sickness as baseline), and the effect of deceleration profile components on drivers' comfort (i.e., comparing different parameters). Experimental studies usually involve 20–40 participants depending on the study design, and we plan to collect more than 4,000 questionnaire data in this WP.

- Approaches: Questionnaires, test track experiments, Wizard of Oz experiments, online survey, open roads experiment in an industrial area, workshop, interviews, driving simulator.
- Baseline & Treatment: Comparing different non-driving related tasks, comparing situations in which participants felt motion sickness to situations without motion sickness, and comparing different deceleration profiles.
- Participant type: Ordinary drivers balanced for gender and with a wide age range, including older drivers (i.e., 18–65).
- Data amount: 20–40 participants per experimental study, and more than 4,000 respondents to online survey across four countries.

Moreover, WP6.5 *User monitoring and related HMI*, which focuses on driver monitoring, will investigate research questions such as how cognitive distraction can affect transitions of control (i.e., comparing different cognitive distraction levels), how monitoring of where the driver is looking can improve the accuracy of the prediction (i.e., comparing drivers' gaze behaviour), how to improve take-over performance using HMIs by comparing different HMI designs, and how to determine teleoperators' workload by comparing different levels of cognitive distractions. Each experimental study will target and collect data from 30–40 drivers.

- Approaches: Driving simulator, real-road experiments, prototypical remote operation workplace, Wizard of Oz experiment.
- Baseline & Treatment: Comparing different cognitive distraction levels, comparing different HMI designs, comparing drivers' gaze behaviour.
- Participant type: Mix of ordinary drivers and ordinary employee drivers balanced for gender and wide age range, including older drivers (i.e., 18–65).
- Data amount: 30–40 participants per experimental study.

WP6.6 *Interaction with other road users*, which focuses on understanding interactions with other road users, will investigate research questions such as what vehicle movement patterns are used by other drivers when interacting with AD (i.e., comparing different vehicle movements), how external HMIs (eHMIs) are evaluated by surrounding drivers (i.e., comparing different eHMIs or no eHMI) and the impact of types and locations of eHMI (i.e., comparing different locations of eHMI, or with no eHMI). Experimental studies conducted in this WP target between 20 and 50 participants each, with a wide age range.

- Approaches: Driving simulator, pedestrian simulator, test track experiment, naturalistic data from intersections, prototypical remote operation workplace, online study, interview on open road, VR headset.
- Baseline & Treatment: Comparing different vehicle movements, comparing different eHMIs or no eHMI, comparing different locations of eHMIs, comparing daytime and night-time, comparing elderly and young pedestrians, comparing different remote operating scenarios, comparing different sensor types.
- Participant type: Mix of ordinary drivers and pedestrians balanced for gender and wide age range, including older drivers (i.e., from 18 to above 60).
- Data amount: 20 to 50 participants per experimental study.

It is important to note that the above information reflects the status of study plans in December 2022 (Project month 18); more detailed information on SP6's methodologies and evaluation plans will be documented in Deliverable D4.4 *User evaluation methods*.

3.4 Feasibility check of research questions from an experimental procedure viewpoint

The work of SP4 *Methodology* includes the definition of research questions for Hi-Drive (see D4.1 *Research questions* by Metz et al. 2023). This task started early in the project and a first draft list of research questions was available by project month 6. Then, feedback on this draft version of research questions was collected from various subprojects and work packages to work towards the final list of core research questions of interest. As a final step, a feasibility check was conducted. The aim of the feasibility check was to ensure that the defined research questions are within the scope of the work planned within Hi-Drive and that the planned data collection is likely to produce the input needed for answering the research questions.

One important source of information in this process was the information provided in this deliverable on Hi-Drive operations and their plans for data collection. The detailed information collected by SP5 *Operations* was sorted and classified with regard to methodological considerations by WP4.5. This was highly relevant for the feasibility check

and was taken into account when evaluating whether research questions can likely be answered within Hi-Drive.

The overview classified the operations by type of enabler, focus of testing, type of planned baseline, etc. (see Chapter 3.2.2). With this information it could be validated that for all finally reported research questions there will be operations contributing data for the analysis. What this analysis looks like in the end will be defined in D4.5 *Effects evaluation methods* and D4.4 *User evaluation methods*. The methodological requirements behind the instructions given in this deliverable are such that the likelihood of finding effects in the planned analysis is maximised. Here, especially the definition of the correct baseline and the requirements regarding quantity of data to be collected are most relevant. Both aim at collecting data in a way that the statistical power of the planned analysis is sufficient to find conclusive answers for the defined research questions.

4 Instructions for Hi-Drive operations

4.1 Implementation of the instructions in practice

A harmonised set of instructions has been provided for all Hi-Drive operations. The instructions and recommendations presented below are mostly for the technical operations to ensure that the data collection is performed in a way that the evaluation team can answer the research questions set for Hi-Drive. Some of the instructions are also applicable to certain user operations and those, too, should comply with the instructions.

Given the significant variability in the focus of the technical operations and in the enablers and use cases present in them, the detailed implementation of the instructions in practice is a task that the operation owners and partners responsible for data processing need to fine tune for their exact operation.

It is acknowledged that due to the variety of studies that will be conducted in SP6 *Users*, which aim to answer very distinctive research questions, there are no 'one-size fits all' instructions that can be given across all user experiments. However, the quality of the study design will be ensured through sharing of information and knowledge in dedicated SP6 meetings to allow peer review and internal discussion before the study is conducted. Similarly, evaluation methods and findings will be shared to ensure an accurate interpretation of each study.

4.2 Test environment

4.2.1 Introduction

The testing of AD systems through diverse and complementary approaches has recently attracted a lot of attention. Various new AD-related standards and regulations are currently being developed worldwide. From the CAD perspective, however, although ETSI reports and plug tests have documented the trialling of C-ITS services, unfortunately no testing guidelines for cooperative AD functions exist today. A first effort to standardise terms and definitions for cooperative automation and its components was made by SAE J3216 (2020), and will continue in order to form the first connected and automated test scenario library. In Europe, the outputs of concluded EU projects, namely MAVEN (see e.g., Rondinone et al., 2018), Enable-S3 (see e.g., Barbier et al., 2019), ICT4CART (see e.g., Pacella et al., 2021), and 5GMobix (see e.g., Shi et al., 2021), are considered relevant to the Hi-Drive use cases.

The testing space that the trials should cover is defined by the AD system[6] (ADS) ODD (introduced by SAE J3016), while the concepts of the operational domain (OD, testing space) and current operational domain (representing the testing space in a specific point in time) describe the inherent gap that exists between the ODD, that describes which conditions the ADS is capable of handling, and the OD that occurs during the ADS testing and deployment. In other words, during testing, the physical infrastructure, environment, and other road users' characteristics of the ODD should be always compared against the physical infrastructure, environment, and occurring road users of the test route.

In Hi-Drive, the ODD of the ADF instance under test is an important aspect of the experimental design for one additional reason: most of the Hi-Drive operations are targeting challenging OD conditions (like road construction, roadway hazards, or tunnels) and OD transitions: for example, an ADS may be operating within its ODD (e.g., two-carriageway motorway, good weather) but then encounter a roadway configuration with lane marking degradation or ambient glare conditions; the resulting OD conditions may exceed the performance limitations of the vision-based sensor of this ADS.

Testing scenarios should, for example, explore the occurrence of conditions that would exceed the nominal ODD to conditions not covered by the ODD. More specifically, the EU regulation[7] splits the nominal scenarios for AD testing into those where the dynamic driving task (DDT), as defined in SAE J3016, is performed under nominal traffic scenarios or those where the DDT is performed under ODD boundaries.

A recently released report by NHTSA applied the new ISO SOTIF standard to a Lane Keeping/Lane Change ADS (Becker et al., 2020) and considered the following types of ODD/scenario variables (from a safety-analysis perspective):

- Permanent variables: These variables are temporally persistent. Examples include roadway functional class, lane type, and permitted types of non-vehicle uses for regional ones, and curves, hills, bridges, and intersections for local ones. Especially in the case of planned test routes, it is likely that these can be known beforehand when deciding the test route.
- Temporary variables: These variables represent events or conditions that are not temporally persistent. They can be further divided into:

[6] According to SAE J3016, an automated driving system (ADS) is the hardware and software that are collectively capable of performing the entire DDT on a sustained basis, regardless of whether it is limited to a specific operational design domain (ODD); this term is used specifically to describe a Level 3, 4, or 5 driving automation system.

7 Draft EU Implementing Regulation, Ares(2022)2667391. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A02019R2144-20220706, accessed in Sept 2022.

- Compounding events or conditions: Given a fixed point in space, these are the aspects of the initial scene that might change. They may persist through an entire trip, only for a portion of the trip, or a change between trips. For example, it might start to rain during one trip but not during another, and this rain may be a short rain shower or it might rain for the rest of the trip.
- Non-typical events or conditions: These events or conditions are unexpected or deviations from normal driving situations that the vehicle must respond to. They can be static, as in the case of a stalled vehicle on the road in a location where it should not be, or dynamic, such as other road users behaving in an erratic manner. Some of the events or conditions, especially some weather-related events, can be considered both compounding and non-typical.

The above considerations about the role of different OD attributes during testing inspired some of the instructions provided below.

4.2.2 Instructions for public road and test track environments

This section presents instructions concerning the features of the selected public road test route and preparation of the test track.

AD is significantly affected by the action of other road participants. Even if the vehicle can perform a manoeuvre isolated from other road users on an empty section of the route or on a test track, it will not shed light on the AD performance in real traffic conditions.

Do not test in isolation: It is expected that interaction with other road participants occurs, and the test environment should reflect that.

The best way this can be ensured is by testing on public roads. If this is realised, the operation owner should not deliberately exclude zones with other vehicles, or pedestrians or cyclists in the urban environment. Exceptions to public road testing should only be made for legitimate reasons, for example if the system cannot guarantee the safety of all the participants because it has not been properly tested to date. On test tracks it is recommended to either perform a joint test with other operation owners or to include other vehicles, pedestrians, or cyclists by the operation owner.

Some of the operation owners have multiple use cases their enabler could enhance or because they have multiple enablers integrated into the system. Additionally, the evaluation goals of the project are deemed to benefit from the inclusion of longer trips. These can be implemented if the multiple use cases are chained together into a single trip.

Use chained use cases: All use cases that can be considered by the enablers should be included in the same trip.

On test tracks, this means that when multiple driving scenarios can be tested, it is good to consider testing them in one driving round and not split the tests per scenario. On public roads, the route should be selected in such a way that as many use cases as possible occur inside it. Exceptions to this should be justified.

Various conditions and features of the test environment may occur incidentally during data collection in the operation, or they may be the focus of the operation, but they may also lead to bias or noise in the results. Thus, close attention should be paid by the operation owner to the conditions of the data collection. If the system is expected to handle the various conditions or they are the focus of the operation, the conditions should be varied in the data collection to make sure that all conditions are represented and can be evaluated. If they are not the focus of the operation, the operation owners should still ensure that the conditions that occur during baseline data collection also take place during treatment data collection in an equal way. Thus, the conditions should be either varied if under test or controlled for if not, and the recommendation is:

Introduce or balance test route/environment variability to ensure different times of day, weather conditions (including transitions between them), road surface conditions, static road infrastructure elements, different characteristics of specific dynamic elements of road infrastructure, and different volumes of traffic.

