Proceedings Paper:
Innamaa, S, Louw, T, Merat, N et al. (2020) Applying the FESTA methodology to automated driving pilots. In: Proceedings of TRA2020, the 8th Transport Research Arena. Transport Research Arena, 27-30 Apr 2020, Helsinki, Finland. Finnish Transport and Communications Agency Traficom. ISBN: 978-952-311-484-5. ISSN: 2669-8781.

Rethinking transport
27–30 April 2020

Proceedings of 8th Transport Research Arena TRA 2020, April 27-30, 2020, Helsinki, Finland

Applying the FESTA methodology to automated driving pilots

Satu Innamaa[a]*, Tyron Louw[b], Natasha Merat[b], Guihermina Torrao[b], Elina Aittoniemi[a]

[a]*VTT Technical Research Centre of Finland Ltd., P.O. Box 1000, FI-02044 VTT, Finland*
[b]*Institute for Transport Studies, University of Leeds, University Road, Leeds LS2 9JT, UK*

Abstract

This paper discusses the methodological challenges related to automated driving (AD) pilots in the real world, providing an overview of some of the solutions offered by the H2020 project L3Pilot. The FESTA methodology was developed in Europe in 2008, as part of the FP7 FESTA project, to provide guidance on a suitable methodology for conducting Field Operational Tests (FOTs). Although the overall methodology has been developed quite extensively for driver support systems, our efforts in the L3Pilot project show that the evaluation process can also be adapted to suit the needs of AD pilot projects, as long as some caveats related to the pilot nature of AD studies are acknowledged. This paper discusses how to use the FESTA methodology, and recommends an adapted FESTA V to be applied to an AD pilot made for the L3Pilot project.

Keywords: Real-world study; automated driving pilot; L3Pilot methodology; FESTA methodology

* Corresponding author. Tel.: +358-40-7610717;
E-mail address: satu.innamaa@vtt.fi

1. Introduction

1.1. L3Pilot project

In a step towards the introduction of automated vehicles on European road, the H2020 L3Pilot project is conducting large-scale piloting and evaluation of automated driving with developed SAE Level 3 (L3) functionality for passenger cars, which are exposed to a range of users in mixed-traffic environments, in different road networks. L3Pilot is coordinated by Volkswagen, and involves 34 partners, including 13 original equipment manufacturers (OEMs). The project has a total budget of 68 M€, and is 48 months in duration (2017–2021).

Extensive on-road testing is vital to ensure adequate operating performance for the automated driving functions (ADFs), which includes understanding the changes in vehicle operations and traffic dynamics, and users' interaction with, and acceptance of, ADFs. Therefore, a multidisciplinary methodology tailored to the assessment of automated driving is needed to ensure the success of the pilots in this project.

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

1.2. The FESTA methodology

The FESTA Handbook (latest version: FESTA 2018) defines a Field Operational Test (FOT) as "*a study undertaken to evaluate a function, or functions, under normal operating conditions in road traffic environments typically encountered by the participants using study design so as to identify real-world effects and benefits*". Here, "*normal operating conditions*" require that the participants use the tested systems during their daily routines, that data logging occurs autonomously, and that the participants do not receive special instructions about how and where to drive. Except in some specific occasions, there is no experimenter/observer in the vehicle, and typically, the study period extends over at least a number of weeks. In order to set up such an experiment, the technology readiness level of the function under investigation must be sufficiently high to allow it to be used by ordinary drivers in real traffic, and without supervision.

The FESTA methodology provides an extensive set of recommendations for developing and implementing an experimental procedure for FOTs. The different steps of an FOT are described by the so-called FESTA V (Fig 1).

1.3. Motivation for this paper

Barnard et al. (2016) wrote a paper on the methodology for automated driving FOTs for the TRA2016 conference and concluded their paper by saying that they did not provide the answer to whether a new methodology for performing automation FOTs is necessary or whether the existing one can be adapted, but raised many questions that would need to be considered when making that decision. Barnard et al. noted that "*FESTA is a living methodology, developed, adapted and adopted by a large community, based on consensus and sharing experiences*" and that that will be the case with a methodology for automation FOTs. They encouraged the use of a common methodology, one which retains the valuable approach and elements of FESTA, but develops new methods to answer new questions, and to bring knowledge on new focus points of the automation FOTs.

