Proceedings Paper:
Louw, T, Merat, N, Metz, B et al. (2020) Assessing user behaviour and acceptance in real-world automated driving: the L3Pilot project approach. In: Lusikka, T, (ed.) Proceedings of TRA2020, the 8th Transport Research Arena. Transport Research Arena, 27-30 Apr 2020, Helsinki, Finland. Finnish Transport and Communications Agency Traficom. Article no: 806, p. 105. ISBN: 978-952-311-484-5. ISSN: 2669-8781.

Rethinking transport
27–30 April 2020

TRA 2020 HELSINKI

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

Assessing user behaviour and acceptance in real-world automated driving: the L3Pilot project approach

Tyron Louw[a]*, Natasha Merat[a], Barbara Metz[b], Johanna Wörle[b], Guilhermina Torrao[a], Satu Innamaa[c]

[a]*Institute for Transport Studies, University of Leeds, University Road, Leeds and LS2 9JT, UK*
[b]*WIVW Würzburg Institute for Traffic Sciences GmbH, Robert-Bosch-Straße 4, 97209 Veitshöchheim, Germany*
[b]*VTT Technical Research Centre of Finland Ltd., P.O. Box 1000, FI-02044 VTT, Finland*

Abstract

The L3Pilot project, funded by H2020, is conducting the first large-scale piloting of SAE Level 3 automated driving in Europe. The main aim of the project is to address a number of key questions in a step towards introducing automated vehicles on European roads. This paper discusses the approach taken by the L3Pilot project, to evaluate user behaviour in, and acceptance of, automated driving in real-world pilots. Although some technical challenges associated with the development and demonstration of such technologies are well-documented, current methodologies, such as those used to evaluate Field Operational Tests (FOTs), offer little guidance about assessing the impact of automated driving on users' behaviour and acceptance. This paper outlines the methods used and developed for assessing user behaviour and acceptance within the project, summarises some of the methodological challenges involved in collecting data during an automated driving pilot, and discusses some approaches we have developed to solve these multifaceted challenges.

Keywords: Real-world study, automated driving pilot, methodology, human factors, user acceptance, user behaviour.

* Corresponding author. Tel.: +44(0)113 343 7882;
E-mail address: t.l.louw@its.leeds.ac.uk

1. Introduction

Over the years, numerous European research projects have developed and tested automated driving (AD) systems. Significant progress has been made, but higher-level AD systems (i.e. SAE Level 3 or higher; SAE, 2018) are not yet ready for market introduction. AD is not realised simply by integrating more and better technology and addressing the numerous legal, ethical, security, safety, and infrastructural considerations, but its success relies on understanding user behaviour, and including the contribution of user needs, limitations, and concerns, into the design cycle (Xiong et al., 2012).

To date, our understanding of user behaviour during AD has mainly been informed by research using driving simulators of different fidelities (cf. Gold, Körber, Lechner, & Bengler, 2016; Louw et al., 2017; Li, Blythe, Guo, & Namdeo, 2018). This type of controlled environment provides a safe, cheap, and repeatable means of investigating users' perception, reactions and limitations, during their interactions with AD systems, particularly in critical situations. Driving simulator studies have also been used for this purpose because it is not yet possible to test such interactions in the real world, but they are unable to capture the entire range of everyday decisions that are likely to occur in the real world (Carsten et al., 2013), such as when and to what extent drivers engage AD, and under what conditions.

The use of an on-road pilot study allows investigations in a more naturalistic setting, also providing more knowledge about the likely effect of different road environments, and surrounding traffic on the behaviour of the automated driving system, and the consequent effect of this on user behaviour and acceptance.