The time of day is mentioned because both lighting and traffic rush hours can affect AD, as they can create situations where the sensors are pushed to their limits, or the increased complexity of the traffic can cause difficulties for the decision-making process of the vehicle. The time of day when talking about rush hours is related to considering different volumes of traffic that may occur in public road tests both in motorway and urban environments.

Similarly, different weather conditions (including adverse weather: heavy rain, snow, etc.) and road surface conditions (e.g., covered by moisture, wet, flooded; or if cold conditions are included, e.g., snowy and icy) can affect the automated driving task. If challenging weather conditions are a focus of the operation, it would be advised to also record data from the transition from nominal weather conditions to the extended ODD weather conditions (e.g., from cloudy to rain). This can best be realised with longer recordings of data that include the whole transition.

The test route should also be considered from the point of view of the static and dynamic road infrastructure—for example, on motorways the different geometrical characteristics of on- and off-ramp segments and tunnels, and varied messaging frequency for dynamic V2X signage for connectivity operation. Examples in urban environments include different

variations of both signalised and non-signalised intersections and the variety of road furniture and adjacent buildings.

It is generally advisable to start from the simplest set of conditions and move towards more complex ones. Explicitly introducing the variability is especially important in test track studies via different physical infrastructure topologies, if they are the focus of the operation, as they cannot be expected to naturally occur in the same way as in public road tests. However, understandably, there are more practical issues with being able to perform the tests in a varied set of test track layouts. Thus, parameterisation of the test scenario variables (ego vehicle's kinematics, other vehicles' kinematics, hazard distance, ego vehicle's initial lane, ego vehicle's target lane etc.) should be carefully considered for a given test scenario to produce as many meaningful test scenario variations as possible.

The effects evaluation requires information about these conditions for the comparison between baseline and treatment and driving scenario extraction. Thus, either the operation owner or the partner responsible for processing the data should ensure that the information is available. If this information cannot be logged, it should be annotated to the data. Thus:

Route and trip annotation should be performed to capture the conditions at the time of data collection.

Route annotation refers to analysing the route and geocoding all permanent variables of the route before starting the operation. Trip annotation, on the other hand, refers to using a secondary passenger (apart from the safety driver, if possible) who can annotate the temporary events or conditions. In addition to those listed above, this could also concern the annotation of temporary events or conditions to which the system may need to respond, especially if the system did not issue a TOR (a static object on the road ahead, other vehicles disobeying signs or traffic controls, or pedestrians darting out into the roadway).

Additionally, for cooperative AD operations, the types of messages, transmission modes, and targeted CAD cooperation classes should be clearly stated and recorded. In these operations there are also additional experimental variables that should be varied while keeping some variables constant: the collaborative agents' initial speed and relative distance, other road user's location and trajectory, the initial position and size and motion of auxiliary artificial dummies[8] when they are used, the location of roadside units, network delays, and so on.

Furthermore, the operations should follow the guidelines in the safety manual of Annex 1.

[8] https://www.euroncap.com/en/vehicle-safety/the-ratings-explained/vulnerable-road-user-vru-protection/aeb-pedestrian/
https://cdn.euroncap.com/media/58226/euro-ncap-aeb-vru-test-protocol-v303.pdf

4.3 Baseline and treatment conditions

The enabler technologies in Hi-Drive can have three different kinds of effects on automated driving: They can enhance the AD performance, they can extend the AD availability, or they can enhance the AD robustness in the nominal ODD as described in Chapter 3.1. Thus, they have a varying portion of the test route where the effect takes place. These effects should be studied both in terms of how the highly automated driving, extended and/or enhanced by the enabler technologies, is different compared to manual driving and which portion of that effect can be attributed to the enablers. In other words: what is the added effect of AD with enablers compared to AD without these enablers. This increases the number of baseline and treatment conditions compared to previous projects.

The detailed recommendations are given below. The simplified and most applicable instruction is:

The operations should collect the manual driving baseline *and* AD-without-enablers baseline, in both the extended ODD and the nominal ODD whenever possible. The treatment conditions should always be AD with enablers.

If realised with an adequate quantity of data, this ought to give the effects evaluation team the necessary data to answer the research questions.

The manual baseline driving data can be substituted with manual baseline data collected in previous projects if the vehicle, the drivers, the manoeuvres performed by the vehicle, the environment, and the traffic conditions are so similar that they do not cause bias to the results, and if all the signals required for the evaluation in Hi-Drive have been logged.

4.3.1 Enabler can have an effect throughout the test route

Some of the enablers either have an effect on the AD at any point in the nominal ODD or can become active at any point. For example, an enabler that improves localisation or an improved machine learning algorithm would typically always be active and always benefit the AD. An enabler which provides vehicle-to-vehicle (V2V) connectivity, on the other hand, would only provide benefits in the presence of other nearby vehicles with such functionality; only then could the ego vehicle receive and use the V2V information to adapt its driving behaviour in a beneficial way.

If the benefits of an enabler concern *an effect on the AD behaviour* (e.g., changes in speed/acceleration profiles, accuracy of driving within the lane, enhanced object detection or decision making), the operation owner should collect *both manual baseline data and AD-without-enablers baseline data*.

This allows to evaluate the driving performance of highly automated driving compared to humans as well as the added value of the enablers.

If the enabler is aiming to *enhance the robustness of AD* in the nominal ODD (that is, trying to reduce the number of take-overs inside the non-extended ODD), the operation owner should collect *AD-without-enablers baseline data*.

This would make it possible to evaluate TOR situations.

It is also recommended that the operation owner collects *manual baseline data* if the takeover was caused by something in the physical infrastructure.

This would allow for building a more complete understanding of the situation that led to the TOR, since the performance of the base AD is one component of it and allows for the comparison between highly automated and human driving.

For all enablers, the operation owners should ensure that 50% of the data collected is treatment data and 50% is equally split between the baseline conditions.

4.3.2 The enabler has an effect only in specific locations

Some of the enabler technologies present in the operations have an effect on the AD only in a small, specific, sub-section of road. For example, they can allow the vehicle to merge from an on-ramp to a motorway or allow for communication with the traffic lights at a signalised intersection. These effects can be seen as either extending the ODD or improving the AD performance in the nominal ODD. This depends on the nominal ODD of the AD.

If the AD without the enabler was already capable of driving at signalised intersections and the enabler (only) adds another layer of redundancy or improved efficiency, then the enabler can be seen as improving the AD performance in the nominal ODD. If, however, the AD without the enabler was not capable of crossing signalised intersections and due to addition of the enabler it does become possible for the AD to cross signalised intersections, then it can be seen as an extension to the nominal ODD.

For the operations that are extending the nominal ODD, it is recommended to collect baseline data under the condition of *AD without enablers* throughout the whole test route (i.e., in the nominal and extended ODD).

This would allow defining the extension and comparison of the AD performance in the extended and nominal ODDs to understand whether there are any deviations in the AD performance between the nominal and the extended ODD.

In addition, the operation owner should collect *manual driving baseline data*.

This will allow for the evaluation to understand how the AD vehicle performs compared to human drivers.

If the motivation for the operation is to improve the AD performance in the nominal ODD with the help of the enablers, operations should collect both the *AD-without-enablers baseline data* and *manual driving baseline data*.

The first baseline serves the evaluation of the added value of the enablers, and the latter the evaluation of the performance level of the highly automated driving so it can be properly understood compared to human driver performance.

It is highly recommended that the manual baseline data is collected with the same vehicle and on the same route as the other datasets.

Since the section of the test route that is affected by the enabler is smaller than the total length of the test route, the split of the total data collection between treatment and baseline might not make sense in terms of kilometres or hours driven. The reason is that the frequency of the sections relevant for the enabler is not the same for all operations. What matters for the evaluation is how many events at the sections of interest are collected. Therefore, it makes more sense to consider the data collection in terms of events that can be compared between the baseline and treatment conditions. These events are defined by the road section the enabler aims to extend the ODD by. It can be, for example, passing through a signalised intersection, driving through a tunnel, or performing a merging manoeuvre from an on-ramp to a motorway.

The operation owner should make sure that of the comparable events, 50% take place under the treatment conditions and 50% are split between the baseline conditions (e.g., 25% in manual driving and 25% in AD without enablers, if both baselines are possible to perform).

In addition, if the events do not take place when driving in AD mode without enablers in the nominal ODD, the total driving time in the AD-without-enablers condition should be approximately equal to the total driving time of AD with enablers in the extended ODD.

4.4 Required quantity of data

The question of how much data the technical operations need to collect cannot simply be answered in terms of one universal amount of data (hours, kilometres, or number of scenarios or events). In research, a classical way to estimate an answer to this question is to conduct a statistical power analysis. Statistical power refers to the ability of a study to find a significant effect of a manipulation, provided that this effect indeed does exist (Cohen, 1988, 1992). In the context of Hi-Drive, this manipulation can consist of adding an enabler to an AD,

or of going from manual driving to AD. The advantages of conducting a power analysis are that it can reduce the risk of undertaking a study that has no chance of producing results of sufficient certainty. At the same time, it also helps to avoid the use of unnecessary resources for collection of too much data.

The following elements are part of a power analysis (Cohen, 1988, 1992):

1. The level of significance used in the statistical test α. A classical value is α=0.05. Output of the appropriate statistical test is a *p*-value; when *p* is below this threshold, we say that there is a significant effect of our manipulation.
2. The statistical power of the study. A classical value is to aim for a level of 0.8.
3. The sample size (*n*).
4. The effect size (*ES*).

The effect size is initially expressed in the same units as the associated performance indicator. However, a standard step in power analysis is to transform the 'raw' effect size into a standardised effect size. Note that the real effect size of a study is unknown. In the power analysis, the value used for *ES* is the (minimum) effect size that we want to be able to detect (if it exists).

There are trade-offs among the four elements listed above.

- The smaller the effect size, the larger the required sample size (given alpha and the required power).
- If we choose a larger value for α (e.g., use 0.1 instead of 0.05), the power of the study will increase (given a fixed choice for the *ES* and the sample size).
- If we choose a larger value for α (e.g., use 0.1 instead of 0.05), a smaller required sample size will result (given the *ES* and the required statistical power).

Another choice to be made is what statistical test is used to compare the conditions, and if this test is to be conducted as a one-sided or two-sided test. This choice is related to how the hypothesis is formulated. A two-sided test is appropriate when testing a nondirectional hypothesis: H0 states 'there is no effect' and H1 states 'there is an effect' (Aron & Aron, 2003). In the context of Hi-Drive, it may well be that the hypothesis is directional because we are expecting an effect (of the ADF or the enabler) in a certain direction. In this case, one-sided tests are appropriate. The zero hypothesis H0 states that ''the effect is less than or equal to zero" and the alternative H1 states that "the effect is larger than zero". As Annex 2 shows, one-sided tests have more statistical power than two-sided tests. Or in other words, fewer observations are required for one-sided tests.

Annex 2 further details several types of power analysis that can offer guidance for selecting a suitable sample size. It provides information about the amount of data needed for specific types of research questions. These types are stationary driving (e.g., reduction in average speed during free driving), percentages of occurrences (e.g., what proportion of attempts of a certain scenario or manoeuvre was successfully attempted), frequency of events (e.g., does the ADF reduce the frequency of take-overs), and duration in time (e.g., the length of time the AD was active).

As a result of the power analysis, the following generic instruction is given for the required quantity of data:

As an overall order of magnitude, the operations should collect "at least several hundreds" of observations of interest.

