university of groningen / faculty of behavioural and social sciences

Framework for Testing Human Factors and Type Approval

December 2021

Frank Westerhuis[1], Arjan Stuiver[1], Karel A. Brookhuis[1], Casper J. Albers[2], & Dick de Waard[1]

University of Groningen,
[1] Traffic Psychology, [2] Psychometrics & Statistics
Faculty of Behavioural and Social Sciences
Grote Kruisstraat 2/1
9712TS Groningen
Phone: 050-363 6761/ 363 6311

f.westerhuis@rug.nl, a.stuiver@rug.nl, k.a.brookhuis@rug.nl, c.j.albers@rug.nl, d.de.waard@rug.nl

http://www.trafficpsychologygroningen.info

Table of Contents

1. Introduction

Automated Driving Systems (ADS) and Advanced Driver Assistance Systems (ADAS) can, if they function properly, make a significant positive contribution to road safety. Safe interaction between the system and the driver is a prerequisite for this. In the past, guidelines have been provided on what such systems should look like and on the conditions they should meet (Kroon et al., 2019, Souman et al., 2021b). However, guidelines do not guarantee that systems, once developed, actually are safe to use or do not introduce issues that compromise safety unintentionally. Examples of ADAS or ADS being used improperly or failing in conditions when the driver expected the system to work adequately have been published (Gold et al., 2016; Merat et al., 2014). There is a need for a testing framework for behavioural effects of systems that manufacturers place in their vehicles. It is however not evident how, and even if, such an assessment framework can be developed. The present report explores the possibilities and obstacles for developing such an assessment framework for ADS and ADAS. The focus for such an assessment framework is on the interaction the system has with the driver.

A difference can be made between driving with ADS/ADAS, without ADS/ADAS, and a transition between these two phases. Especially in the case of ADS where the driver may be (partially) out of the loop, the transition could be a difficult and potentially dangerous transfer of control. The driver needs to be aware of his or her surroundings and have an appropriate understanding of how the system functions. Creating the (situation) awareness takes some time, while depending on the traffic environment this will take more, or less time. For an assessment framework, these uncertainties are important concepts. Many of the safety-related factors in the end may be described in terms of whether the driver understands what the system is doing (has the appropriate mental model), what his or her role at that moment is and will be in the near future (related to transferring control) and knowing the situation on the road (situation awareness). These are the fundamental concepts to understanding the interaction between driver and system. It must be noted, however, that there is more to behaviour of drivers than just these concepts. There are other concepts, e.g. those related to behaviour, such as drivers misusing or wilfully abusing a system (using it for what it is not intended for). An assessment framework needs to determine whether the issues associated with these concepts are properly addressed by or in the ADS/ADAS and thus whether the interaction between driver and system is safe.

To indicate what is necessary for the development of an assessment framework, in this document these concepts are described and it is explored how they can be used to test the safety of the interaction between driver and ADS/ADAS, and the resulting behaviour of the system/driver as a whole. Therefore, for each concept, the important factors related to safety are described, after which, if possible, the impact it may have on the development of an assessment framework is given.

The goal of an assessment framework should be determined more precisely before designing, developing, let alone implementing it. In general, it can be said that the main focus of the assessment should lie on safety. Factors such as comfort may have an indirect impact on safety, but safety itself should be the focus. Although accidents, injuries, and fatalities are not directly measurable within an assessment framework, the overall effect of assessment should have the reduction of accidents, injuries, and fatalities as main focus.

1.1. Structure of document

This document describes several important concepts (e.g., safety, interaction, situation awareness, transition of control, statistical implications for assessment, etc.) and what they mean for the development of an assessment framework for the safety of ADS and ADAS. Below is a list of sections in which these concepts are described. Note that there is a large interconnectedness between these concepts and that therefore some relations and overlap in topics between sections exist.

For each concept the impact for the framework is given when possible. In the final chapter a conclusion about the feasibility of developing (parts of) the assessment framework is reported. The main concepts that are described in this document are respectively:

- Safety, interaction, and cooperation
- Safety assessment
- Human-Machine Interface (HMI)
- Situation Awareness
- Mental Models
- Transfer of Control
- Testing with participants
- Power analyses: calculating the minimum required sample size
- Determining criteria
- Test methods
- Context: exceptional or critical situations
- Other considerations

2. Interaction

2.1. Safety, interaction, and cooperation

Because of the ongoing developments of ADAS and ADS, driving increasingly becomes a shared task between a human driver and in-vehicle automation. This means that driving is no longer the sole responsibility of a human driver, and cooperation between the human driver and in-vehicle automation will be required. Drivers now (also) have to interact with their in-vehicle automation systems, being (largely) redeemed from controlling the vehicle and interacting with other road users. As with any interaction, there are two parties involved, but in this case, one of those is designed by a (vehicle) manufacturer. This means that in-vehicle automation manufacturers not only have to make decisions about how automation independently operates a vehicle, but also about how it interacts with the human driver.

In November 2021, the Society of Automotive Engineers released the Automated Vehicle Safety Consortium (AVSC) Best Practice for Evaluation of Behavioral Competencies for Automated Driving System Dedicated Vehicles (ADS-DVs; AVSC, 2021). This document provides guidelines to evaluate system behaviour of ADS within its Operational Design Domain (ODD). In short, they conclude that evaluating system functioning (competence) requires defining specific (system) behaviour, measuring this behaviour within a specific context, applying metrics, and comparing these with acceptance criteria and ODD-relevant thresholds. Even though these guidelines are useful for the evaluation of ADS, they do not provide concrete thresholds for 'sufficient' competence and evaluation is limited to individual system behaviour. In practice, however, the interaction with the ADS/ADAS is a part of all the interactions between vehicle and human driver. It is therefore important that the system's interaction is evaluated within the context of all interactions with all available types of (support/automated) systems in the vehicle. Indeed, actual vehicle driving behaviour on the road is determined by emergent behaviour from the human and automated drivers, and the quality of this combination will decide the eventual safety of a vehicle on the public roads. The assessment of ADS/ADAS should be made with the driver included, as a complete system where vehicle, ADS/ADAS, and human work together. Emergent behaviour from this combination (behaviour that can only be expected during the interaction of driver and systems) should be evaluated. Moreover, because these interactions include all possible in-traffic situations, they are not limited to one system's ODD as decided by the manufacturer. Systems can often be activated outside their ODD. It should be clear to a driver that the system is not intended for use there and that it can be unsafe. Even better would be to make it impossible to use the system outside its ODD. However, currently most systems can be used outside their ODD and it is up to the driver to be aware of that. For an assessment framework, a decision must be made about how to deal with the usage of systems outside their ODD.

Framework

Because terms such as 'interaction' and 'cooperation' are very broad and generic, they are difficult to use as part of an assessment procedure in practice. The main aim of this document is to describe the obstacles for the development of a framework for evaluating the quality of interaction processes between human drivers and in-vehicle automation. Relevant aspects that constitute these interactions are described for the different obstacles.