1.1. L3Pilot, the first large-scale piloting of automated driving

The L3Pilot project (https://www.l3pilot.eu/) is co-funded by the European Commission, coordinated by Volkswagen, and includes 34 partners, incorporating 13 Original Equipment Manufacturers (OEMs). The project started in October 2017, has a total budget of 68 M€, and is 48 months in duration. L3Pilot is conducting the first large-scale testing and piloting of automated driving with SAE L3 and L4 functionality for passenger cars, in a step towards introducing automated vehicles on European roads. The project is generating a large amount of technical and behavioural data, from a range of vehicles (~100) and participants (~1000) being tested in different traffic and road environments at 13 sites across Europe. Within L3Pilot, the systems that are being tested are referred to as Automated Driving Functions (ADF), which is defined as an *"activity or purpose of a vehicle to enable automated driving*" (Penttinen et al., 2019). An AD system, on the other hand, is defined as a *"combination of hardware and software required to realise an ADF"* (Penttinen et al., 2019). There are four different types of ADFs being tested within L3Pilot (for more details see Metz et al., 2019):

- The *motorway ADF* can operate on motorways and other two-carriage way roads in uncongested conditions, at speeds up to 130 km/h.
- The *traffic jam ADF* can operate in high-density traffic or congestion situations and at speeds up to 60 km/h.
- The *urban ADF* can operate on urban roads at speeds ranging between 25 km/h and 50 km/h.
- The *parking ADF* can perform parking manoeuvres and also drive to the parking spot, and will be tested in a controlled study in a closed environment, such as a private parking lot or designated area.

The overall objective of the L3Pilot project is to test and study the viability of AD as a safe and efficient means of transportation, gain knowledge for exploring and promoting new service concepts, and provide inclusive mobility. To that end, the project has four primary focus areas. First, *technical and traffic evaluation* assesses the effect of the ADF on vehicle performance, and the surrounding traffic, based on data logged directly by the test vehicles in the on-road tests. Second, the *user and acceptance evaluation* assesses users' experience in, acceptance of, and behaviour with, the ADFs. Third, *impact assessment* extrapolates these results and assesses the potential impacts of so-called mature ADFs[†] on factors such as personal mobility, traffic safety, traffic efficiency, and the environment. Finally, *socio-economic impact assessment* utilises the above analyses to determine monetary values for the estimated effects, and weighing expected costs and benefits of the ADFs on a societal level (EU28) with different penetration rates.

[†] What we expect the AFDs and their Operational Design Domains (ODD) to be like when they are widely penetrated into the market

The objective of this paper is to outline the data collection methods we have developed within the project to conduct the *user and acceptance evaluation*, including the real-world pilots and supplementary studies, and to provide an overview of some of the challenges we have faced and solutions we have developed during this process.

2. User and acceptance evaluation approach

Since L3Pilot is conducting the first large-scale testing and piloting of ADFs across a range of vehicles, and test sites, new assessment approaches have been required. For example, the existing FESTA (Field opErational teSt supporT Action) methodology provides an extensive set of recommendations for developing and implementing an experimental procedure for assessing market-ready driver support systems in Field Operational Tests (ARCADE, 2018). However, this methodology has had to be developed for automated driving pilots, since the ADFs in L3Pilot are still prototypes (see Penttinen et al., 2019; Innamaa et al., 2020).

One aspect of developing the methodology for L3Pilot was to develop a series of detailed research questions for each of the four evaluation areas outlined above. The first step in this process involved creating a wide-ranging set of research questions, uninhibited by the limitations of any single data collection methodology. These were based on the established literature, and had input from project members, based on their experience in previous, related, work. The second step involved rating the importance of each question against criteria where were relevant to each evaluation area. For example, for the user and acceptance evaluation, a question was rated highly if it advanced our understanding of the safety and acceptance of the tested system. The third step involved rating the feasibility of answering each question, based on the methods and resources available within the project. Step three was an iterative process, because questions that could not be answered by the pilot studies could potentially be answered using supplementary methods (described below). Research questions that were rated as highly important and highly feasible were automatically adopted by the project. Research questions that were rated as highly important, but ultimately not feasible, were not adopted. Research questions that were rated as low importance were not generally adopted by the project, irrespective of their feasibility, to allocate the resources for the areas with highest priority.

Following the process described above (and in Hibberd et al., 2018), the research questions were organised into a number of key themes, including user acceptance and trust of the systems, willingness to use and pay for the functionalities, measures of driver state (stress, distraction, fatigue, workload), user risk perception, driver engagement in non-driving related tasks, user behaviour during, and after, take-over situations, and user motion sickness. The sub-set of research questions for the user and acceptance evaluation are shown in Table 1, with more detailed derivatives of each presented in Metz et al. (2019).