Here, an observation can be any scenario or trip for which a performance indicator can be calculated for comparison between baseline and treatment.

4.5 Selection of participants

The user operations and the technical operations that have any kind of user related aspect studied in them (e.g., filling in questionnaires regarding their experiences or having a driver monitoring system) should consider the selection of test participants in line with the guidelines below.

The operations should use the best class of test participants allowed by their company rules, country legislation, ethical aspects, and other limiting factors. The preference of test participant classes is in the order of representativeness of real life: (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.

Suppose externals or employees of the OEM (without specific training) are not allowed to drive the test vehicle. In that case, they are recommended to 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 the ADFs in use).

All test participants should regularly drive (in their daily life). If ordinary drivers are used, it is recommended that the demographic factors reflect the driver population of the future customer or user base.

For instance, include all age groups, also young (<25) and older (60+) drivers, and preferably both male and female participants in all age groups in line with their proportion among the overall driver population.

4.6 Instructions related to the common questionnaire

The overall aim of SP6 *Users* is to evaluate CAD from the users' perspective. This includes the onboard users', external road users', and teleoperators' perspectives. Common questionnaires will be distributed to participants of all the operations testing the ADF. Common pre-drive and post-drive questionnaires have been created to assess user acceptance of automated driving systems and related factors in WP6.3 *User acceptance and awareness*.

The pre-drive questionnaire should be administered before the respondent experiences the tested ADS. This applies especially to participants who are 'ordinary drivers'. If the respondent is already familiar with the tested system (e.g., a safety driver), the questionnaire can be filled in before the data collection begins, whereas the post-drive questionnaire should be administered after the respondent has experienced the tested ADS. If the respondent experiences the ADS over multiple drives, the post-drive questionnaire should be filled in after the last drive. Alternatively, the post-drive questionnaire can be filled in multiple times by the same respondent.

The questionnaires have been designed specifically for those SP5 operations which test ADS. The respondents can be either a driver or a passenger in the vehicle.

Printed copies can be used to collect the answers, if preferred, over an online implementation. The responses are uploaded to the corresponding Hi-Drive database set for subjective data.

The questionnaires have elements that apply to both user studies and technical operations. Thus, they should make sure to use the same phrasing in the questions. The recommendation, therefore, is:

User studies should use the applicable parts of the same questionnaires as the technical operations.

The questionnaires mentioned do not cover all the data needs of the various studies. Each study can also use additional questionnaires. The other work packages in SP6 will actively exchange information regarding the questionnaires they are planning to use to facilitate the use of similar questionnaires within the project.

Questionnaires collected during SP5 operations will be amalgamated and evaluated by WP6.3 *User awareness and acceptance* to provide an overview of drivers' willingness to use the

system, acceptance, mobility impact, system's performances, and so on. Similarly, a common set of the questionnaire will also be used to investigate the interactions between the system and other road users.

5 Conclusions and outlook

This deliverable covers the work of WP4.5 *Experimental procedure*. The goals of this work package were to define how the experiments conducted in the later stages of the Hi-Drive project, especially in SP5 *Operations*, should be performed. The outcome of WP4.5, and thus the content of this deliverable, are summaries of the plans of the Hi-Drive operations and instruction and recommendations on the operations' procedures. For all the following points, there was close interaction with the other relevant work packages across various subprojects.

For the first goal, the summary of operations, the amount and diversity of operations planned in the Hi-Drive project posed a challenge. It was overcome by a structured approach based on multiple resources. As detailed in Chapter 2, one resource was operation description templates filled in by the operation owners. Another resource was bilateral discussions between WP4.5 and the operation owners. From the information, the operations were categorised based on the main characteristics and summarised. An overview was also made of user operations grouped according to research topic.

The second goal, the formulation of recommendations for the experimental procedures, has been worked out. The recommendations focus on what amount of, and how, data should be collected to be able to assess the effects defined by the research questions. A challenge in this regard was the variation of enabler technologies set to be developed and tested in the Hi-Drive project. It was especially challenging from a methodological standpoint that some of the enabler technologies aim at extending the ODD of AD while others enhance the robustness of AD in terms of decreasing the need for take-overs or increasing AD performance. The high number of different kinds of operation setups also increased the difficulty of determining how much data must be collected per operation or per topic of interest. The formulation of the instructions and recommendations was achieved by leveraging the gathered information while compiling the summaries and in close cooperation with the other methodology work packages.

This deliverable serves as a direct input to the operation owners in SP5 *Operations* and SP6 *Users*. As this deliverable is one of the three deliverables in SP4 *Methodology* that define basic requirements and goals of data collection from a methodological point of view, it also serves as input to the open work packages in SP4. Specifically, the other two deliverables, in addition to this one, are the deliverables on research questions (D4.1 *Research questions* by Metz et al. 2023) and data requirements (D4.2 *Data for evaluation* by Fahrenkrog et al. 2022). Using these three deliverables as a basis, detailed analysis plans for user evaluation (D4.4 *User evaluation methods*) and effects evaluation (D4.5 *Effects evaluation methods*) will be developed.

In WP4.7 *Methods for effects evaluation*, the content and lessons learned from this deliverable will be used in various tasks. One of them is the driving scenario definition task, for which the summary and grouping of operations included in this deliverable are especially helpful. Another task of WP4.7 that will benefit from this is the definition of relevant performance indicators for each research question and being able to link the operations with driving scenarios, research questions, performance indicators, and ultimately effects.

Due to the sheer variety of operations, not all data from all operations can be pooled together to answer each research question. Therefore, grouping of operations is required to find the candidates for meaningful pooling of data or for merging of results. Among these groups, different analysis methods might be needed to assess the effects of interest. The grouping of operations presented in this deliverable is already a basis for consideration of this operation pooling. However, the final version of the pooling depends on the definition of the driving scenarios, the collection of relevant performance indicators, and analysis concepts intended for the evaluation of effects. These points are part of WP4.7 *Methods for effects evaluation*. Thus, WP4.7 will use the grouping of this deliverable and might refine it based on the aforementioned factors.

References

Aron, A. & Aron, E. N. (2003). *Statistics for Psychology*. 3rd edition. Prentice Hall.

Cohen, J. (1988). *Statistical power analysis for the behavioural sciences*. Lawrence Erlbaum Associates.

AVSC00002202004, Automated Vehicle Safety Consortium (2020). AVSC Best Practice for Describing an Operational Design Domain: Conceptual Framework and Lexicon. *SAE Industry Technologies Consortia*.

Barbier, M., et al. (2019). Validation of Perception and Decision-Making Systems for Autonomous Driving via Statistical Model Checking. *IEEE Intelligent Vehicles Symposium (IV)*, Paris, France, 2019, pp. 252-259, doi: 10.1109/IVS.2019.8813793.

Becker, C. J., Brewer, J. C., & Yount, L. J. (2020). Safety of the intended functionality of lane-centering and lane-changing maneuvers of a generic level 3 highway chauffeur system (Report No. DOT HS 812 879). National Highway Traffic Safety Administration.

BPI PAS 1883:2020: Operational Design Domain (ODD) taxonomy for an automated driving system (ADS) – Specification

Bolovinou et al. (2023). Hi-Drive Deliverable D3.1: Use cases definition and description.

Cao, Y. et al. (2022). L3Pilot - Code of Practice for the Development of Automated Driving Functions. Available at: https://l3pilot.eu/fileadmin/user_upload/Downloads/Project_Information/L3Pilot_CoP-ADF_v1.0.pdf [Date Accessed: 13.1.2023]

Cohen, J. (1992). A power primer. Psychological Bulletin 112 (1), 155-159.

Fahrenkrog et al. (2022). Hi-Drive Deliverable D4.2: Data for evaluation.

FOT-Net, CARTRE & ARCADE (2021). FESTA Handbook, Version 8, September 2021. 227 p. https://www.connectedautomateddriving.eu/wp-content/uploads/2021/09/FESTA-Handbook-Version-8.pdf

Innamaa, S., Aittoniemi, E., Bjorvatn, A., Fahrenkrog, F., Gwehenberger, J., Lehtonen, E., Louw, T., Malin, F., Penttinen, M., Schindhelm, R., Silla, A., Weber, H., Borrack, M., Di Lillo, L., Merat, N., Metz, B., Page, Y., Shi, E. & Sintonen, H. (2020). L3Pilot Deliverable D3.4: Evaluation Plan. <https://l3pilot.eu/fileadmin/user_upload/Downloads/Deliverables/Update_28042021/L3Pilot-SP3-D3.4-Evaluation_plan-v1.0_for_website.pdf>

ISO 21448:2022. *Road vehicles — Safety of the intended functionality*.

L3Pilot consortium (2021). Final Project Results. L3Pilot Deliverable D1.7. 162 p. https://l3pilot.eu/fileadmin/user_upload/Downloads/Deliverables/Update_10082022/L3Pilot-SP1-D1.7-Final_project_results-v1.0_for_website.pdf

Metz et al. (2023). Hi-Drive Deliverable D4.1: Research questions.

Pacella, F., Bonetto, E., Castillo, G. A.G., Brevi, D., Scopigno, R. (2021). Implementation and Latency Assessment of a Prototype for C-ITS Collective Perception. *IEEE International Mediterranean Conference on Communications and Networking (MeditCom)*, Athens, Greece, 2021, pp. 100-105, doi: 10.1109/MeditCom49071.2021.9647572.

Palin, R., Ward, D., Habli, I., & Rivett, R. (2011). ISO 26262 safety cases: Compliance and assurance.

Penttinen, M., Rämä, P., Dotzauer, M., Hibberd, D., Innamaa, S., Louw, T., Streubel, T., Metz, B., Wörle, J., Brouwer, R., Rösener, C. & Weber, H. (2019), Experimental procedure: Deliverable D3.2 of L3Pilot. https://l3pilot.eu/fileadmin/user_upload/Downloads/Deliverables/Update_07102021/L3Pilot-SP3-D3.2-Experimental_procedure-v1.0_for_website.pdf

Rondinone, M., Walter, T., Blokpoel, R., Schindler, J. (2018). *V2X communications for infrastructure-assisted automated driving*. In Proceedings: IEEE International Symposium on" A World of Wireless, Mobile and Multimedia Networks" (WoWMoM), 12 June 2018 Chania, Greece

SAE J3016:2021. *Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles*.

SAE J3018:2020. *Safety-Relevant Guidance for On-Road Testing of Prototype Automated Driving System (ADS)-Operated Vehicles.*

SAE J3216:2021. *Taxonomy and Definitions for Terms Related to Cooperative Driving Automation for On-Road Motor Vehicles.*

Sauvaget et al. (2022). Hi-Drive Deliverable D5.1: Description of "Operations".

Shi. Y., Yu, H., Guo, Y., Yuan, Z. (2021). *A Collaborative Merging Strategy with Lane Changing in Multilane Freeway On-Ramp Area with V2X Network*. Future Internet. 13(5):123. https://doi.org/10.3390/fi13050123

Thorn, E., Kimmel, S. C., Chaka, M., & Hamilton, B. A. (2018). *A framework for automated driving system testable cases and scenarios (No. DOT HS 812 623)*. United States. Department of Transportation. National Highway Traffic Safety Administration.

Toyota (2020). *Toyota Voluntary Safety Self-assessment for SAE Level 4 and Level 5 Automated Vehicle Technology Testing on Public Roads.* Available at: https://amrd.toyota.com/app/uploads/2022/03/VSSA.pdf (Accessed: 9.1.2023).