During an exploration into shared driving tasks for the Dutch National Road Authority Rijkswaterstaat (RWS), Petermeijer et al. (2021a; 2021b) defined requirements for successful cooperation between in-vehicle automation and the human driver, which were divided into seven dimensions: (1) compatible goals, (2) shared mental models, (3) shared situation awareness, (4) distribution of responsibility, power, and authority, (5) adaptability, (6) conflicts, and (7) communication. The project gave a description of what an assessment framework could look like and which aspects would be important and how they might be potentially measured. They described how external changes influence both the driver and the system, and defined indicators of quality of the shared driving task. According to Petermeijer and colleagues, the key concept is interaction and cooperation between driver and system, which is the basis of their framework. In the end however, they had to conclude that to implement such a framework, a very specific description of criteria and measurements is in many cases difficult to give and complexity and multiplicity of criteria and aspects to assess is very ambitious if not overambitious. In the final section of their report, Petermeijer et al. (2021a; 2021b) provide a preliminary list of aspects and issues that will be encountered when implementing such a framework. To actually determine the feasibility of an assessment framework, those issues need to be addressed and described in more detail. In this document, a more complete set of issues is described and how they impact the feasibility of an assessment framework.

When analysing the issues that arise when trying to develop an assessment framework, as we have done in this work, it becomes clear that the implementation of such a framework is quite difficult in reality. On the one hand, with a very generic approach, criteria and measurements remain at a theoretical level and are difficult to apply. On the other hand, with a very specific approach, criteria and measurements quickly multiply in number and become so specific that they are only applicable in very specific situations. It is clear however that there are aspects of ADS and ADAS that may be evaluated to be safe or unsafe. Combining the knowledge from Petermeijer et al. (2021a; 2021b), the AVSC best practice document, other projects, and realising that there are many issues unresolved leads to the conclusion that there are aspects that are very important for an assessment framework, which can be discussed separately and will each lead to important implications for the assessment framework. In this document, relevant aspects for evaluating what the interaction between driver and system must achieve and how to measure this will be discussed and the impact of the issues surrounding these aspects are given as points of attention for future work on the assessment framework. In the next sections important topics such as transition of control, situation awareness, risk assessment, testing with participants, statistical impact of the assessment etc. are discussed in separate sections, although links from one section to the next are made.

2.2. Safety assessment

The goal of evaluations of any (automated vehicle) type approval should be the general safety on the public road. For in-vehicle interactions specifically, positive evaluations should guarantee that human drivers are able to use and interact with in-vehicle automation in a safe way. The crucial question that follows is: what is 'safe'? It could be that, for example, 'safe driving' means 'driving without accidents'. If so, this should mean that testing is based on evaluating that the likelihood automation has on the occurrence of traffic accidents, approaches zero (i.e., 95%, 99%, 99,9% safety?). Even though it is beyond the scope of this document to provide such a result, a threshold should eventually be set on what is 'sufficiently

safe', before safety effects of any automated system, human driver, and these two combined can be evaluated. This is an ethical matter: until 100% safety is the attainable threshold, any other percentage means that the remaining risk should be determined and accepted.

2.3. Human-Machine Interface (HMI)

The main interactions between a human driver and vehicle automation takes place through the Human-Machine Interface (HMI). With this interface, in-vehicle automation can send information about its availability, (operational) status, and about the driving environment, for example. This information is mainly presented to the human driver by means of visual, audio, and/or haptic signals. Vice-versa, automation can also receive information from the human driver. A human driver may provide input to the system intentionally (e.g., via physical controllers, buttons, voice control) or unintentionally (e.g., a system may observe human behaviour such as eye movements, steering wheel touching and/or movements).

For a human driver, the need for information differs depending on the level of automation: during manual driving, drivers tend to prefer receiving information about the driving task while during automated driving, mostly information about the automation status is preferred (Beggiato et al., 2015). This means that the functional requirements for an HMI can change while driving and will be different based on the system functions that are used. Souman et al. (2021b) defined guidelines that might be used to evaluate in-vehicle HMIs of different ADAS and ADS functions. These guidelines are founded on basic HMI guidelines and are further specified for (1) information functions, (2) warning functions, (3) assistance functions, and (4) automated functions. A clear distinction between levels is not possible, guidelines for information functions largely apply to warning, assistance and automated functions as well. A warning is also a form of information and assistance and automated functions also provide information and warnings. In other words: guidelines for relatively simple systems (lower level) are largely applicable for more complex systems (higher level) as well.

In-vehicle HMI guidelines may be used to evaluate whether a human driver perceives and correctly understands signals and/or messages from in-vehicle automation and how conveying the information can be improved (Souman et al., 2021b). A large limitation is that HMI is a broad concept: there are many guidelines ranging from fairly general (e.g., overall display size) to very system-specific aspects (e.g., system X menu settings Y). Higher level systems (ADS) are becoming increasingly complex, in particular compared to more 'conventional' systems, which also results in a need for more (specific) evaluation guidelines. For this reason, Souman et al. (2021c) concluded that most variables still require very specific definitions, operationalisations, and test criteria before they can be measured reliably and consistently for evaluation purposes. Based on the extensive number of studies performed on in-vehicle HMI (e.g., Campbell et al., 2016; Campbell et al., 2018; Kroon et al., 2019; Schömig et al., 2020; Souman et al., 2021b), consensus about general standards for HMI elements should be reached, also with regard to a standardised HMI-testing protocol as a whole.

Distraction

An additional factor with regard to signal perception concerns distraction. Regardless of whether information and/or signals are perceived and correctly interpreted, it is important that these interactions do not distract the driver from the main driving task. Even if attracting attention is the main purpose of a signal (e.g., a warning), it should be prevented that these

signals attract attention to such an extent (too long) that the driver does not pay enough attention to the road anymore (i.e., reducing situation awareness). Distraction can be determined with eye-tracking, for example, by measuring the amount of time that a driver takes his or her eyes off the road due to a specific signal (Khan & Lee, 2019). In addition, unreliable messages/warnings (false alarms) can also be distracting, potentially leading to driver annoyance and turning off automation altogether, discarding all potential safety benefits while driving.

Impact for framework

HMI testing should be performed to determine the perceivability of signals and/or messages from in-vehicle automation, and how well these are understood by the human driver (Souman et al., 2021b). This can be performed with traditional methods, based on the extensive literature about HMI guidelines.