Based on the specific research questions, and the fact that pilot studies cannot provide the data to answer all research questions, the project developed a multifaceted assessment approach, to form a holistic view of users' behaviours with, and acceptance of the ADFs. These include data from a combination of quantitative and qualitative data collection methodologies, centred primarily around the pilot studies, including user questionnaires, videos of the driving scene, recordings of drivers' head, hands, and posture during the pilot, and vehicle-based data. Data is also collected from supplementary studies, including driving simulator and Wizard-of-Oz studies, and a large-scale international survey. Each of these methods of data collection is introduced and discussed in the sections below.

This multifaceted data collection and analysis approach is used regularly in Field Operational Tests (FOT) or Naturalistic Driving Studies (NDS) studies investigating user behaviour. For example, the UDRIVE (Lai et al., 2013; van Nes, Bärgman, Christoph, & van Schagen, 2019), SHRP2 (Dingus et al., 2015), ecoDRIVER (Jamson, Kappe, & Louw, 2014) and DRIVEC2X (Brizzolara et al., 2014) all relied on both subjective (attitude and behaviour questionnaires) and objective (vehicle and video) data in their evaluation, which were supplemented by interviews, focus groups, or self-confrontation sessions.

Table 1. User and acceptance research questions in L3Pilot and the methods used to address them. Vehicle and video data includes TOC rating.

Research Question	Sub-Research Question	Real-world pilot				Other methods		
		Vehicle -based analysis	Video-based analaysis	Pilot site questionnaire	Interview/Focus group	Driving Simulator	Wizard-of-Oz	Annual Survey
What is the impact of ADF use on user acceptance & awareness?	Are drivers willing to use an ADF?	x	x	x	x			x
	How much are drivers willing to pay for the ADF?			x	x			x
	What is the user acceptance of the ADF?			x	x	x	x	x
	What is the impact of ADF on driver state?		x		x	x		
	What is the impact of ADF use on driver awareness?		x	x	x	x		
	What are drivers' expectations regarding system features?			x	x	x	x	x
What is the impact of ADF use on user experience?	What is drivers' secondary task engagement during ADF use?		x	x		x	x	
	How do drivers respond when they are required to retake control?	x	x	x	x	x	x	
	How often and under which circumstances do drivers choose to activate/deactivate the ADF?	x	x		x	x	x	
	What is the impact of ADF use on motion sickness?			x				
	What is the impact of motion sickness on ADF use?			x				

2.1. Pilot site questionnaires

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

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

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

The second part of the questionnaire is administered immediately after the pilot drive concludes, or the final pilot drive if a participant participates in more than one drive. It examines test participants' initial reactions regarding their experience while using the particular ADF, including acceptance, safety and comfort, among others. To

examine whether participants felt that they would change any of their behaviours should they have access to that particular ADF in their daily lives, they were re-asked questions about vehicle use and purchasing decisions, driving history, in-vehicle system usage, engagement with non-driving tasks, trip choices, and mobility patterns. The questions in this section are phrased to address the specific ADF under investigation, with the only exception being motorway and traffic jam ADFs, which utilise the same questions, because they have similar Operational Design Domains (ODD).

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

2.2. Video-based analysis

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

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

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

2.3. Vehicle-based analysis

In L3Pilot, vehicle-based data is mainly used to answer research questions in the area of technical and traffic evaluation, for example, examining the performance of the ADF during lane-keeping situations. However, it can also be used to address some user-related topics, since there are scenarios where the driver interacts with the system and controls the vehicle. For example, examining how often, and under what circumstances, drivers choose to activate and deactivate the ADF. Vehicle-based data can also be used to analyse take-over situations, thus supplementing the TOC-rating with more specific indicators, such as take-over times, or more specific vehicle controllability measures, including the magnitude of lateral or longitudinal accelerations (cf. Louw et al., 2017). Whether such analyses are a useful approach within L3Pilot depends on the experimental protocol at each pilot site, because the TOC-rating approach requires similar take-over scenarios across drivers and trips and there may be slight variations between pilot sites in terms of the traffic conditions that drivers are exposed to during the pilot.