Weber, H., Hiller, J., Eckstein, L., Metz, B., Landau, A., Lee, Y. M., Louw, T., Hogema, J., van Weperen, M., Lehtonen, E., Sintonen, H., Streubel, T., Svanberg, E., Bolovinou, A., Rigos, A., Andreone, L., Bellotti, F. & Zlocki, A. (2022). L3Pilot Open Data. Retrieved from https://doi.org/10.5281/zenodo.5874765. [Date accessed: 12-12-2022].

List of abbreviations and acronyms

Abbreviation	Meaning
AD	Automated driving
ADAS	Advanced driver assistance system
ADF	Automated driving function
ADS	Automated driving system
AV	Automated vehicle
AVSC	Automated Vehicle Safety Consortium
BSI	British Standards Institution
CAD	Connected and automated driving
CAV	Connected and automated vehicle
C-ITS	Cooperative intelligent transport systems
CoP	Code of Practice
DoA	Description of Action
ETSI	European Telecommunications Standards Institute
eHMI	External human-machine interface, i.e., HMI outside the vehicle
EU	European Union
FESTA	Field opErational teSt supporT Action
FOT	Field operational test
GNSS	Global Navigation Satellite System
GPS	Global Positioning System
HAD	Highly automated driving
HMI	Human-machine interface
IFTD	In-vehicle fallback test driver
ISO	International Organization for Standardization
L3	SAE level 3 (driving automation)
NDRT	Non-driving related task
NHTSA	National Highway Traffic Safety Administration
OD	Operational domain
OEM	Original equipment manufacturer
ODD	Operational design domain
PAS	Publicly available specification

Abbreviation	Meaning
PI	Performance indicator
SAE	Society of Automotive Engineers
SDV	Software defined vehicles
SOTIF	Safety of the intended functionality
SP	Subproject
TOR	Take-over request
V2I	Vehicle-to-infrastructure (communication)
V2V	Vehicle-to-vehicle (communication)
V2X	Vehicle-to-everything (communication)
VRU	Vulnerable road users
WP	Work package

Annex 1 Safety Manual

Background

The Safety Manual deals with one of the WP4.5 *Experimental Procedure*'s objectives, in particular its aim is to develop and instruct on strict safety procedures for the road tests. This should address the fundamental question: "What is the best practice for testing automated driving (AD) on public roads in the EU"? Under this perspective, the goal of the Safety Manual is to offer guidance on effective safety procedures for the road tests.

A specific Safety Manual Working Group was set up, including several members from academic and research institutes and original equipment manufacturers (OEMs) such as CRF, FKA, TME, BMW, and VTT. The starting point was the experience of these partners from L3Pilot and the best practices of the involved OEMs.

Therefore, this annex provides the steps needed so that a system is acceptably safe for the technology under test in the applicable ODD. For other safety related aspects, L3Pilot (Cao et al., 2022) and Hi-Drive Code of Practice (CoP) documents may be referred to.

Methodology for the Safety Manual

In this section, we describe the methodology adopted for the creation of the Safety Manual, remembering that it gives guidance on strict safety procedures for the road tests. The starting points for the Safety Manual are listed below:

- Information from OEMs about their procedures, especially those used in the previous project, L3Pilot.
- ISO26262 (Functional Safety), ISO21448 (SOTIF).
- Research questions and details on the operation plans from SP4 *Methodology* and SP5 *Operations*, respectively.

Considering the last point, Figure A1.1 illustrates how the Safety Manual interacts with other subprojects and WPs.

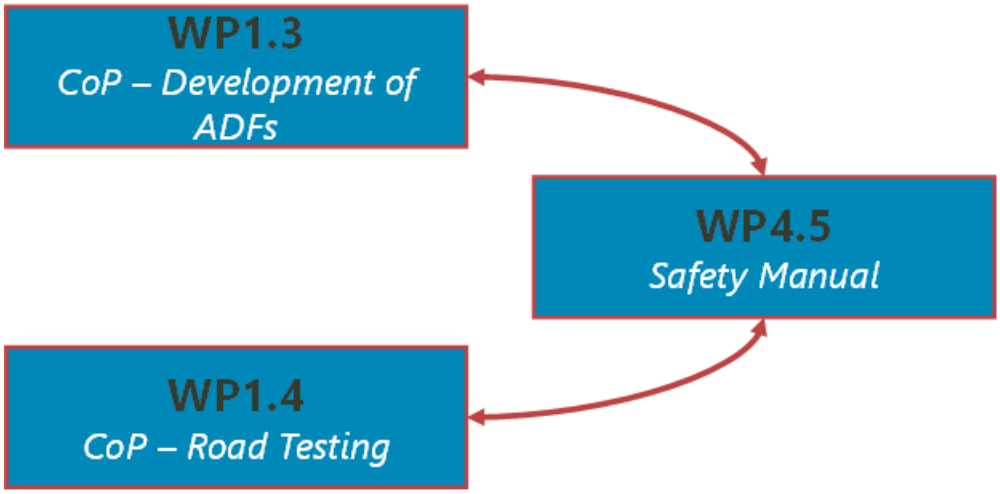


Figure A1.1: interaction of safety manual with other SPs and WPs.

The most relevant interaction is with the development of the CoP, both in WP1.3 for the development of ADFs and in WP1.4, specifically addressing the road testing (even if in a more "procedural and organisational" way, such as how to obtain permissions in a cross-border scenario).

This document aims at providing the necessary steps related to operational safety and covers two main areas:

- *In-vehicle fallback test driver* (IFTD), also called "safety driver".
- *Operational guidelines for testing and piloting* (OGTP).

Both are described in the following sections.

It should be noted that this document does not cover aspects such as product safety, as there is no intended use for consumers nor organisational safety, since this aspect may be specific to each company.

In-Vehicle Fallback Test Driver

One of the key points when testing ADFs with a high level of automation (from SAE-L3 up) is to guarantee safe driving, characterised by the following factors:

- Reliable emergency driving operation (at low or high speeds, depending on the specific function under test)
- Calm and balanced driving in any situation
- High safety awareness in recognising risks.

In this perspective, the role of the IFTD, or "safety driver", is essential to assure safe driving in automated mode and the correct reaction to (possible) misbehaviour. Particular attention by the IFTD should be placed on the following:

- Know the system you are using and its limitations.
- The driver is always right (namely: if the system is not performing within your safety zone, you should take over control immediately).
- Never take AV behaviours for granted (e.g., system and sensor performance can be affected in certain weather conditions).
- In case of unexpected machine behaviour, always abort.
- Strictly respect the traffic code and rules.
- Follow the emergency procedure in case of an incident or accident.

In order to define and correctly select a person for the role of IFTD, the following criteria should be considered:

- For selection
  - Several years of driving experience
  - Tests on proving grounds and real roads
  - A general safety mindset
- For training
  - Basic knowledge of AD technology
  - Ability to handle scenarios on a private test track (e.g., control unstable vehicle in specific AD-related scenarios)
  - On a private test track, fault injection tests are performed to demonstrate that the IFTD has the skill to perform emergency/evasive handling manoeuvres (for example, during normal AD driving a fault in longitudinal or lateral control is injected and the IFTD must react and control the fault)
  - Introduction to the specific AD system (e.g., what are the capabilities and limitations of this specific vehicle/system).
- For assessment
  - Fail/pass criteria
  - Periodic re-training for updates (optional and dependent on practice)

The following sections provide a couple of examples originating from the experiences gained from testing the AD functions in the previous L3Pilot project.

Finally, the IFTD task requires collaboration with SP1 *Collaboration* for the CoP (WP1.3 *CoP for the development of ADFs* and WP1.4 *CoP for road testing*, in particular).

Training of In-Vehicle Fallback Test Driver

After selecting the IFTDs for testing of the early ADF prototypes, training them is crucial. The training is recommended to be designed according to SAE J3018 and "AVSC Best Practice for In-Vehicle Fallback Test Driver Selection, Training, and Oversight Procedures for Automated Vehicles Under Test".

Based on the proposed selection process, it is assumed that only experienced and professional drivers capable of controlling the vehicle in highly dynamic situations will be considered.

As a first step, the IFTDs will attend a classroom session during which they will be familiarised with the ADF and its limitations (e.g., based on sensor performance or algorithm capabilities), ODD, specific regulations that apply, and so on. Furthermore, the IFTD must understand the HMI design and the procedures for engaging and disengaging the ADF. In addition, emergency and evasive handling manoeuvres will be introduced.

In the second step, the IFTD will be trained in a vehicle on a closed private road or proving ground. Here, the IFTD will be familiarised with the pre- and post-trip procedures, the ADS HMI, and the procedures for engaging and disengaging the ADS. Besides normal operation, the IFTD will also be trained in emergency and evasive manoeuvres. It is recommended to perform fault injection here, whereby the ADS is intentionally brought to an unsafe state requiring the intervention of the IFTD. The IFTD must take over control and follow emergency or evasive manoeuvre procedures.

Training in a driving simulator, in addition to the in-vehicle training above, can be beneficial for including high-risk scenarios and fault injections into the training. This training mainly contributes to learning the system limits and correct take-over procedures, as driving simulators are mostly limited in terms of vehicle dynamics. Leaving the ODD (e.g., because of changing weather conditions or temporary road works) can also be done in a driving simulator.

After completing all of the above training steps, the IFTD performs their first drive on public roads. To verify the success of the training, an experienced IFTD rides along with the new IFTD and evaluates and corrects their behaviour.

Through all training segments, the IFTD shall be evaluated and must show the required capabilities. Additionally, after the initial training, the training shall be repeated periodically and the IFTD must be continuously monitored and re-evaluated.

Examples of Definitions for the In-Vehicle Fallback Test Driver

This section covers three examples of IFTD instruction, based on experiences from CRF, BMW, and FKA (mainly from the L3Pilot project).

CRF procedure for definition of the in-vehicle fallback test driver

Two main assumptions are made in this case:

- An Automated Driving (AD) system is defined as any driver assistance system capable of maintaining both vehicle speed and lane position on roadways without direct control inputs by the driver.
- Low-speed functions (such as automated parking) are not considered as AD systems for the purposes of the presented training.

The safety drivers (or IFTDs) of CRF are professional drivers specifically trained for AD and who have undergone dedicated courses. Figure A1.2 shows the procedure.