Individual HMI elements can be tested: which information does the automation show, how is the timing, is it well-detectable, understandable, necessary, clear, reliable, and standardised, for example (see e.g., Kroon et al., 2019, Souman et al., 2021b). The problems that arise with this type of testing is that many times a rating will depend on the type of information presented. This is even more the case for functions that are automated to a higher degree (Souman et al., 2021b). Indeed, the more automation of the driving task is incorporated in a system, the more numerous and complex the guidelines for their functions become. To evaluate the "higher-order" guidelines, system and/or function-specific testing becomes required, which greatly expands the number of factors that should be considered. Reaching consensus about selection and prioritization of the most important functions and guidelines is therefore required (Souman et al., 2021b; 2021c). Furthermore, guidelines are mostly suggestions about what could or should be done to improve an HMI. Apart from some exceptions, most guidelines do not specify criteria about when a system scores 'sufficient', which also highlights the necessity to form and reach consensus about these criteria (Souman et al., 2021c).

3. Situation Awareness, Mental Models, and Transfer of Control

3.1. Situation Awareness

Driving is a dynamic task that requires monitoring the environment and operating within that environment at the same time. A driver's awareness of the environment (and itself) is defined as Situation Awareness (SA; Endsley, 1995).

According to Endsley (1995), SA is achieved in three levels: (1) perceiving individual elements in the environment, (2) combining these elements into one holistic comprehension, and (3) making predictions of the future state of these elements. All three levels require cognitive processing time to complete. In terms of driving, SA traditionally consisted of monitoring the (road) environment, other road users, and anticipating these by operating the vehicle (i.e., driving), all performed by the human driver. Since the introduction of information systems (e.g., navigation), human drivers also rely on information about the road environment that is provided by these systems (e.g., routes, speed limits, traffic jams). In addition, with ADAS and ADS, the role of the driver can change from full operator of the vehicle to (partial) supervisor or even passenger, all possible within one drive. This has vast implications for SA, because the more information is perceived through a system and/or the more aspects of the driving task are performed by an automated system, the less it is necessary for a human driver to perceive and stay aware of the surrounding traffic situation. Currently however, the vehicle in relation to the driving environment should still be monitored. When considering automation up to SAE level 3, which means the driver is expected to keep monitoring the driving task, automation still has limitations and situations it cannot handle. For this reason, it is important that a driver not only remains aware of the driving situation with these systems, but also of which tasks the automation is performing (mode awareness). Moreover, if the driver is allowed to perform Non-Driving Related Activities (NDRAs; SAE level 3 or higher), the driver should be given sufficient time to (re)build SA before having to make decisions or operations that are important for driving safely.

Although Situation Awareness is a very useful concept that can help identify aspects of the interactions between humans and automation, it is a continuously changing concept that can comprise a very large part of the human understanding of the situation. As a term, it is useful to understand cognitive processes during full manual driving as well as during (supervised) automated driving. It is, however, difficult to measure SA outside of experimental settings and, perhaps more importantly, it is very difficult to provide criteria for what is sufficient SA. Indeed, SA should not be too low, but what is sufficient differs for every system and context. For information purposes, a driver should be able to perceive the information provided by any system and interpret this correctly. For ADAS and ADS, it is also important that SA remains sufficient to enable a driver to intervene in a critical situation to prevent a conflict or accident.

Impact for framework

It can be concluded that in-vehicle automation may impact a driver's Situation Awareness (SA). The type of information and the way of presenting this information to the driver is important: assessments should evaluate whether automation-provided information is relevant, reliable, perceivable, understandable, and contributes to a driver's SA (see also section 2.3). The necessity to (re)build SA also influences the time needed to safely take over control from automation. Evaluations should determine whether SA can be kept at a sufficient level while

using in-vehicle automation, or that sufficient time (budget) is provided for (re)building SA before a Transfer of Control (ToC) takes place. For this reason, the framework could assess SA by means of presenting critical driving scenarios in which the automation fails and a human driver has to take over from automation: a so-called 'Transfer of Control' scenario, which will be discussed in section 3.3.

3.2. Mental Models

A driver's mental model of a system (such as ADAS or ADS) is a personal, internal representation of the system's functional behaviour. In other words: a driver's mental model contains this individual driver's knowledge and expectations about a system's operation and limits (Johnson-Laird, 1994). Mental models play a major role in forming and shaping the behaviour while driving with ADS, for example. Humans hardly have a complete and correct mental model of a system (Souman et al., 2021a). This can lead to problems when interpreting what a system does and predicting what it will do. However, a perfect mental model is often not necessary to use a system correctly. For this reason, Norman (1983) described mental models as "messy, sloppy, incomplete, and indistinct knowledge structures".

When starting to drive a vehicle with ADS for the first time, driver behaviour, acceptance, and trust develop and change over the different phases of experiencing the system (Dunn, Dingus & Soccolich, 2019). At the start (the first phase), trust in ADS could be low if the driver has little or no experience with ADS. After some time familiarising with the system, gradually moving to a post-novelty operational phase, driver behaviour should begin to adapt to the system's operation. In this phase, a mental model of the system may gradually develop while trust grows, including becoming familiar with the limitations of the system. While interacting with a system, the mental model evolves further while also being shaped and/or restricted by earlier experience with comparable systems and/or the (technical) background of the user (Souman et al., 2021a).

Over time and with increasing experience, gained trust could lead to overreliance on ADS and underload of the driver, the latter because there may be little to do for the driver except monitoring the system's operation (Souman et al., 2021a). Also over time, people might forget details of a system: in particular details regarding system features or limitations that are rarely experienced. Another problem is that, with prolonged experience, drivers may engage in secondary tasks if that is in line with their working mental model of the system's operation, as Dunn, Dingus & Soccolich (2019) demonstrated. They found that when people drive more often and longer with ADS, the proportion of eye glances (>2s) on secondary (non-driving related) tasks increases. This could mean that drivers feel increasingly confident that their mental model is accurate and that the automation is able to successfully control the vehicle laterally and longitudinally. Moreover, it is expected that drivers of automated vehicles may be even more inclined to engage in non-driving related tasks (e.g., watching videos or sleeping) when vehicle automation develops towards SAE level 4 (De Winter et al., 2014).

In principle, a manufacturer also has a "model" of its users (i.e., the human drivers) and their mental models. In practice, this will likely lead to different systems and interactions in different vehicles, which might well lead to unexpected and unwanted differences. If so, experience with automation in one vehicle will usually have positive effects (such as recognition of, and confidence in actions by the system) when confronted with new automation in another vehicle,

but this might have negative effects as well (such as unexpected, different communication by the system). Taking all this into account, the question is how to measure a driver's situation awareness, in other words: how to gather information about perception, understanding, explanation, and predictions of the situation by the driver. There are several possibilities to collect these data, including (driving simulator) experiments, verbal protocols of reports by the (learning) driver, interviewing the driver in various stages of the learning process, and questionnaires after having used a system. A well-known and widely accepted form of studying the generation of SA is the Situation Awareness Global Assessment Technique (SAGAT), as developed and introduced by Endsley (1985).