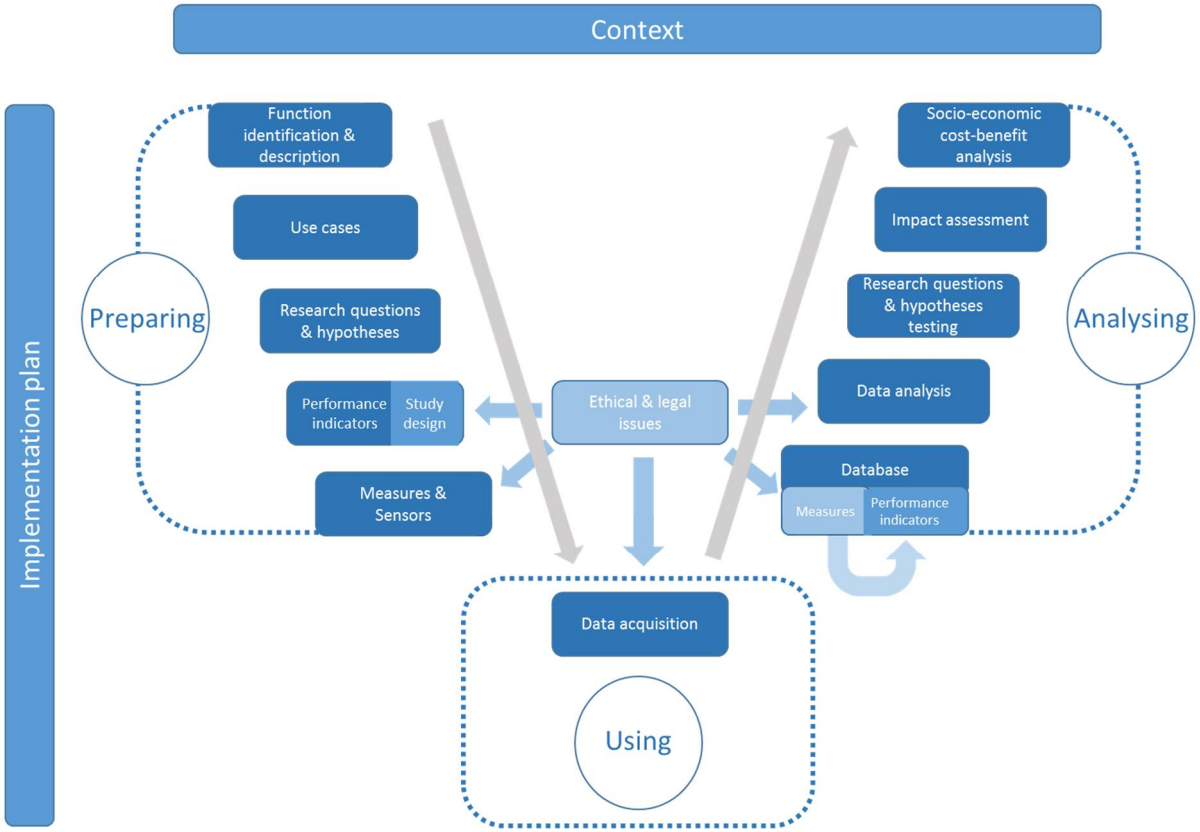

Fig. 1 FESTA V (FESTA Handbook, version 7, redrawn)

FESTA was designed to be applied to FOTs with market-ready products. Therefore, in its current state, it does not necessarily apply to studies with prototypical ADFs - as already suspected by Barnard et al. (2016). Thus, some adjustment of the "V" is needed for it to accommodate testing prototype functionalities, such as ADFs, in real traffic. The pilot nature of the tests in L3Pilot and other automated driving pilots brings some practical and ethical limitations, regarding the use of the automated vehicles, and limits any firm conclusions drawn about their implementation in the real world, or, indeed, their expected impacts. Therefore, we needed to make some adjustments to the methodology created for FOTs. To generate valid conclusions regarding the impacts of the ADFs, principles used to collect the evaluation data, and any ensuing conclusions, need to be considered carefully.