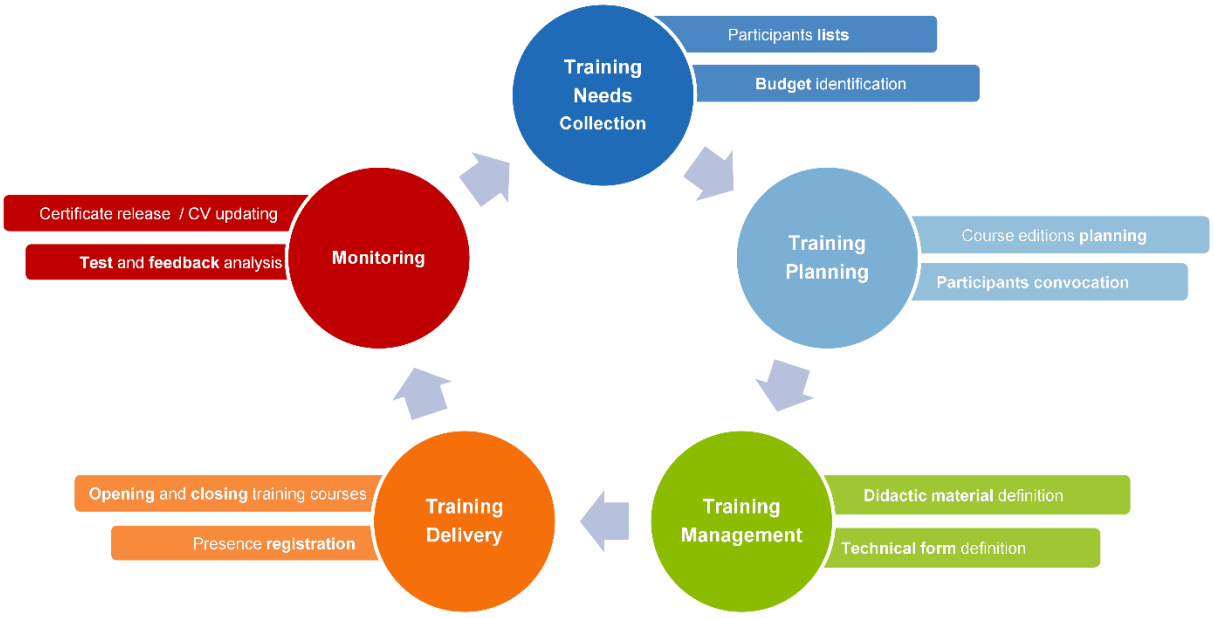


Figure A1.2: Training Management Process in CRF.

This training procedure defines the vehicle operation procedures necessary to that ensure AD development vehicles are in compliance with applicable governmental regulations, in a safe

manner and within their ODD. It also ensures that the safety driver can take manual control of the vehicle and place it in a safe state if it will be necessary during the ODD.

As sketched in the figure, first the needs for training are collected, the contents of the training are planned, and participants are selected. The management phase includes the definition of the didactic material (slides, manuals, etc.) and how the course is delivered. After its completion, there is the monitoring phase, with test of participants and related feedback analysis.

The scheme is presented in Figure A1.3.

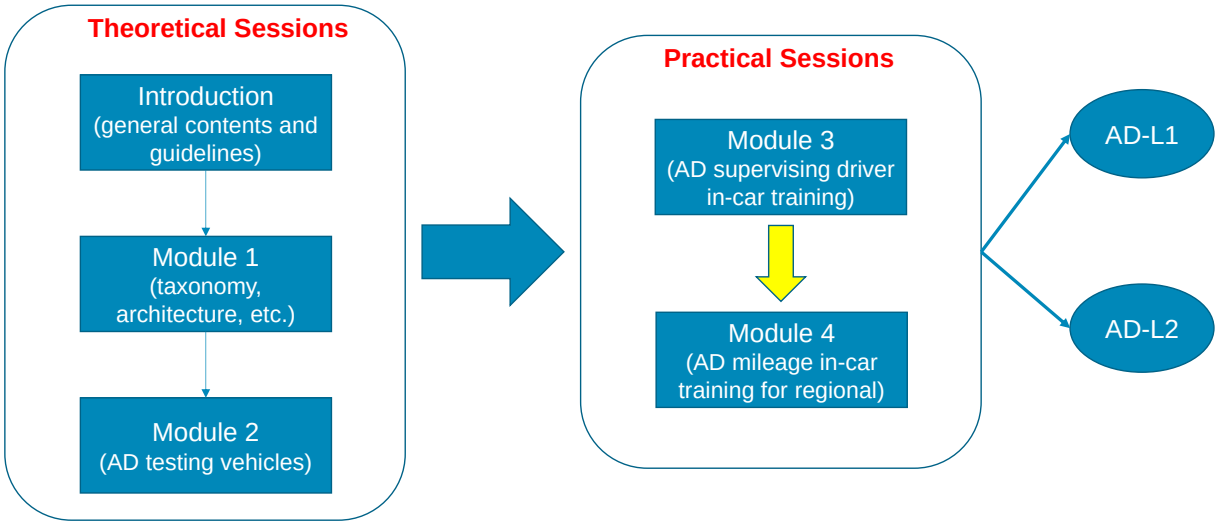


Figure A1.3: Schematic representation of how the training course is structured at CRF, with related sessions.

As represented in the figure, there are two parts: theoretical and practical. The first one, after a general introduction, provides information and details on the use cases and scenarios of ADFs, their architectural schemes, and specific elements of the system under evaluation. The second part deals with the test on the car, first static—more related to HMI topics—and then dynamic (both on test track and on real roads).

The AD Course is structured on two levels of training:

- AD-1: Sufficient for permission to drive both on both private roads and proving grounds.
- AD-2: This level authorises driving AD vehicles on public roads, in case a specific regional regulation requires "extra driving training or mileage". AD-1 is a prerequisite for AD-2 training.

This AD course must be considered specific training for expert drivers so they can be authorised to drive and/or supervise cars equipped with AD features.

Finally, an additional operator (e.g., the experimenter) can also be present on board. This operator provides operational support to the safety driver and/or leads the experiment(s). In this case, there must be a clear separation of task and communication guidelines established between the operator and the safety driver:

- Clear separation of tasks: executing monitoring of the ADS, including annotating logs, monitoring system status, and providing information to the safety driver about system status and planned behaviour. The operator shall also monitor the state of the safety driver (e.g., distractedness, illness/impairment, non-compliance to procedures, etc.)
- Communication guidelines: the operator is communicating verbally with the safety driver so they can be aware of upcoming vehicle behaviours. Additionally, the operator may also support the safety driver with the meaning of visual and audible signals to facilitate its interpretation and anticipate vehicle behaviours.

All in all, the safety of people on board should be assured for everyone (including operator and safety driver).

BMW procedure for definition of the in-vehicle fallback test driver

As part of the Hi-Drive project, BMW focused their operations on user-related testing. The basis for this type of testing is that the driving of the vehicle and the experience of the ADF is done by a study participant (e.g., an ordinary driver) sitting in the driver's seat. This in turn has a significant impact on the tasks as well as the requirements for the definition of the IFTD located in the passenger seat.

The following situations have been defined to clearly determine the seating constellation in regard to the responsibilities in real road traffic:

- Non-Automated Driving: while driving manually or partially automated, the responsibility and driving task lies with the subject in the driver's seat. The IFTD is only monitoring the situation and, if necessary, intervenes in the event of misconduct of the participant or ADS.
- Automated Driving: At the moment of ADS activation, responsibility passes to the IFTD. The IFTD then has the task of monitoring the traffic situation and the ADS behaviour. In the case of ADS malfunction or traffic-based critical situations, the IFTD needs to intervene and bring the situation to a safe and stable state.
- System transition: The activation and deactivation need to be observed and handled by the IFTD. The responsibility between subject and IFTD and vice versa changes only when the IFTD is satisfied that the situation and the ADS are in a safe state and that awareness of the participant is assured.

Based on this definition, BMW elaborated a process to identify organisational, technical, and IFTD-based training requirements. The starting point of the process is the ODD requirements used in a hazard and risk analysis to identify critical key points. From this analysis, the corresponding measures, such as driving school pedals, are derived for longitudinal control since the IFTD is sitting in the passenger seat.

This process is shown in Figure A1.4.

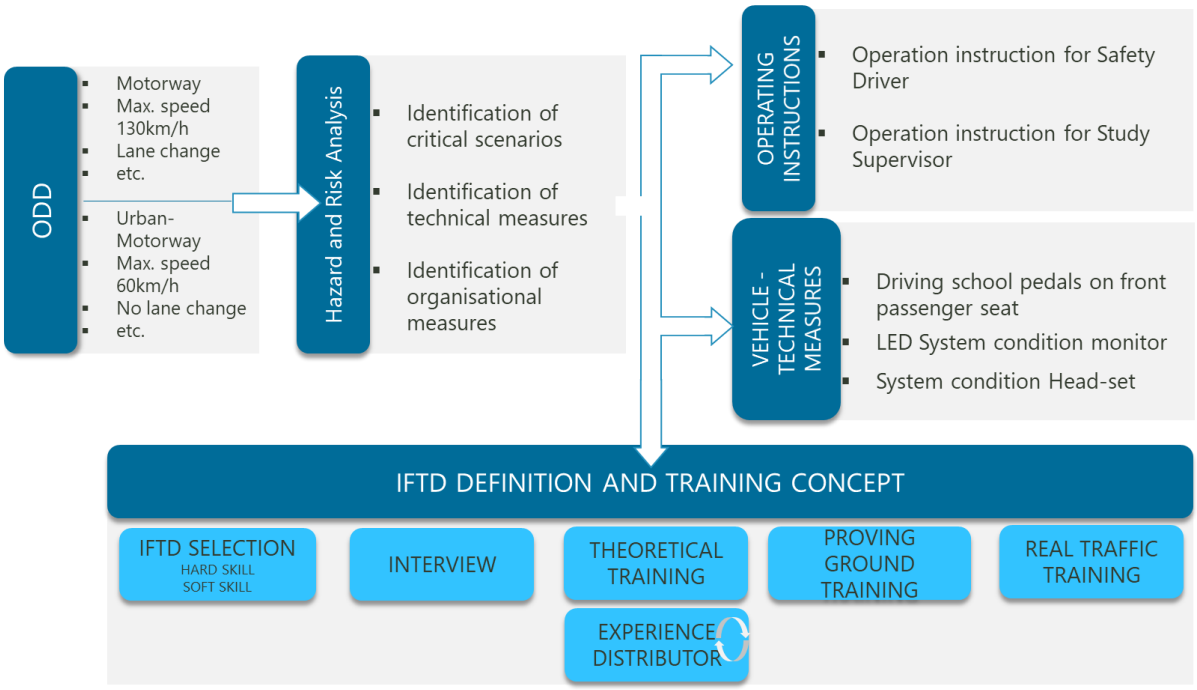


Figure A1.4: BMW process to identify the IFTD training concept.

The process step "IFTD definition and training concept" shown in the figure consists first of the selection of suitable personnel in alignment with the general selection criteria mentioned in the chapter above. BMW extended the requirements by adding its own internal licences for "high dynamic driving" and "handling with prototype vehicles" as needed hard skills. In regard to soft skills, the potential IFTD must have shown a high capacity to focus during long drives and a strong sense of responsibility in their daily vehicle test operation. Especially the latter is confirmed through interviews on whether the potential IFTD has the necessary attitudes towards responsibility and safety awareness.

After passing the above-mentioned selection criteria, the candidate IFTD starts ADS and ODD relevant training, consisting of theoretical, proving ground and real traffic training. Each training module is accompanied by a special instructor who will also rate and finally approve the candidate and issue a certificate. The modules are generated to train and test the

understanding of ADS behaviour and limits and the driving skills needed for special ODD scenarios from the passenger seat. The modules are itemised in Figure A1.5.

Figure A1.5: BMW Training modules.

When seated in the passenger seat, the IFTD needs to demonstrate the ability to control situations on the proving ground even in very challenging scenarios like the ISO (3888-1, 3888-2) double-lane change. Finally, the IFTD is trained in real traffic on public roads, especially in the handling of system failures.

FKA procedure for the definition of the in-vehicle fallback test driver

At FKA, the task of operating a prototype vehicle with an active ADS is separated into two roles: the IFTD in the driver's seat and the operator in the passenger's seat.