There are limitations for assessing a person's true mental model (Souman et al., 2021a). According to Endsley (2000), mental models mainly contain static, long-term stored knowledge. This differs from the (earlier introduced) concept of Situation(al) Models, which mainly represents knowledge of the current state of a system (Endsley, 1995). Because Situation Models are closely related to Situation Awareness (SA), the Situation Awareness Global Assessment Technique (SAGAT) can be used to measure both SA and Situation Models. With this method, a (driving) simulation can (suddenly) be stopped at different moments to ask the human driver specific questions about their SA or Situation Model at that exact time. The outcome is however difficult to interpret because it may be incomplete, evolving, unstable, and without clear boundaries (Souman et al., 2021a). Because mental models should "contain" a driver's estimations of actual system functioning and allow generating explanations of observed states, it should minimally reveal the following elements to enable "measuring" the mental model: (1) a set of (production) rules (IF … THEN …) and (2) a set of state variables (e.g., the mode of an ACC: off, stand-by, active; or set speed). Most of the latter information should be continuously communicated to the driver via displays (visual or other modalities) and be directly available for the user. The rule sets and state variables combined should make it possible for the driver to not only observe current system states, but also clarify system functioning, and predict future states of the system (Souman et al., 2021a).

Several studies shed light on how aspects of mental models of ADS may have measurable effects on driving behaviour (Souman et al., 2021a). For example, Orlovska et al. (2020) found that drivers who use ADS longer and more frequently show more confidence and trust in ADS compared to drivers who use such systems less often. Furthermore, the people who used ADS less often only tended to use it mainly in well-known situations and were reluctant to use it in unknown situations. Based on a driving simulator study in which an ADS would eventually fail to cope with a presented driving situation, Kircher, Larsson, and Hultgren (2014) found that drivers who expected that a system was unable to handle the (traffic) situation took over control from automation before it reached its limits. This implies that their mental model "noticed" the upcoming limitations of the system and enabled the driver to anticipate.

Impact for framework

Interactions between human drivers and in-vehicle automation are very dependent on mental models. Human drivers should know what automation can and *cannot* do (its goal), what it is currently sensing, *not* sensing, doing, *not* doing, and what that means for the human driver him or herself (i.e., the goal of automation, its perception, the division of labour, and the resulting responsibilities and behavioural adjustments). For this reason, it is recommended to include mental models in the assessment procedures for human-automation interactions. In principle, a major factor should be whether the interactions provided by the automation (i.e., adequate information and feedback; HMI) allow the human driver to form, maintain, and calibrate a

‘sufficient’ mental model of the automation to drive safely. There are several factors that should be considered for these measurements.

When driving with ADS for the first time, in the earliest learning phase of automation, drivers’ mental models are unstable, volatile, and subject to changes and refinements until the driver (begins to) feel comfortable with the operation and information provided by the automation. In this phase, the essential question is to what extent the interactions provided by the automation contribute to forming a 'correct' (first) mental model by the human driver, and what could be implemented further by the manufacturer to support this sufficiently. After the learning phase, when drivers feel confident and comfortable with their ADS, the mental model becomes steady and is hard to adapt. For this reason, it should be taken into account whether a system is new for a driver, or that he or she already has experience with it. A system could provide more or different (forms of) information to novice or experienced drivers, for example. Either way, short-term and long-term effects of the automation on mental models should be evaluated.

Finally, the more drivers become confident with operating in-vehicle automation, the more they tend to forget its limitations. Sufficient awareness of limitations is however vital to be conscious of all the looming (infrequent) dangers, particularly if these seldomly occur. As confident and comfortable drivers have been found to become more inattentive, the framework should include assessments of attentiveness (SA and/or Situation(al) Models) and/or reactions to (infrequent) dangers. Furthermore, because adequate mental models go hand in hand with trust in the system, and low trust may easily lead to low use of the system, the effects of mental models on system use, disuse, misuse, and abuse should also be considered.

3.3. Transfer of Control

In a ‘traditional’ driving context, the human driver is the only one who has full control over the vehicle at all times. Even though it is not applicable for all in-vehicle systems, several forms of ADAS and all forms of ADS allow automation to take over (parts of) the driving task if requested or needed: this is called the Transfer of Control (ToC). This means that within one drive from A to B, the driving task can be performed by two entities: the human driver and/or the automation and this can change while the vehicle is moving. Lu et al. (2016) defined six types of control transitions (ToC types):

1. Optional Driver-Initiated Driver-in-Control;
2. Mandatory Driver-Initiated Driver-in-Control;
3. Optional Driver-Initiated Automation-in-Control;
4. Mandatory Driver-Initiated Automation-in-Control;
5. Automation-Initiated Driver-in-Control;
6. Automation-Initiated Automation-in-Control.

A human driver may consciously take control from (ToC type 1) or give control to (ToC type 3) automation voluntarily if both automation and/or human driver are able to manage the current driving task without any problems. The next options are more critical: if a human driver believes that either the automation or the human driver is not able to manage the current driving task, he or she will mandatorily take control from automation (ToC type 2) or give control to the automation (ToC type 4), respectively. In all these cases, the human driver is supposed to be aware of what the automation is capable of and what the automation is not capable of, also in relation to the human driver's own driving capabilities and state. It is also possible that

automation initiates a transfer of control, either by giving control to the human driver (ToC type 5) or taking control from the human driver (ToC type 6). Automation-initiated ToCs are by definition mandatory, regardless of the reason. For example, the automation may detect a driving context it is not designed to handle (i.e., outside of its ODD), or an internal system failure, and request the driver to take control of the vehicle. Vice versa is possible as well: an automation may detect driver inattention, inability, or even unconsciousness, and take control of the vehicle to prevent an emergency (Lu et al., 2016).

From the literature, it is known that transferring control takes time, which not only depends on the type of transition, but also on the state, capabilities, and responsibilities of the driver and/or the automation, and the road environment at hand (see e.g., Lu et al., 2016; Gold et al., 2016; Naujoks et al., 2018). In principle, for all types of transitions, there should be sufficient time available to make sure that either the human driver or the automation builds the necessary situation awareness and takes over full control before reaching one's operational limits: the so-called 'time budget' (Gold, Happee, & Bengler, 2018). How much time exactly is 'sufficient' for every situation is currently unclear and difficult to determine. From a literature review, ToC time budgets range from 0 up to 30 seconds, while 'required' take-over times, which includes perceiving a takeover request and preparing both mentally and physically to drive the vehicle (perception + reaction time), run from 1.14 up to 15 seconds (Eriksson & Stanton, 2017). Both types of timings vary greatly between transfer types: if the used automation allows drivers to perform Non-Driving Related Activities (NDRAs), for example, more time budget should be provided for transferring control compared to systems that do not allow NDRAs (see e.g., Yoon et al., 2021). In addition, it is also known that driving behaviour is affected directly after a human driver has retaken control of a vehicle (Merat et al., 2014). For example, driving parameters such as Standard Deviation Lateral Position (indicating swerving; SDLP) and Mean Lane Position may increase temporarily, depending on the workload of the human driver. Even though the peak in safety critical values occurs approximately 10-20 seconds after the actual transition, it may take up to 60 seconds for driving behaviour to fully stabilize (Merat et al., 2014; Melnicuk, Thompson, Jenning, & Birrell, 2021).