2.4. International user acceptance survey

The pilot site questionnaires provide a unique opportunity to understand how drivers who have used the ADFs perceive and accept them. However, the pilot sites have unavoidable differences in terms of the participant population and experimental design. This means that the sample size available for analysing some question or topics may not have the statistical power for drawing robust conclusions about the interactions of various socio-demographic factors, user groups and their influence on acceptance of conditionally automated driving (Price, Daek, Murnan, Dimmig, & Akpanudo, 2005). For example, since some questions related to acceptance are tailored to the specific ADF being tested, responses across all ADFs cannot be combined in the evaluation. Therefore, the sample size for some questions may be less than 100, or less, if responses from participant types cannot be combined. Therefore, a much larger sample is needed. To address this shortcoming, L3Pilot is conducting a two-phase international questionnaire study on ~24,000 car drivers, to investigate the factors that influence acceptance of SAE L3 AD (i.e. conditionally automated vehicles) in this group of ordinary drivers.

The first phase of this survey was administered between April and November 2019, and included 78 items. To develop a global perspective, the survey included respondents from eight European countries (UK, Sweden, France, Germany, Italy, Hungary, Finland, Spain), and eight non-European countries (China, USA, Brazil, India, South Africa, Turkey, Japan, and Indonesia), with each country represented by 1,000 respondents. These countries were chosen because they represented different geographical regions in Europe and the other continents, and have large car markets. The survey addresses factors such as user acceptance, mobility, privacy, trust, perceived safety, willingness to pay and use, technology readiness, experience with road vehicle automation, and knowledge of ADFs. The survey also includes an adapted version of the Unified Theory of Acceptance and Use of Technology (UTAUT) questionnaire, initially developed by Venkatesh, Morris, Davis, & Davis (2003), which provides a comprehensive synthesis of research to model technology acceptance. In L3Pilot, this questionnaire will be used to understand the factors that influence respondents' acceptance of L3 AD, and build a more comprehensive model, which incorporates aspects such as technology readiness and socio-demographic factors, including age, gender, driving experience, country of residence, education, household, employment status, and income.

The second phase of this exercise will include an updated survey, informed by the results of phase one, and will be administered in the last quarter of 2020 to respondents from eight countries across all continents. Where feasible, the survey uses items from the pilot site questionnaire so that we can examine the extent to which real-world exposure to the ADFs affects respondents' perceptions.

2.5. Driving simulator and Wizard-of-Oz studies

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

2.6. Interviews and focus groups

Within L3Pilot, interviews and focus groups are used to assess drivers' views of the tested ADFs and the potential implications of the use, or availability, of these systems on users' daily travel behaviour (cf. Pudāne et al., 2019). For example, whether individuals feel more comfortable using an ADF during particular types of trips, such as for commuting or leisure trips. Interviews allow participants to express more detailed views related to their

experiences during piloting, while focus groups allow participants to develop ideas and insights into their own experiences, based on views expressed by others (Stewart & Shamdasani, 2014). These methods may also be used to gather participants' views on future scenarios or concepts, such as ride-sharing in an automated vehicle, or to discuss the impacts on, or acceptance of, AD, by specific user groups. Focus groups can also be used to gain a more comprehensive insight into the potential impact of how access to an ADF affects travel behaviour (personal mobility). The outcome of the interviews and focus groups will also be used in the impact assessment evaluation. For example, by combining results from participants' views of AD with publically available travel behaviour data and statistics, it is possible to assess the implications of AD at an EU level.

3. Challenges for user and acceptance evaluation in AD pilots

Conducting real-world automated driving pilots introduces a number of methodological challenges. While the broad methodological challenges faced within L3Pilot are discussed in detail in Innamaa et al. (2019), the sections below discuss aspects specifically related to the user and acceptance assessment. One limitation for AD pilots is related to the external validity of the results obtained, especially regarding user data, and particularly when it is part of a large-scale, multi-partner project, involving real traffic. Challenges related to this include differences in system maturity, variations in the test environment and experimental design, and the type of user recruited (e.g. professional (test), or ordinary driver, or passenger). These elements of the pilot are discussed further below.