The IFTD is responsible for vehicle control and stays responsible while ADS is active. The IFTD must monitor the vehicle movement and the ADS actions. In case the ADS is about to enter an unsafe condition, the IFTD must take over the vehicle control and disengage the ADS. The same applies to any unexpected or inappropriate behaviour or unclear traffic scenarios as well as the violation of any traffic code or ADS-specific regulations. In addition, the IFTD must validate whether the ADS is still within its ODD. Secondary tasks and non-driving related tasks may not be performed during driving. This includes, for example, interaction with tertiary control elements or control of the in-car PC or measurement equipment. Communication with the operator or other passengers shall be limited to the absolute minimum required for safe vehicle operation.

The operator is responsible for operating the ADS, which includes starting and supervising the ADS software, monitoring the ADS' planned actions, and validation of the computer

vision output. In case of any violation of the required ADS behaviour, the operator must inform the driver to disengage the function. Furthermore, the operator is responsible for answering the IFTD's questions and monitoring the IFTD's wellbeing.

At FKA, there are two types of IFTDs: those allowed to drive with active ADS on closed proving grounds and test tracks and those allowed to drive on public roads. Within the scope of this section only the second is considered. Only professional drivers who are trained to control the vehicle at its handling limits are selected. These abilities of a potential IFTD are assessed by a third party. A general health check is required for IFTDs. The IFTD's general safety mindset must be confirmed by their having had the relevant driver's licence for several years and a required minimum score at the German federal "Register of Driver Fitness". Furthermore, the IFTD is required to have a general understanding of ADS and ADAS. Usually, the IFTDs are part of the ADS development team and closely involved in the development.

The IFTD training consists of three elements: classroom training, test track training, and supervised training on public roads.

Classroom training: the potential IFTDs are trained to understand the general ADS and vehicle design and the sensor setup. Especially, the limitations which apply regarding the ADS capabilities are discussed. In the second step, the IFTDs are familiarised with their tasks, responsibilities, and the role of the operator. Pre- and post-trip procedures are presented, while emergency manoeuvres and procedures are trained in theory. The third step includes the regulations which generally apply for usage of ADS on public roads. Finally, a description of the specific pilot site is presented. The IFTDs are trained to understand the ODD, the specific limitations which apply to the pilot site, and challenging scenarios. The IFTD is retrained if there are any changes to the previous content, such as software updates or extensions of the pilot sites.

Test track training: on a closed proving ground, the IFTD is familiarised with pre- and post-drive procedures and the interaction with ADS-specific HMIs and engagement and disengagement of the ADS. Interaction and communication with the operator are also practised. The ODD and system limits are shown. Emergency and evasive manoeuvres and procedures are practised. As the last step of the test track training, the IFTD must demonstrate the correct activation and engagement of the ADF. After a first section of normal driving with activated ADS, fault-injection tests are performed. The instructor will inject faults into the ADS and vehicle. The IFTD must show the capability of identifying the injected fault and reacting correctly within a set amount of time. The performance of emergency manoeuvres and procedures is evaluated by the instructor. An example is shown in Figure A1.6. While driving with activated ADS on a straight section (solid blue line), the instructors inject a fault causing a rapid lateral motion of the vehicle (orange dashed line).

The IFTD must recognise the unsafe behaviour of the vehicle and intervene. In this specific situation, the IFTD must take over the lateral and longitudinal control and prevent the vehicle from leaving its lane (blue dashed line).

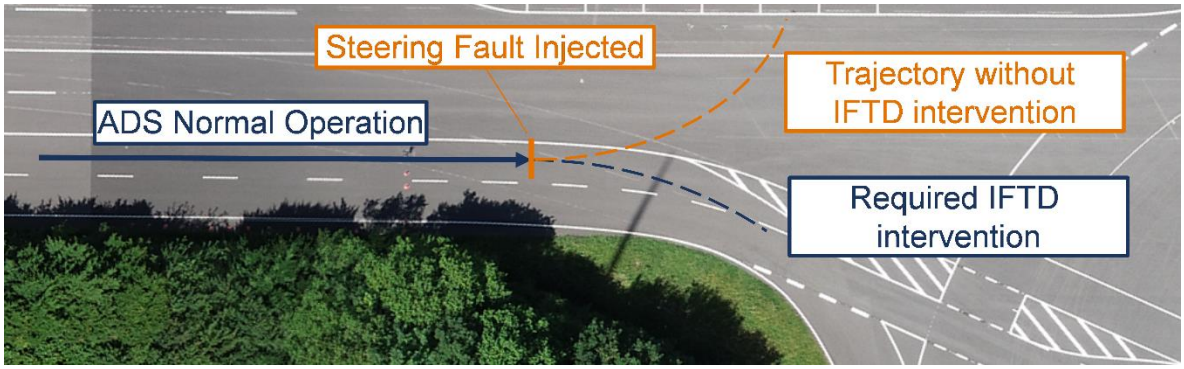


Figure A1.6: Example of fault injection test on test track

Driving Simulator training: Within the scope of the L3Pilot project, FKA did not implement any driving simulator tests in its IFTD training. But within the scope of Hi-Drive, FKA aims to use this tool to include high risk and more complex scenarios in its IFTD training.

Public road supervised training: Before including the IFTD into the IFTD pool for test drives, a new IFTD must perform the first drive on a public road with an experienced IFTD alongside.

Operational guidelines for testing and piloting

The second area of the safety manual concerns the Operational Guidelines for Testing and Piloting (OGTP), which requires collaboration with SP3 *Vehicles* and SP5 *Operations*. It is worth noting here that this document does not aim to be specific to any particular system or function. However, it is necessary that the ADS under scope is clearly defined, identifying its subsystems and units. A functional decomposition of the ADS may help to visualise these elements.

It is necessary to differentiate between the subsystems that belong to the vehicle itself (such as steering system, braking system, etc.) and those inherent to the ADS, since it might be that a production vehicle is used as a basis vehicle.

The operational guidelines cover four areas: i) preparation or general safety aspects, ii) system validation, iii) on-board human safety, iv) handling critical situations, and v) data logging. They are described in the following paragraphs.

Preparation or general safety aspects

In order to have a common safety approach, considering that the subsystems of a production vehicle may be used, it is necessary to consider ISO26262 (definition of safety goals, safety

case, verification & validation, etc.). Each of these shall be applied to the overall safety strategy, which involves a layered approach, which shall consider the machine related aspects (e.g., self-diagnostics, monitoring of fault messages, and so on), as well as the human related aspects (e.g., safety driver training, HMI, incident response, etc.). Additionally, the ADS shall consider requirements based on ISO/PAS 21448 (SOTIF), as well as address aspects such as behavioural competency, self-diagnostics, and code management processes.

This means that for *Behavioural Competency*, it is necessary that the system requirements are defined based on relevant signals and whether the behaviour is contextually safe during testing (e.g., code driving regulations such as posted speed limits, or keeping a certain distance around other road users). *Self-diagnostics* is necessary to identify when there are faults in the system or when certain manoeuvres would be contextually unsafe.

Finally, the *Code Management Process* considers software code reviews, specific tests to promote code quality, and continuous software testing.

System validation after system modifications

During testing of an ADS, there may be changes that affect the hardware, software, vehicle platform, or even the ODD. In the case of public road testing, and assuming that the vehicle platform, ODD, and hardware are already defined, a common aspect that undergoes changes is the software, such as improvement of existing features or newly developed features.

First, it is necessary to make an impact analysis of the changes to determine the implications of such changes. Next, the following steps must be applied:

- Step 1: Proof of concept. Based on the changes done, the level of testing must be defined so that it is appropriate for such a change, considering aspects such as criticality, magnitude of change, etc. This includes code verification (usage of code standards, expert code reviews, etc.) and specific testing such as unit tests. After these steps a complete build of the software will be available.
- Step 2: Simulated environment check. The built software with implemented changes must be then verified within a simulated environment.
- Step 3: Proving ground. The next step is validation on a proving ground, which is performed by an expert driver. The expert driver's role is to check that the intended change is having the expected effect on the real vehicle. After this step, the vehicle is ready to go out on public roads.
- Step 4: Test drives with low traffic volume. When testing on public roads in real-life conditions, the system must first be tested in low-density traffic.

- Step 5: Test drives in real and operative scenarios. The vehicle is ready for public road testing in any real-life traffic situation.

The steps described can be identified as safety gates that have to be passed. If at any of them the test is not passed, the issues need to be identified, solved, and the whole cycle executed from Step 1.

Handling critical situations

In order to properly address potential critical situations, the following aspects must be considered and established:

First, it is necessary to have a *Risk Management Process*, which is based on risk assessment and shall be executed at different levels (e.g., ODD, vehicle, safety drivers). The frequency and severity of the risks shall be properly assessed, to establish the appropriate measures and therefore prevent potential issues.

Then, the *Incident/Near-Incident Investigation and Reporting* needs to be considered. It is an incident response procedure for AD test vehicles on public roads that shall be put in place, defining the company's organisation and procedures for providing a response, in case of an incident involving an AD test vehicle on public roads. The incident response procedure follows the existing incident reporting rules in the company and adds procedures required for internal information sharing and organisation related to the AD test vehicle on public roads. An organisation structure with clear roles and responsibilities must be defined to connect all parties, from safety driver/operator through the company emergency window person and a dedicated AD Incident Investigation Team.

Data logging

During testing, data logging is a necessary part of the safety procedure to be able to retrieve information on several aspects that can be used for later system evaluation, such as critical events data, state of the ADS, and other data collection performance indicators that may be needed. Some data logging recommendations are suggested below:

- Permanent logging or ~30s before incident/critical events: It must be ensured that data from an incident are stored appropriately.
- Activation state of automation.
- Other data collection aspects: GPS, vehicle state (e.g., velocity, acceleration, steering angle), driver's condition.

It should be noted that the requirements for data logging in Hi-Drive go beyond this.

Besides the addressed topics, there are other aspects related to operational guidelines such as maintenance procedures of the vehicle platform and its subsystems (e.g., vehicle controller area network (CAN bus) system checks, tyre pressure, fuel maintenance, and so on), calibration procedures (for the different sensors equipped), and troubleshooting procedures that must be considered but are not addressed in this document.

During testing activity, it is recommended to have a daily usage check to cover aspects such as sensor state and driver licence status.

Outcomes and conclusions

This Annex presents the Safety Manual, which provides a strict safety procedure for the road tests. It included two main topics: In-Vehicle Fallback Test Driver (IFTD) and Operational Guidelines for Testing and Piloting (OGTP), including examples of the former.

The outcomes of the Safety Manual deal with the following topics:

- Operational guidelines for testing and piloting (which include safety driver and related training procedure, handling incidents, with investigation and reporting, operation permission, data logging, and so on).
- Guidelines for technical realisation (including vehicle modifications and system integration, assuring the availability of some features, such as pedestrian protection, override possibility, emergency stop, etc.).

In addition, the ODD enhancement should be carried out "step-by-step", considering the following phases:

- Simulated environment check.
- Proving ground (especially for new features).
- Real road conditions, but with less traffic density (e.g., other routes or calm traffic hours).
- Final tests in real and operative scenarios.

The entire Safety Manual should be applicable to the three operational states: Manual driving/ Baseline; Testing/Preparation phase; and Piloting/Treatment. On the other hand, since we are considering a prototype system solution that is not yet ready for market introduction in its current state, the proposed framework may deviate in regard to specific solutions implemented by partners.