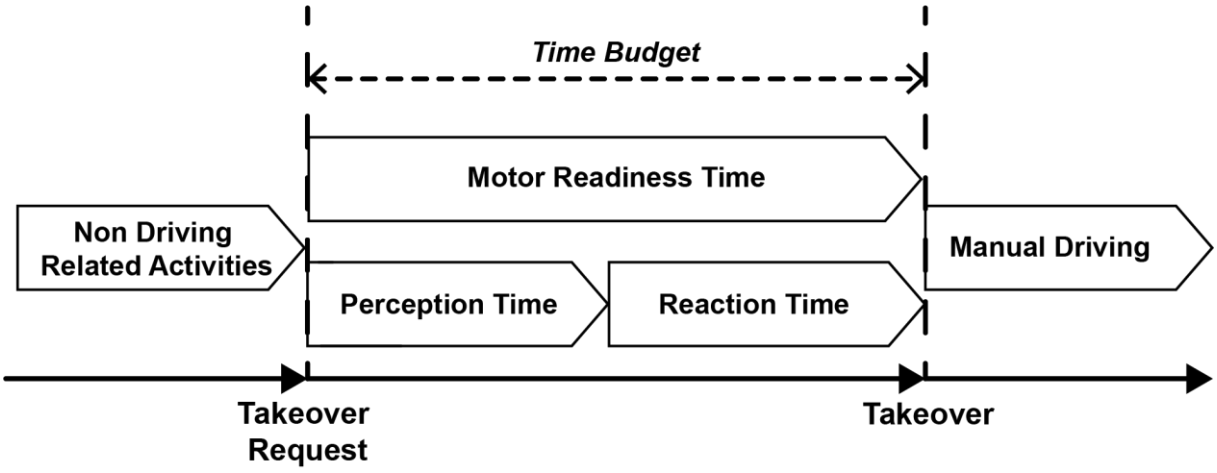

Figure 1: Overview of the ToC process. Adapted from Yoon et al. (2021).

Impact for framework

For the framework, the abovementioned findings indicate that the quality of transferring control may be determined by (1) the amount of time that is allowed to transfer control (time budget), (2) the amount of time that is required for the human driver to prepare for taking full control (perception + reaction time), and (3) the effects on driving behaviour after taking over control (manual driving; see figure 1). In general, the time budget should be appropriate (i.e., not too short, but also not too long) and the mean required 'readiness' time should be as low as possible. During post-transfer manual driving, driving behaviour should be similar to the pre-transfer phase or to ‘normal’ driving of the human driver. If values do increase, these should stabilize within safe boundaries as fast as possible.

In conclusion, due to the large variance in transition types, drivers, types of automation, and driving contexts, the possibilities for testing ToCs are rather large and selection of safety-critical contexts seems appropriate. One norm for ToC is not available and it is also expected that this will not be sufficient for testing all possible interactions while driving a vehicle with automation in real traffic. There is an ISO standard (ISO/TR 21959-2) available for testing ToC performance that suggests scenarios, measurements, and test environments to implement in the evaluation framework. Mean reaction time, post-transition driving behaviour (e.g., vehicle lane positioning and swerving), and collision avoidance are variables that may be used to determine ToC safety. Eventually, consensus should be reached about 'acceptable' ToC times for safety-relevant driving scenarios and even though 'more time' may serve as a rule of thumb, it should be specified what the safest time budgets really are.

4. Behavioural assessments

4.1. Testing with participants

Even though it may not be the first step in the assessment, part of testing could be based on participants. In that case, the first question that should be asked is “which participants should be included?” In principle, this could be anyone who holds a valid driving licence. Within this group, however, there is great variation in knowledge, skills, and experience, for example. Indeed, not only variables such as age, (basic) knowledge level, driving experience, and personality should be considered, but also perceptual, cognitive, and motor abilities are factors that determine each individual driver, their (personal) requirements for vehicles, in-vehicle automation systems, and interactions (see e.g., Borowsky, Shinar, & Oron-Gilad, 2010; Falkenstein, Karthaus, & Brüne-Cohrs, 2020; Larsson, Kircher, & Hultgren, 2014; Starkey & Isler, 2016).

The consensus is that driving skills differ between people and also change within people because of ageing, for example. Indeed, older drivers may experience decreasing cognitive, perceptual, and/or motor skills compared to younger drivers (see e.g., Anstey & Wood, 2011; Falkenstein, Karthaus, & Brüne-Cohrs, 2020; Depestele et al., 2020). Older drivers, however, tend to have more driving experience and use this to anticipate traffic events and drive more defensively (e.g., adopting longer headways; Andrews & Westerman, 2012). Even though in theory, younger drivers have “better” perceptual and motor skills, they have less driving experience, less experience with detecting hazards (Borowsky, Shinar, & Oron-Gilad, 2010) and are also more likely to exert risky behaviour (Simons-Morton et al., 2015). Regardless of age, a driver’s skills and capabilities may also change within different contexts of a drive and the resulting cognitive load on the driver (Biernacki & Lewkovicz, 2021).

Education

Upon receiving a driving licence, people are expected to have the required knowledge and basic skills for safe driving. What is learned exactly, however, depends on the moment of examination. In particular regarding the more recent developments, drivers will differ in terms of knowledge, skills, and experience with automotive (support) systems and particularly new systems will be unknown for many drivers. In addition, there is great variance between drivers: many may have used multiple systems often, while others might never have used one single support system, even if it is installed in their vehicles (Harms, Bingen, & Steffens, 2020). Also, the information and education that people receive when first being introduced to an in-vehicle automation system differs and influences drivers' prior-expectations and the way they use such systems (Souman et al., 2021a). In practice, many people only receive a limited amount of information about their in-vehicle automation systems such as the vehicle owners' manual (Souman et al., 2021a). Other sources include individual in-vehicle information messages provided by the automation, or more sophisticated in-vehicle education by means of a tutoring system, for example (see e.g., Boelhouwer, 2021).

Impact for framework

For the framework, it is advised to incorporate assumptions about which drivers will be included in the testing procedure to form a representative sample. It is useful to consider drivers' prior-knowledge, skills, and experience before performing an actual test. Which level of prior-knowledge about in-vehicle automation is expected; are drivers aware of the limitations of systems and/or the used technologies, are relevant questions to consider. Furthermore, it could be argued that including only the 'low skill' drivers, i.e., drivers that are legally allowed to drive but score relatively low on cognition, perception, and/or motor skills, could provide different test results compared to including the 'average' or 'strong' drivers. What constitutes a relatively 'low', 'average', or 'high' skilled driver is however difficult to determine. Also, assessments should determine to what extent the information that is provided by car manufacturers (e.g., documentation/manuals, in-vehicle information, tutoring systems) supports driver education and forming a driver's mental model. In principle, the most preferred option would be that a design leads to safe behaviour without having to educate or inform people (i.e., an intuitive system).