3.1. Selecting test participants

The evaluation methods developed for the L3Pilot project cover different types of test participants, including ordinary driver, professional (test) driver, safety driver, and passenger. Within L3Pilot, ordinary drivers are defined as, "*individuals who hold a licence granting them permission to drive on public roads, but do not have any additional driving qualifications or permits, such as racing licences, and do not drive or test vehicles as part of their work*." Pilot sites are encouraged to use ordinary drivers as test participants in every instance possible. However, due to the prototypical nature of the systems, for safety, privacy, and legal reasons, this requirement can not always be fulfilled. Instead, safety/professional drivers will operate the vehicles in most cases. Within L3Pilot, professional (test) drivers are defined as, "*individuals who drive vehicles as a profession, or as part of their day-to-day work, for remuneration, and have typically extensive driving experience. As part of their training, they have been trained to e.g. handle cars in critical situations. These drivers can be deployed to operate prototype vehicles undergoing road tests*". Some of the professional (test) drivers have also been trained to operate the specific ADF being tested, and, therefore, are referred to as safety drivers. Safety/professional drivers' perceptions of, and behaviour while using, ADFs are likely to be influenced by their specialised training and more in-depth understanding of the systems' functioning. Therefore, responses from each participant type will be treated separately in the evaluation.

Based on the rules at the respective pilot sites, ordinary drivers may be allowed to drive, but will be accompanied by safety drivers who will intervene (only) in dangerous situations. The presence of the safety driver in the car might influence the perceptions and behaviour of the ordinary driver, as research shows that participants perform more poorly on tasks in an experimental setting when being watched by an experimenter (Belletier et al., 2015). Careful consideration must, therefore, be given to the instructions that the professional (test) driver receives, in terms of how they interact with the participant. However, ethical matters must also be considered carefully, because while it may be preferable for the technical and traffic evaluation if the tested system was allowed to operate to the edge of its limits (or ODD), it is more important that professional (test) drivers' actions prioritise safety of the occupants inside the vehicle, and that of the other road users.

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

3.2. Variations in ADF maturity

The vehicles that test ordinary drivers use during the pilots contain prototype human-machine interfaces and automated driving control systems. These systems are still under development, and their maturity may inevitably vary between pilot sites, and potentially within pilot sites, should any updates be required during the prolonged

testing schedule at some pilot sites. The use of "imperfect" prototypes and any unexpected behaviour of the systems may occasionally result in unpleasant driving or interaction experiences for users. These encounters may well affect user experience, and thus acceptance, since a development system that is prone to errors is likely to elicit different acceptance ratings, compared to a market-ready system.

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

3.3. Variations between test environments

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

The approach adopted within L3Pilot to deal with the above issues is two-fold. First, experimental guidelines have been developed to align experimental approaches across pilot sites, with an attempt to control for the study design, instructions for the selection of test participants, experimental protocol, and participant instructions and information. These have been described in detail in L3Pilot Deliverable D3.2 (see Penttinen et al., 2019). Second, where possible, information is collected about the variations between, and within, testing environments (i.e. confounding variables), which will be considered by the project-wide user and acceptance evaluation.

3.4. Standardised application of methods

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

3.5. Regulation and site-specific rules

The lack of regulatory alignment across the different European countries, such as differences in the permission granted for the tests, affect the possibilities for achieving consistent results on user testing across pilot sites. For example, there may be variations across pilot sites in terms of the type of drivers that are permitted to be participants (see section 3.1), which provides challenges for comparison across the project. Furthermore, pilot sites may vary in terms of participants' specific or unrestricted roles and permitted activities when operating prototype vehicles. For example, should drivers at a particular site not be permitted to engage in non-driving related tasks during the pilot, this may affect our ability to answer some questions at a project level. The same applies to the willigness to pay questions and TOC-rating.

4. Conclusions

Due to the prototypical nature of ADFs, FOTs and NDSs are not yet possible for automated driving. Therefore, controlled piloting of automated driving functions in real-world traffic represents a valuable first step in understanding how these systems might be used and accepted by members of the public. While real-world piloting may not provide the most comprehensive understanding of the safety and user impacts of automated driving, users' first experiences of ADF in a real traffic context are still valuable, and can inform the ongoing development of these systems. Therefore, it is possible to answer many of the user and acceptance related research questions, so long as caveats are included when results and recommendations are presented.