Finally, the outputs can be used by WP5.4 *Operations* for safe operation execution and by SP7 *Effects* to assess especially how safety procedures can have an impact, such as the presence of a safety driver.

Despite the work done in building this Safety Manual, some important challenges remain. For example, different countries have varied regulation in this area that should be considered, and this should be harmonised; with reference to the previous point, cross-border activities should also be considered, because they are a kind of “free zone” where it is often not clear which regulation is valid and which rule should be applied.

Annex 2 Examples of different types of power analysis

Background

This Annex details several types of power analysis that can offer guidance for selecting a suitable sample size. Some are directly based on the work of Cohen (e.g., *Stationary driving* and *Percentages of occurrences* below). Other types of performance indicators are not covered in standard statistical textbooks (e.g., *Frequency of events* and *Duration of "time AD active"*). Here, Monte Carlo analysis is used to obtain relationships between study design parameters and statistical power.

Stationary driving

This is for research questions like "Does ADF reduce the average speed in free driving?" or "Does ADF change the average time gap in car-following?" Typically, this type of research question is not the most critical in terms of statistical power, since the associated performance indicators can easily be collected in large quantities in normal driving.

When comparing two means, a t-test for independent means can be appropriate. This case is covered by Cohen (1988, 1992). The first step in the power analysis is to obtain the normalised effect size d. This is defined as:

$$d = \frac{|m_{BL} - m_{TR}|}{\sigma}$$

where m_{BL} and m_{TR} are the means in baseline (BL) and in treatment (TR), while σ is the standard deviation (assumed identical for BL and TR here). Conventional levels for d from Cohen are 0.20, 0.50 and 0.80, respectively, for a small, medium, and large effect.

To obtain some feeling for the orders of magnitude involved, some of the L3Pilot Open Data (Weber et al., 2022) were analysed. With respect to the average speed in free driving, the results in Table A2.1 were obtained. Expressed in km/h, the effect size was 8.8 km/h.

Table A2.1: "Mean_v_" in free driving: averages, standard deviations, and number of observations in baseline and treatment (source: L3Pilot Open Data, Weber et al., 2022).

Condition	Average of Mean_v [km/h]	s.d. of Mean_v [km/h]	n
Baseline	113.7	22.8	92491
Treatment	104.9	25.1	179510

The pooled standard deviation equals 24.4 km/h. Using this value for the equation above gives a value of d=0.36. Using the same value, Figure A2.1 shows the normalised effect size as a function of the raw effect size ($m_{BL} - m_{TR}$).

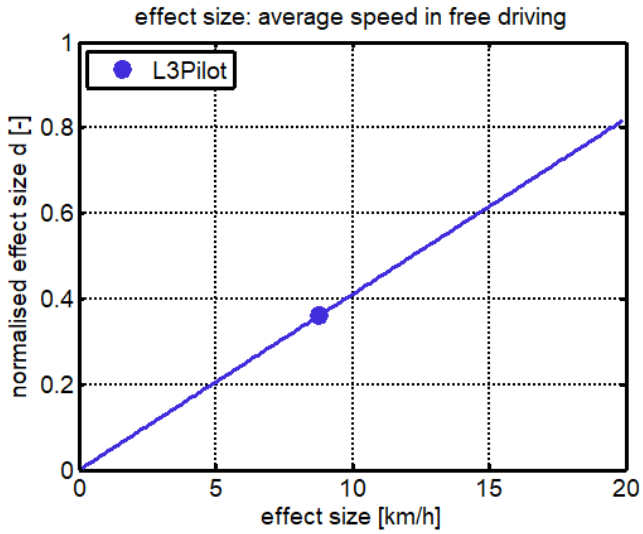


Figure A2.1: Average free-driving speed: normalised effect size as a function of raw effect size (based on L3Pilot Open Data). Also showing the effect size obtained from the same source.

Statistical power as a function of the normalised effect size is shown in Figure A2.2 for various levels of alpha and test types. It shows the required number of observations to reach a certain level of statistical power. Using the classical value of P=0.8, Table A2.2 shows the results for various levels of alpha, d, and test type. The most demanding numbers (small effect size, two-sided test, alpha = 0.05) are in the order of magnitude of 1500. Here it should be noted that an 'observation' consisted of a 10-second chunk of free-driving data. Thus, with experiments that cover multiple days of driving in normal traffic, this order of magnitude can easily be obtained.

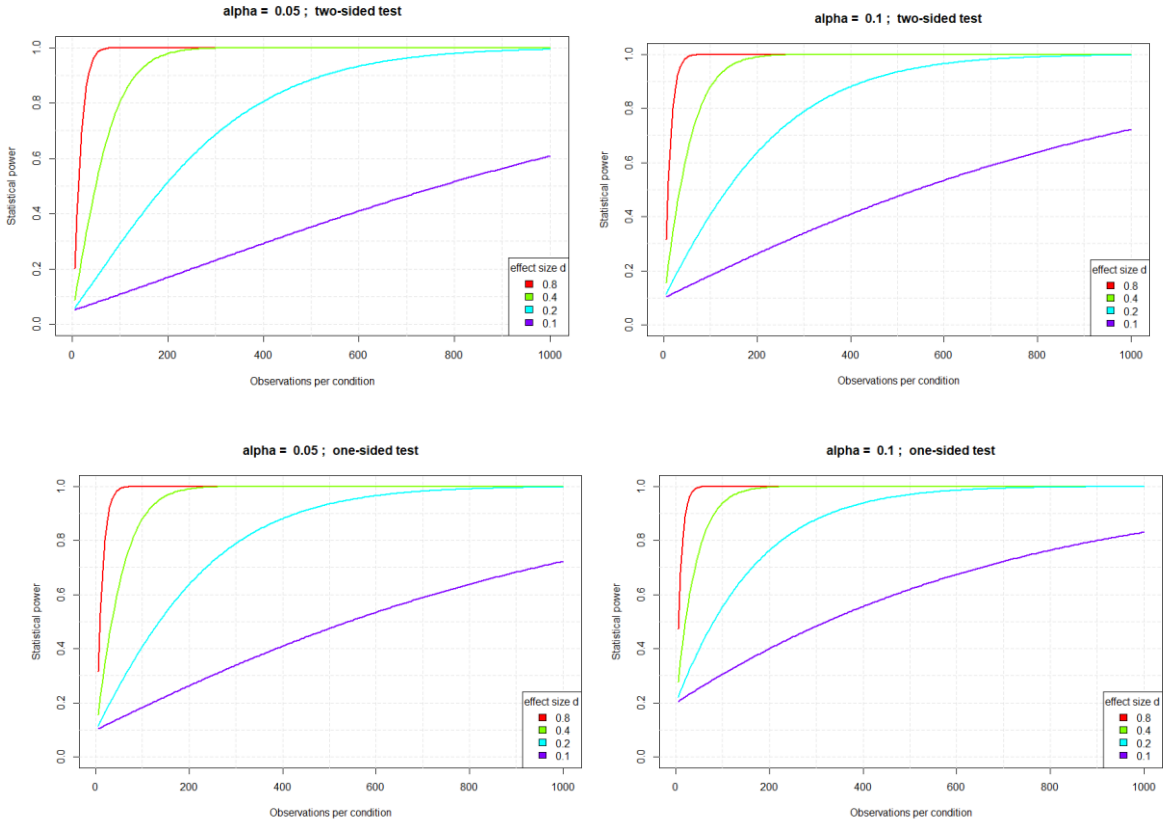


Figure A2.2: Statistical power as a function of the number of observations per condition and effect size d, for several levels of alpha and test types.

Table A2.2: Required number of samples per group (n) as a function of level of significance alpha, effect size d, power P, and test type.

alpha	d	P	type	n
0.05	0.8	0.8	two-sided	26
0.05	0.4	0.8	two-sided	100
0.05	0.2	0.8	two-sided	394
0.05	0.1	0.8	two-sided	1571
0.05	0.8	0.8	one-sided	21
0.05	0.4	0.8	one-sided	78
0.05	0.2	0.8	one-sided	310
0.05	0.1	0.8	one-sided	1238
0.1	0.8	0.8	two-sided	21
0.1	0.4	0.8	two-sided	78

alpha	d	P	type	n
0.1	0.2	0.8	two-sided	310
0.1	0.1	0.8	two-sided	1238
0.1	0.8	0.8	one-sided	15
0.1	0.4	0.8	one-sided	57
0.1	0.2	0.8	one-sided	226
0.1	0.1	0.8	one-sided	901

Percentages of occurrences

This section deals with performance indicators that express a percentage or proportion. This will usually be the proportion of attempts where a certain scenario or manoeuvre was successfully executed. Examples are

- successfully merging onto the motorway, for motorway merging functions,
- passing an intersection without having to stop for a red traffic light, for GLOSA (green light optimal speed advisory).

Most of the time, the analysis will consist of a comparison of two conditions: baseline and treatment (e.g., with versus without ADF, or with versus without an enabler). This requires a test for the difference between two independent proportions.

The first step in the power analysis is to obtain the normalised effect size h. The relationship between the (assumed) proportions in baseline and treatment (P_{BL} and P_{TR}, respectively) is as follows: (Cohen, 1988, 1992)

$$h=\left|2\left(arcsin\sqrt{P_{BL}}-arcsin\sqrt{P_{TR}}\right)\right|$$

This relationship is visualised in Figure A2.3. A few numerical examples are presented in Table A2.3.

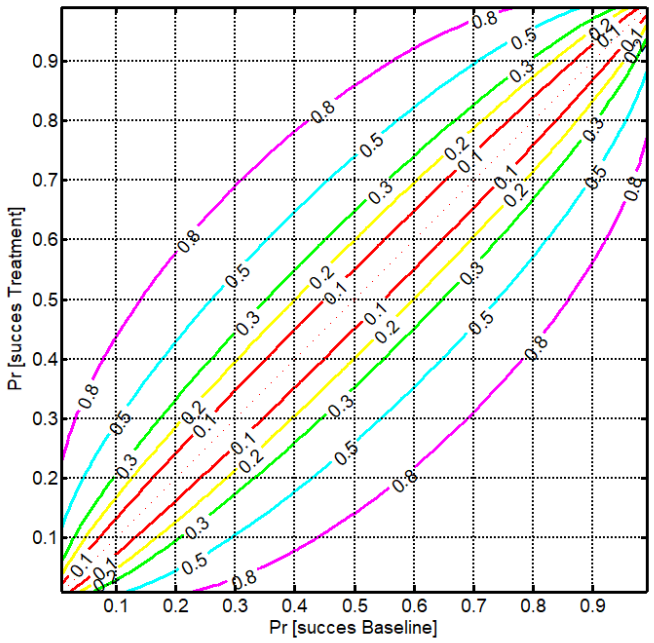


Figure A2.3: From proportions of success in baseline and treatment to effect size h

Table A2.3: Some examples of proportions of success in baseline and treatment and the corresponding effect size h.

P_bl	P_tr	h
0.50	0.55	0.10
0.50	0.60	0.20
0.80	0.85	0.13
0.80	0.90	0.28
0.90	0.95	0.19

For proportion tests, the classical values for a small, medium, and large effect according to Cohen are 0.2, 0.5, and 0.8, respectively. Looking at the examples in Table A2.3, it seems reasonable to want to be able to show that h=0.2 effects can indeed be found.