4.2. Power analyses: calculating the minimum required sample size

Once it is decided which instruments are to be used and what measurements are to be collected, one has to decide the sample size for the study. When this sample size is too low, no accurate conclusions can be drawn from the data, rendering the data collection effort void. On the other hand, collecting more data than minimally sufficient can be seen as a waste of resources.

Whether one is interested in the performance of an automated lane keeping system or in the safety of a fully self-driving vehicle, the steps behind the statistical power calculations are surprisingly similar.

In its essence, the key question is: "How large should the sample size be in order to be able to assess the safety of the system with sufficient accuracy?"

Whatever the context, the study will pan out as follows: in total n tests are being performed, yielding test scores x_1, …, x_n. A higher test score indicates more safety, and we need to know whether or not the average test score exceeds a certain threshold. In case x is a measure of unsafety, simply reverse all the arguments in the following.

Oftentimes, we can safely assume that the test scores stem from a normal distribution around a certain mean μ: $x \sim N(\mu, \sigma^2)$. This assumption can be made even in case the distribution isn't normal. When the sample size is large enough, the central limit theorem implies approximate normality; and the consequences of non-normality are much smaller than often anticipated (Ernst & Albers, 2017). Furthermore, in some situations more sophisticated models can be used. These models generally have higher power (Cundill & Alexander, 2015), so the calculations here will provide an upper bound for the required minimum sample size.

What value is 'safe enough' should be decided upon prior to data collection. This threshold value can be denoted by τ. This is a decision that has to be made based on substantive arguments; i.e. by taking the value that one needs to obtain a driver's license as a threshold.

In a standard statistical testing approach, one would want to test the null hypothesis H_0: $\mu \leq \tau$ versus the one-sided alternative hypothesis H_1: $\mu > \tau$. However, null hypothesis significance testing is set up to reject the null hypothesis in favour of the alternative. The goal here is not necessarily to reject H_0 in favour of H_1, as this would imply proving that the system is safer than the threshold. The goal should be to prove the system is not more unsafe than the threshold; or that it is at least as safe as the threshold value.

This can be done via so-called equivalence tests, also known as non-inferiority tests (Rogers et al., 1993; Snapinn, 2000). In essence, the procedure is as follows. First, a second threshold, δ, needs to be set with interpretation "If we can prove that the safety μ is at least $\tau - \delta$, then the system is safe enough". This yields the statistical hypotheses

H_0: $\mu \leq \tau - \delta$

H_1: $\mu > \tau - \delta$

From this notation, it is clear that the specific values of τ and δ aren't directly relevant, all that matters is their difference $\tau - \delta$. The number of parameters can be reduced by defining $\delta^* = \tau - \delta$, and only working with this difference. The interpretation of the parameter δ^* needs to be done relatively to the mean (μ) and standard deviation (σ): $(\delta^* - \mu)/\sigma$, that is, the number of standard deviations difference from the mean μ is interpreted in the same way as Cohen's effect size measure d (in our case δ^*). This effect size can be interpreted as standardized mean difference: it quantifies the number of standard deviations that δ^* and μ are apart from each other.

Apart from the value $(\delta^* - \mu)/\sigma$ as standardised effect size, two additional values are needed for a power calculation: α and $1 - \beta$. The value α represents the significance level (which is usually set to .05 in scientific practice): if H_0 is correct, the probability of falsely rejecting H_0 (and, thus, claiming sufficient safety for the system) is α. Reversely, $1 - \beta$ is the power, which is the probability of incorrectly *not* claiming sufficient safety, even though H_1 is correct.

For the computations, statistical software is needed. The choice of software is (largely) arbitrary. Commonly used software packages include G*Power (Faul et al., 2007) and the R package pwr (Champely, 2000). For example, in R the computations for $\delta^* = .2$ (d in the commands), $1 - \beta = .8$ and $\alpha = .05$, the commands to get the sample size are:

```
library(pwr)
pwr.t.test(d=0.2, power=.80, sig.level=0.05, type="one.sample",
alternative="greater")
```

and this gives as output `n = 155.9256`. Thus, for this example, the sample size must be at least $n = 156$.

Figure 2 shows the power calculations for two choices of alpha (1% and 5%), four choices of $1 - \beta$ (80%, 85%, 90%, 95%) and different effect sizes. For standardised effect size $\delta^* = .2$, the resulting minimal sample size for the six combinations of α and $1 - \beta$ range from 156 to 397. For smaller effect sizes, the minimal sample size quickly escalates. For instance, for $\delta^* = .1$ it ranges from 620 to 1580. Thus, the more accuracy that is desired (to find a smaller effect size), the more information one needs to collect.