The objective of our paper is to discuss the FESTA V adapted for the L3Pilot project and the methodological challenges related to automated driving pilots in the real world. The paper is written from the evaluation viewpoint.

2. "FESTA V" for automated driving pilots on open roads

L3Pilot project adapted the original FESTA V (Fig. 2) to describe better the key-steps of the project. Specifically, the changes made to the original "V" are the following:

- The first step in the adapted FESTA V is the description of *Functions & use cases*. These two aspects are separated in the original FESTA V. Because in L3Pilot the use case description is linked to the operational design domains (ODD) of the tested automated driving functions (ADF), in our adapted FESTA V, these two aspects have been combined.
- The second step, *Research questions & hypotheses*, was kept the same in the adapted FESTA V as in the original.
- In the original FESTA V, defining performance indicators were combined with devising the study design, and defining measures and sensors was a separate step. In the L3Pilot project, we combined the data specification as one step (*Performance indicators & measures*) and handled *Study design* as a separate step because those working with data were (mostly) different persons from those planning the test routes and participants at the pilot sites, and belonged to different sub-project.

- The original FESTA V goes directly from *Measures & Sensors* to *Data Acquisition*. In the adapted version, we divided this into four steps: *Test site set-up* as part of preparation, *Pre-tests* and *Test* as phases of the actual data collection and *Test site wrap-up,* which will be conducted after the tests.
- The phases *Database* and *Data analysis* in the original FESTA V were combined into *Data processing* in our adapted version of the V.
- The original FESTA V phase *Research questions & hypotheses testing* was divided into three phases in the adapted FESTA V: [evaluation of] *Technical performance & cybersecurity*, *User acceptance* and *Driving & travel behaviour*.
- The evaluation phase *Impact assessment* was specified as *Impact on safety, mobility, efficiency and environment* in our adapted FESTA V for the L3Pilot project. The last phase of the evaluation was called *Socio-economic cost-benefit analysis* in the original FESTA V. In the adapted version for the L3Pilot project, we called it [the evaluation of] *Societal impacts* to also enable the incorporation of other analysis not considered the traditional cost-benefit analysis.

The subsections below describe each step of the adapted FESTA V.