For a wide range of effect sizes h, Figure A2.4 shows the statistical power as a function of the number of observations per condition. These show how statistical power increases with increasing effect size and with increasing numbers of observations.

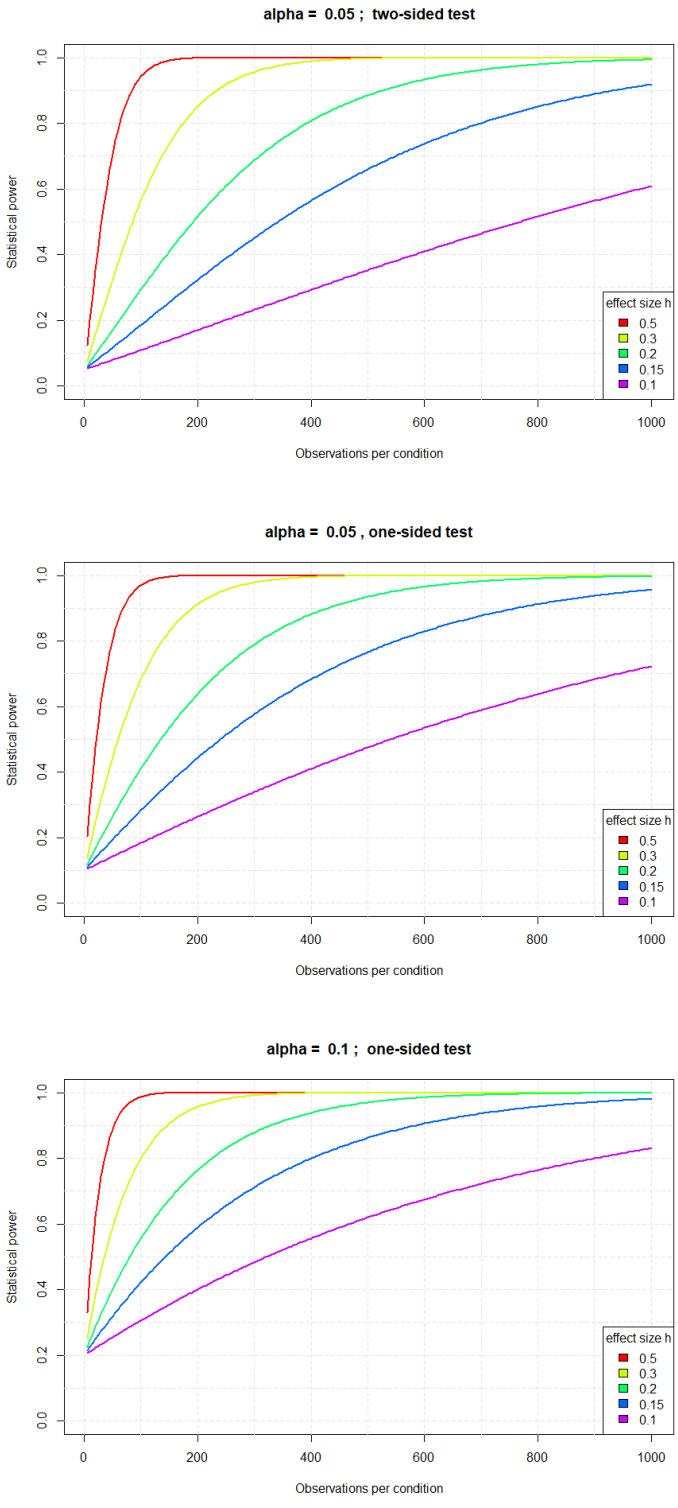


Figure A2.4: Statistical power as a function of the number of observations per condition and effect size h, for several levels of alpha and test types.

Figure A2.4 can also be used to assess how many observations are needed, such that a statistical power of 80% is reached. These results are presented in Table A2.4.

Table A2.4: Required number of samples per group (n) as a function of level of significance alpha, effect size h, power P, and test type.

alpha	h	P	type	n
0.05	0.5	0.8	two-sided	63
0.05	0.3	0.8	two-sided	175
0.05	0.2	0.8	two-sided	393
0.05	0.15	0.8	two-sided	698
0.05	0.1	0.8	two-sided	1570
0.05	0.5	0.8	one-sided	50
0.05	0.3	0.8	one-sided	138
0.05	0.2	0.8	one-sided	310
0.05	0.15	0.8	one-sided	550
0.05	0.1	0.8	one-sided	1237
0.1	0.5	0.8	one-sided	37
0.1	0.3	0.8	one-sided	101
0.1	0.2	0.8	one-sided	226
0.1	0.15	0.8	one-sided	401
0.1	0.1	0.8	one-sided	901

As Table A2.4 shows, an effect size of h=0.2 requires 226 observations per condition, even if we are settling for a liberal level of alpha = 10% and when a one-sided test is appropriate. Table A2.5 shows the required number of observations for even more liberal values of alpha, and the effect of lowering the required power to 70%.

Table A2.5: Required number of samples per group (n) as a function of level of significance alpha, effect size h, power P, and test type.

alpha	h	P	type	n
0.2	0.2	0.8	one-sided	139
0.2	0.2	0.7	one-sided	88

Frequency of events

This is for research questions like "Does ADF reduce the frequency of event X?" This can be the number of occurrences per unit of time or per distance driven. The event under consideration can be system-related (e.g., the frequency of take-over requests) as well as traffic scenario-related (e.g., the number of passive cut-ins experienced by the ego vehicle).

In the analysis presented below, the unit of observation is arbitrarily taken as one hour. To obtain power estimates, the occurrence of events is modelled as a Poisson process. This is characterised by one parameter, Lambda, which represents the expected number of events per hour. The research question is whether $Lambda_{BL}$ differs from $Lambda_{TR}$. In the Monte-Carlo analysis, the effect size ES was taken as a parameter with several levels, ranging from 5% to 20%. For a given level of $Lambda_{BL}$, the corresponding $Lambda_{TR}$ was chosen as $Lambda_{BL} * (100\% + ES)$. Using these parameters as input, thousands of virtual experiments were run, with various settings for the number of replications and the effect size. For each virtual experiment, a statistical test was conducted to assess whether or not the effect was significant. Statistical power was then estimated as the proportion of experiments where a statistically significant effect was found. Two-sided tests were applied. The results are shown in Figure A2.5.

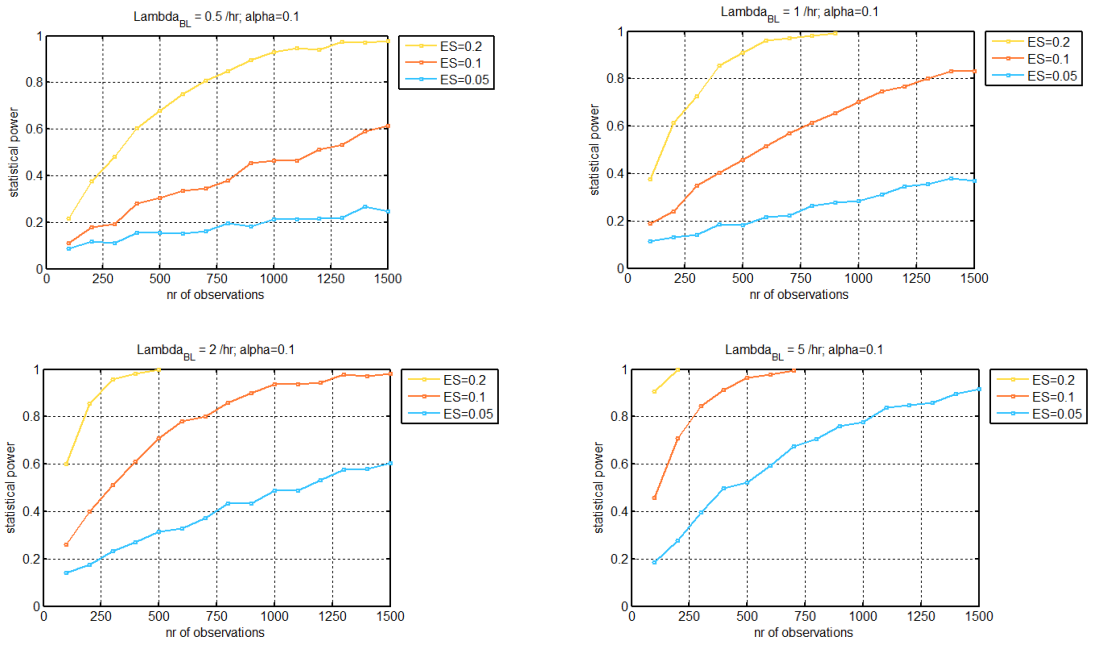


Figure A2.5: Statistical power as function of the number of replications per group, Lambda, and the Effect Size (alpha=0.05).

The choice of hours as the basic unit means that the 'number of observations' can also be read as the 'number of hours'. Figure A2.5 shows that depending on the frequency of events

and on the change thereof, the required number of observations (hours) needed to obtain a statistical power of 80% can vary from several hundreds to several thousands of hours.

Duration of "time AD active"

This is for a specific kind of research question, like "Do the enablers increase the time that AD remains active, once it has been activated?" This research question cannot be answered by investigating the percentage of time that the AD was active alone. If the driver is quick enough to activate AD, large proportions of trip time with AD active can be obtained, but that does not tell us if that consisted of a few long activation phases, or of many short ones. Therefore, the primary data for this analysis should consist of the durations of the individual "AD active" phases.

Monte-Carlo analysis was applied to obtain relationships between study design parameters and statistical power. The basis element in the simulations was an exponential distribution for the "duration AD active" phases. One parameter, mu, described the average time between occurrences (of AD activation). Based on L3Pilot data, an initial value of mu=4 minutes was used. This value is expected to increase as an effect of enablers, which justifies the use of one-sided tests. It is assumed that data are collected in runs that last 60 minutes. In the simulations, a run was always stopped at this end time, cutting the activation of that moment short even if AD could have continued.

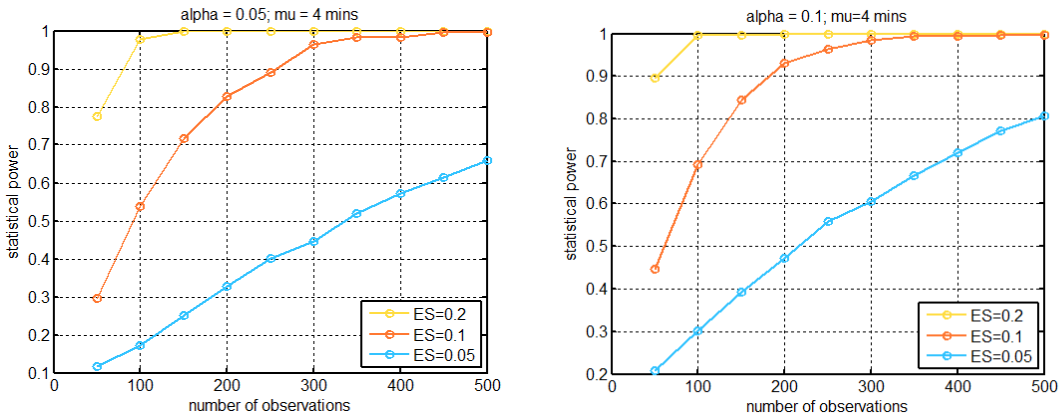


Figure A2.6: Statistical power as a function of the number of observations (hours) per group for various effect sizes.