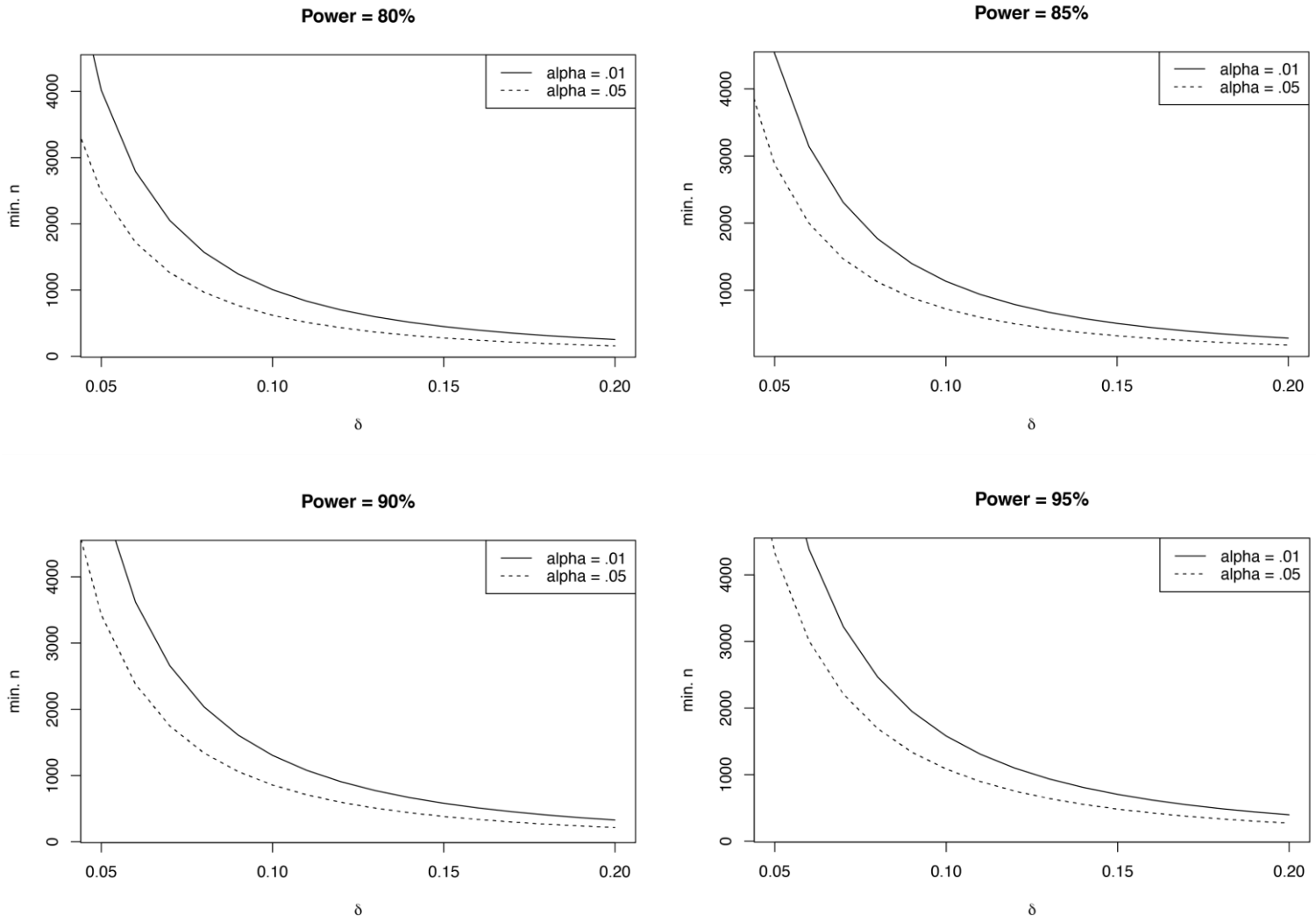

Figure 2: Power calculations for two choices of alpha (1% and 5%) and four choices of 1 – β (80%, 85%, 90%, 95%).

Whether it is feasible in practice to run hundreds, or perhaps, thousands of tests for a single automated system improvement, remains to be seen. This requires a decision based on substantive arguments.

The main challenge in the power analyses lies in a proper justification of the choices for α and β (cf. Lakens et al, 2018), which need to be made on substantive arguments. Parameter α is concerned with the false-positive rate: if this value is set too high, unsafe ADAS/ADS could be admitted to the roads too easily, which clearly is a risk for road safety. Perhaps α = .05, which is conventional in many scientific studies, is too high in this setting, and α = .01 or even α = .001 is better.

Parameter β is concerned with the false-negative rate: setting this value too high hinders safe ADAS/ADS to be admitted to the road network. Although this does not negatively affect road safety, it does have financial consequences for the manufacturers.

Ideally, both types of error are set (very) low. However, the consequence of this, in turn, is that a much larger sample size is required – which could be so large that it is infeasible or too expensive to carry out the tests. Thus, α and β need to be set in such a way that a balance is found between road safety, the financial interests of ADAS/ADS manufacturers, and the feasibility of the ADAS/ADS safety tests.

Impact for framework

There is a need to find substantive arguments regarding the minimal power level, the maximum level of significance, and the minimal standardised effect size. The challenge lies in finding and defending these values. The statistical calculations themselves are straightforward. Choosing too liberal values might yield feasible sample sizes, but at the same time too lenient targets (e.g., having the false positive rate too high). Choosing more conservative values will correspond to acceptable targets, yet might lead to unrealistically large sample sizes.

4.3. Determining Criteria

The need for criteria to decide when drivers' reactions to the system's operation are still acceptable or not, imposes a heavy load on examination bodies and inspection institutes. The need to draw a solid dividing line for acceptability, of the time a driver needs to take over control, for instance, urgently asks for standards with respect to safety. Let alone defining thresholds for constructing laws in this respect. One example where criteria have been developed and consensus has been reached by different stakeholders, is alcohol level and driving capability, where research has resulted in standardised criteria that are used in practice.

After World War Two, motorized vehicles quickly became common for everyday use. The same holds for drinking alcohol, not only at home but in traffic as well, which led the police (in the USA) to the suspicion that the alarming rise in traffic accidents was partly due to driving under influence. Several studies indeed showed that the risk of being involved in an accident increases with rising blood alcohol concentrations (BACs; Borkenstein et al., 1964; 1974; Compton et al., 2002; Blomberg et al., 2009). Borkenstein et al. (1964; 1974) demonstrated in a grand epidemiologic study (6000 people involved in motor vehicle collisions, over 7000 controls, see also chapter 4.2) that the increase in relative accident risk begins at BAC levels as low as 0.04 ‰ and increases notably at BACs over 0.08 ‰, continuing exponentially. Borkenstein's so-called Grand Rapids study is recognised as the golden standard for the effect of alcohol on driving performance. This knowledge has provided a solid basis for legislative decisions about driving under the influence of alcohol in many countries; the underlying thought being that alcohol at a BAC over 0.05 ‰ causes driving impairments that lead to accidents.

Additionally, laboratory studies have confirmed the thought that alcohol impairs cognitive operations that are needed to competently drive a vehicle as well. Many scientific studies have been conducted since long, measuring the objective and subjective effects of alcohol on cognitive functioning. Lately, they included specifically aspects of driving, in particular reaction time and keeping the vehicle in a straight course (Louwerens et al., 1987). The latter was measured as the Standard Deviation of the Lateral Position SDLP (O'Hanlon et al., 1982; O'Hanlon, 1984), a golden standard now for admission of medicinal drugs in traffic (see Ramaekers, 2017), and assessing the risk of recreational drugs (see Veldstra et al., 2011).

Impact for framework

In order to find comparable standards for evaluating the safety of in-vehicle automation, similar measures may be useful for determining "impaired driving" while interacting with in-vehicle automation, for example, during the transition of control phase (Brookhuis et al., 2003; see also section 3.3). An important factor for impaired driving will be inattention, for example resulting from introspective behaviour (daydreaming) by the driver, distraction, or performing NDRAs. The operational result is that the driver makes a delayed response, an inappropriate response, or no response at all (Sussman et al., 1985). The accident likelihood of delayed responses is still acceptably low up to within 2.5 – max 4 seconds (Godthelp & Käppler, 1988). However, one study demonstrated that, when required to resume control, drivers were able to return into the loop only after 5 to 7 seconds (Gold et al., 2016). Inappropriate responses may lead to dangerous manoeuvres for quite some time after resuming control by the driver, since it might take about 30-45 seconds to stabilise behaviour (Merat et al., 2014). Not (timely or accurately) responding may lead to serious accidents, implying that choices should be made about which 'golden' standard variables should be included in evaluations for in-vehicle

automation. In short, variables such as reaction time (RT) and swerving (SDLP), but perhaps also following distance (Time Headway), time to collision (TTC), and/or collision avoidance could be used, depending on the type of automation and (traffic) context. For each of these variables, a rigorous test application comparable with the alcohol standard may lead to the requirement for collecting data from many ADS-equipped vehicles.