For this project, the limitations of the planned on-road tests were taken into account prior to data collection, in order to find suitable solutions for achieving meaningful and scientifically valuable conclusions at the end of the project, regardless of the practical limitations that pilot studies with prototype systems pose. Within L3Pilot, a combination of various methods, followed by a merging of results across different testing environments, is the main approach used to handle the variations across pilot sites, and testing environments. This broad methodological setup provides an advantage that is more powerful than the drawbacks and limitations of a single method. For instance, the pilots will not necessarily provide a representative response, for example, for questions regarding willingness to pay, since a large proportion of the sample includes professional (test) drivers. Therefore, especially for this type of research question, where a more representative sample is vital, the on-road tests are supplemented by the results from a large international user acceptance survey. Other complementary data collection methods have also been developed to supplement the data collected as part of the piloting, with the aim of providing a holistic view of users' behaviours with, and acceptance of, ADFs.

In conclusion, we hope this overview illustrates the sorts of challenges and lessons related to collecting data in such real-world pilots. Hopefully, our experiences in L3Pilot will not only be used to answer our project research questions, but also to help future projects, while also deriving recommendations for future on-road tests for L3/L4 ADFs.

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

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

Belletier, C., Davranche, K., Tellier, I. S., Dumas, F., Vidal, F., Hasbroucq, T., & Huguet, P. (2015). Choking under monitoring pressure: being watched by the experimenter reduces executive attention. *Psychonomic bulletin & review*, 22(5), 1410-1416.

Brizzolara, D., Fischer, F., Koskinen, H., Koskinen, S., Heinig, I., Follin, P., Barbier, C., Ojeda, L., Visintainer, F., Fernandez, J., Rämä, P., Rosé, H. (2014). *Report on FOT Operations, DRIVE C2X Deliverable D11.3.* European Commission.

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

ARCADE (2018). *FESTA Handbook. Version 7.* Retrieved from https://connectedautomateddriving.eu/wp-content/uploads/2019/01/FESTA-Handbook-Version-7.pdf

Fridman, L., Brown, D. E., Glazer, M., Angell, W., Dodd, S., Jenik, B., ... & Abraham, H. (2017). Mit autonomous vehicle technology study: Large-scale deep learning based analysis of driver behavior and interaction with automation. *arXiv preprint arXiv:1711.06976.*

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

Gold, C., Berisha, I., & Bengler, K. (2015, September). Utilization of drivetime–performing non-driving related tasks while driving highly automated. In *Proceedings of the Human Factors and Ergonomics Society Annual*

Meeting (Vol. 59, No. 1, pp. 1666-1670). Sage CA: Los Angeles, CA: SAGE Publications.

Hibberd, D., Louw, T., Aittoniemi, E., Brouwer, R., Dotzauer, M., Fahrenkrog, F., ... & Penttinen, M. (2018). *From Research Questions to Logging Requirements: L3Pilot Deliverable D3.1.* European Commission.

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

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

Jamson, S., Kappe, B. & Louw, T. (2014). *Performance Indicators and acceptance analysis plan. ecoDriver Project Deliverable 42.1.* European Commission.

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

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

Li, S., Blythe, P., Guo, W., & Namdeo, A. (2018). Investigation of older driver's takeover performance in highly automated vehicles in adverse weather conditions. *IET Intelligent Transport Systems*, 12(9), 1157-1165.

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

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

Metz, B., Rösener, C., Louw, T., Aittoniemi, E., 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.

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

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

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

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

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.

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

Pudāne, B., Rataj, M., Molin, E. J., Mouter, N., van Cranenburgh, S., & Chorus, C. G. (2019). How will automated vehicles shape users' daily activities? Insights from focus groups with commuters in the Netherlands. *Transportation Research Part D: Transport and Environment*, 71, 222-235.

SAE International. (2018). *J3016-2018: Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles.* SAE International.

Stewart, D. W., & Shamdasani, P. N. (2014). *Focus groups: Theory and practice* (Vol. 20). Sage publications.

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

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

Xiong, H., Boyle, L. N., Moeckli, J., Dow, B. R., & Brown, T. L. (2012). Use patterns among early adopters of adaptive cruise control. *Human factors*, *54*(5), 722-733.