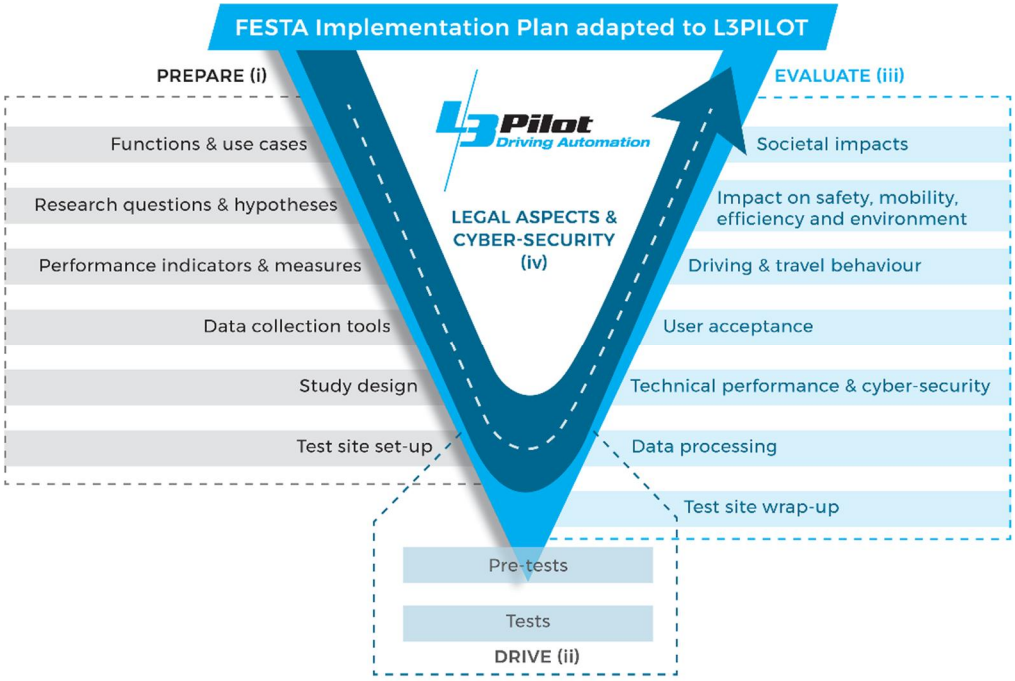

Fig. 2 FESTA V adapted to L3Pilot project

2.1. *Functions & use cases*

In the first step *Functions & use cases*, we make (first an early) description of the ADF(s) of the test vehicles from the user viewpoint:

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

Current automated driving pilots use vehicles with prototype ADFs in their field tests. These are used to collect vehicle technical data, as well as to evaluate test users' experience. However, if the tested ADFs have a low-level TRL (technology readiness level), then it would be inaccurate to use this data to scale up the results to high penetrations on a regional level (e.g. EU28). Therefore, in such cases, it is important to define the *mature ADFs* (i.e. what we expect the AFDs and their ODDs to be like when they are widely penetrated into the market) for the impact assessment. The description of these mature ADFs should be done together with those developing them. Thus, for the evaluation, it is important to describe ADFs

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

In practice, there are likely differences

- between test vehicles
- over time within a single test vehicle if the ADF development continues to unfold during the pilot project
- between the prototype ADF and the mature ADF
- between the test environment(s) & conditions and the entire ODD

In addition, there will also be inherent differences in the driving code and regulatory framework of the respective region or country, which could influence the way that the test pilot can be carried out.

For the evaluation, it is essential to understand and account for all these differences. Otherwise, there is a risk that the reliability of the user and acceptance, and impact assessment are compromised. Therefore, it is an iterative process to update the function and use case descriptions to facilitate accurate evaluations.

Information on functions tested in L3Pilot can be found in the deliverables D4.1 (Griffon et al., 2019) and D3.2 (Penttinen et al., 2019), and on the mature ADFs in the deliverable D3.3 (Metz et al., 2019).

2.2. Research questions & hypotheses

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

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

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

The process for formulating the research questions within L3Pilot project is described in detail in the L3Pilot deliverable D3.1 (Hibberd et al., 2018), while the final list of questions following the feasibility check described above, is presented in deliverable D3.3 (Metz et al., 2019).

2.3. Performance indicators & measures

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

The performance indicators and derived measures were set for the research questions in the L3Pilot deliverable D3.1 (Hibberd et al., 2018). These led to a common data format (Hiller et al., 2019) for the signal level data. This format was a product of a dialogue between the evaluation team (data needs) and the vehicle owners (data providers). The common data format derived from the information needs of the evaluation enables the use of common data processing tools which harmonise the data used in the evaluations.

2.4. *Data collection tools*

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

This step also includes the definition and development of databases into which the data is stored for different phases of processing and analysis. It may also be necessary that these tools can pseudonymise and anonymise the pilot sites without compromising the validity of evaluation results.

As not all the data is logged from the vehicles, this step also includes planning additional supplementing data collection methods to facilitate answering the research questions. In L3Pilot, this includes questionnaires, wizard of Oz vehicles, driving simulators, etc (See Metz et al., 2019, and Louw et al., 2020, for a detailed description).

2.5. *Study design*

As part of the fifth step, selecting the baseline is an important task in the planning of tests and evaluations. The evaluation team must consider for each research question, if a baseline is needed - and, if so, what a meaningful baseline is. Alternative baseline options include, e.g. fully manually driven car, SAE level 0 vehicles, today's situation or future ADAS situation. The inclusion or exclusion of active safety systems and other ADAS and the implications of their inclusion on the impact estimates must also be carefully considered. In the selection of the baseline, also the feasibility of the baseline data collection is a relevant factor.

An important step in the preparation of the pilot is to agree on the study design to be implemented at each pilot site. This includes selecting the test route and test participants, running the user tests, collecting baseline and treatment data, etc. The adoption of automated driving will change the mobility ecosystem, and therefore, new innovative assessment approaches may be needed. One example of the need for adjustment of the FESTA V relates to the FESTA guideline that the "ordinary user" should be able to operate the tested function in real traffic without supervision. However, owing to the practical, safety, and ethical issues related to a pilot, the L3Pilot approach restricts the operation of ADFs to predetermined test routes. Moreover, the L3Pilot approach needed to be adjusted for the type and role of the test participant (professional driver, safety driver, and ordinary driver). For example, when dealing with prototype systems in mixed traffic on open roads, the use of specially trained safety drivers is a requirement for safety, legal and ethical reasons. Consequently, it is not possible to conduct the same nature of research as in a naturalistic driving study or FOT. This has implications for the scope of user and acceptance evaluations and studies on travel behaviour, and the collection of baseline data. Therefore, the evaluation team needs to anticipate and plan for how to utilise the controlled tests on open roads and supplementing studies to get the data needed for the evaluation.

As there are differences in the rules and regulations for field-testing for each vehicle manufacturer and region, an active dialogue between the evaluation team and the pilot sites during the study planning is needed. Otherwise, the possibility for assessing generalised results across all pilot sites may be difficult. The solutions we developed for experimental procedures for the pilot sites are described in L3Pilot deliverable D3.2 (Penttinen et al., 2019).

2.6. *Test site setup*

For the *Test site setup* step, an active dialogue between the evaluation/methodology team and the pilot site is recommended to ensure that the agreed experimental procedures are correctly implemented. In the L3Pilot project, we have dedicated one evaluation partner for each pilot site to plan the practical tests together. Each evaluation partner is also responsible for the analysis and processing of the field data (common data format data) at its respective pilot site. Therefore, (s)he needs to fully understand the details of the field experiments.

2.7. *Pre-tests*

Pre-tests step involves running all the phases of the project on a small scale to ensure that all the processes and tools chains function as intended. This should include all the steps in the evaluation, related data flows, test participant handling, etc. The verification of all the steps is vital to the success of the project. The possibilities for

corrective actions after the start of data collection are limited. Hence, pre-tests are useful to anticipate potential caveats and to validate the designed study before the start of the real pilot.

2.8. Tests

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

2.9. Test site wrap-up

Test site wrap-up includes the delivery of data, including situational data and other metadata. In this phase, it is also important to report in detail all the deviations from the plan and any system updates made during the data collection phase (which aspects was updated when and what change this update might have had on users' experience and the ADF behaviour).

2.10. Data processing

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

2.11. Technical performance & cybersecurity

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

2.12. User acceptance

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

Furthermore, common criteria for recruiting and selecting the test participants across pilot sites must be defined, such as: all test participants should regularly drive, demographic factors should reflect the driver population of interest, for example, the future customer population (depending on evaluation scope).

The detailed plans for user and acceptance evaluation can be found in the L3Pilot deliverable D3.3 (Metz et al., 2019) and Louw et al. (2020).

2.13. Driving & travel behaviour

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

The detailed plans for driving and travel behaviour evaluation can be found from the L3Pilot deliverable D3.3 (Metz et al., 2019) under methods for technical & traffic evaluation and under mobility impact assessment.

2.14. Impact on safety, mobility, efficiency and environment

This phase of evaluation assesses the impacts on safety, mobility, efficiency and environment and scales them up to EU28. These results are needed in the last phase of evaluation which is the assessment of the societal impact.

The impact assessment phase utilises data or results from the *Driving & travel behaviour* evaluation and the *User acceptance* evaluation, like results of differences in driving behaviour and of willingness to have, etc. Additional inputs include the descriptions of the mature ADFs and their ODDs and the penetration rates that will be used in the assessment of the *Societal impacts*.

The impacts are first assessed on the level of a single event (single driving scenario), then on traffic flow and region. In the L3Pilot project, the impact assessment results are scaled up to the EU28 level.

An assessment in an automated driving pilot includes more uncertainty compared to a traditional FOT of a market-ready product. The sources for that uncertainty include but are not limited to:

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

Naturally, the reliability of impact assessment results also depends on the sophistication of the simulation tools.

The detailed plans for impact assessment can be found from the L3Pilot deliverable D3.3 (Metz et al, 2019).

2.15. Societal impacts

Key factors for the success of the assessment of the societal impacts of automated driving is the accuracy of statistics and how the details of existing statistics meet the needs of this phase of evaluation. If the statistics do not provide information on how many road crashes today in the EU occur on roads and in conditions matching the ODD of the automated vehicles, or how much vehicles' CO2 emissions are produced while driving in environments and conditions when ADFs could be used. This brings uncertainty to the scaling up of the results and reduces the reliability of the impact estimates.

For the cost-benefit analysis, a further source of uncertainty is the lack of reliable (publically available) estimates of the costs of these future systems, since they are still under development.

The detailed plans for socio-economic impact assessment can be found from the L3Pilot deliverable D3.3 (Metz et al., 2019).

3. Conclusion

This paper aimed to discuss the process of setting the methodology, collecting data and performing the evaluation in an automated driving pilot. Specifically, this paper described the FESTA V adapted for the L3Pilot project. It provides recommendations and discussed the methodological challenges related to automated driving pilots in the real world, and it provides an overview of some of the solutions to these challenges offered by the L3Pilot project, presented through an adapted FESTA V framework, and from an evaluation viewpoint.

The main differences between the original FESTA V and that adapted here for the L3Pilot project originate from the difference between the traditional FOTs of (nearly) market-ready products and a pilot study of systems still in a prototype phase, in a highly competitive market. On the one hand, the process needs to protect the intellectual property rights of the vehicle manufacturers and their competitive position. On the other hand, the outcomes of the evaluation need to be meaningful and meet the aims of the project.

In practice, one clear difference between an FOT and an AD pilot is that the latter one requires more iteration of

the steps in the preparation phase in order to align the evaluation plans with the practicalities at the pilot sites. In addition, the rules and regulation related to the AD pilots on open roads are more strict than those related to the first FOTs. Thus, the planning of all the test phases needs to be planned in close collaboration between the evaluation team and those implementing the pilot sites.

Even though the large-scale impact estimates of automation cannot be fully guaranteed, the impacts of automation are potentially so far-reaching that it is important to get these early phase estimates. These first large-scale field tests provide an excellent opportunity for conducting these assessments.

We recommend the use of the FESTA V adapted for the L3Pilot project as described in this paper also for other AD pilots conducted on open roads.

Acknowledgements

The research leading to these results has received funding from the European Commission Horizon 2020 program under the project L3Pilot, grant agreement number 723051. Responsibility for the information and views set out in this publication lies entirely with the authors. The authors would like to thank partners within L3Pilot for their cooperation and valuable contribution.

References

Barnard, Y., Innamaa, S., Koskinen, S., Gellerman, H., Svanberg, E., Chen, H. (2016). Methodology for Field Operational Tests of Automated Vehicles. In: Rafalski, L., Zofka, A. (eds). Part of special issue for Transport Research Arena TRA2016. Transportation Research Procedia, Volume 14, 2016, Pages 2188-2196.

FESTA (2018). FESTA Handbook, version 7. https://connectedautomateddriving.eu/wp-content/uploads/2019/01/FESTA-Handbook-Version-7.pdf

Griffon, T., Sauvaget, J.-L., Geronimi, S., Brouwer, R. (2019). Description and Taxonomy of Automated Driving Functions. Deliverable D4.1 of L3Pilot project, version 2.0.

Hibberd, D., Louw, T., Aittoniemi, E., Brouwer, R., Dotzauer, M., Fahrenkrog, F., Innamaa, S., Kuisma, S., Merat, N., Metz, B., Neila, N. (2018). *From research questions to logging requirements. Deliverable D3.1 of L3Pilot project.* European Commission.

Hiller, J., Svanberg, E., Koskinen, S., Bellotti, F., Osman, N. (2019). The L3Pilot common data format – Enabling efficient automated driving data analysis. ITS European Conference, Eindhoven, 10-13 June 2019.

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

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

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