An example of a very pragmatic standard is the two-second rule, which is a rule of thumb for a driver to maintain a safe following distance at any speed on any road. The rule is that a human driver should ideally stay at least two seconds behind any vehicle that is directly in front of the driver's vehicle. The reaction time of a human driver is composed of perception, decision, lifting-moving-pressing the foot, braking, of which the first three take close to a second. Therefore, an ADS would still be safe at the following distance of one second, however, drivers (or passengers) might feel uncomfortable in that case.

5. Measures and methods

5.1. Test methods

Apart from defining which concepts are relevant for testing, another step is to determine the required methods for performing the tests in practice. There are multiple methods available, each with its own strengths and limitations: mostly comprising a trade-off between experimental control (internal validity) and realism (ecological validity). In general, one can distinguish between paper-based testing and use(r)-based testing, of which several suggestions are briefly provided below. For a more elaborate overview, the reader is referred to Souman et al. (2021c).

Paper-based assessments could be performed both by trained or untrained evaluators. In theory, well-designed checklists (e.g., including ISO-standards) could lead to efficient and standardised evaluations because anybody should be able to perform these assessments. This method gives less realistic results, compared to for example user-based testing, because it does not incorporate any results of true vehicle/automation usage, and may also require (pre)defining very strict evaluation points and criteria about very specific systems or (sub)components for different contexts. This will lead to a very large number of evaluation points if all the systems in a vehicle have to be evaluated. Appointing trained experts to provide their opinions may solve part of this problem, as they are able to take broader impacts of smaller details into account. This could come at the expense of objectivity, however.

An alternative for paper-based testing is use(r)-based testing with participants. If a representative sample can be formed, multiple methods can be applied to evaluate emerging interactions between human drivers and in-vehicle automation. First, driving simulators may be used. Even though a driving simulator is a virtual assessment tool, it does include participants performing the (simulated) driving task and which increases realism compared to checklist testing. Moreover, with driving simulators, much more (experimental) control is possible: a driving simulator allows exposing every participant to the exact same vehicle, automation, environment, traffic conditions, etc., to an extent that is not possible with non-virtual assessment methods. The largest limitation is that simulators are not real vehicles, which limits the ecological validity. To improve ecological validity and realism further, participants can be tested in a real vehicle, but within a controlled environment: on a test track. The most realistic test method is driving the vehicle within real traffic (i.e., on-road testing), which is also the least controlled environment (Souman et al., 2021c).

Test measures

There are several variables suitable to measure the effects of driver-automation interactions and specific variables can be used for measuring one or multiple safety-related aspects of the framework (for example: situation awareness, mental models, transfer of control). Variables such as reaction time, lateral position, swerving (standard deviation lateral position), speed, time headway, time to collision, and/or obstacle/accident avoidance can be measured and used as input to evaluate the overall safety and performance of driver-automation interactions objectively (see also section 4.3). The largest advantage of these variables is relevance and realism, because measured driving parameters have a direct link to driving behaviour in real traffic.

Apart from measuring parameters that are directly related to the driving task, psychophysiological measures such as EEG, ERP, heart rate, heart rate variability, blood pressure, skin conductance, thermal imaging, and/or pupillometry, data about arousal, (cognitive) workload, and mental effort may be acquired while driving and interacting with automation (see e.g., Lohani et al., 2019). Measuring eye movements (e.g., fixations, % gaze time off-road) can give insights in perception, attention, distraction, and situation awareness, for example (see e.g., Khan & Lee, 2019). Even though these measurements provide relatively objective information, they often require the use and installation of advanced measuring equipment that needs to be operated by trained experts and, because the link with driving behaviour is less direct, interpretation of results is often necessary.

With self-reports, drivers can be asked to rate their own performance (e.g., how well were you able to drive) or required mental effort, or what their situation awareness is at a certain moment while driving. Participants can be asked with self-reports about their subjective opinions about the interaction processes and acceptance of systems and interactions. Drivers' mental models may also be investigated with self-reports (see sections 3.1 and 3.2). Several examples of self-report questionnaires are the Acceptance Scale (Van der Laan et al., 1997), Rating Scale Mental Effort (RSME, Zijlstra, 1993), and SAGAT (Endsley, 1985). Although self-reports can provide very detailed information about an individual's experience, it may be difficult to generalize the results to a larger population.

Monitoring the performance on non-driving related tasks (NDRA's) while driving with automation can provide information about the influence of distraction on driving safely. A secondary task potentially decreases situation awareness, and may therefore be used to determine whether specific interaction procedures allow a driver to regain sufficient situation awareness to perform a safe transfer of control.

Impact for framework

The measures listed above form a typical but non-exhaustive list for evaluating human (driving) behaviour. They are often used in very particular settings and interpreted in that context. To use them to evaluate driver behaviour when driving with ADS or ADAS requires that they are specified for that context. This is not impossible, but also not very easy and may require, as indicated earlier, many separate specifications for different systems, aspects of systems or contexts.

Because the different assessment methods each have their own strengths and limitations, it should first be determined in which part of the development process the evaluations should take place. One option could be to include multiple testing methods in the total evaluation to guide a vehicle's development process and its interactions. Evaluation could start with largely controlled methods for concept testing, and the more a vehicle and its automation are deployed in a real vehicle, the more realistic evaluation methods could be used. Whether it is indeed desirable or necessary to apply a full step-by-step evaluation of the design process, to evaluate a selection of steps, or only the final step (i.e., the completed vehicle) has yet to be determined. In addition, choices should be made about the responsible parties for (each part of the) evaluation procedure and what the role and responsibilities of the OEM and/or a dedicated testing agency are.

5.2. Context: exceptional or critical situations

Even though driving always carries some risks, not all traffic situations and contexts are equally dangerous. Indeed, the riskiest situations may be those that occur very scarcely and unexpectedly, which require a fast response from the driver. Such so-called 'critical' or 'near-crash' scenarios are very useful for assessing safety effects of in-vehicle automation. Specifically, scenarios in which automation reaches ODD limitations and/or runs into (technical) failures may be examined. In addition, driving while deliberately using automation outside its ODD could be investigated as well. Based on the information provided by the system, the interaction that follows should enable the human driver to intervene and prevent the event from resulting in a crash.

Impact for framework

There are several limitations for using (exceptional) critical scenarios in a testing framework. Firstly, because of the risk involved in critical situations, the possibilities for testing these scenarios in real (naturalistic) traffic are limited. Therefore, driving behaviour during critical situations may only be investigated on a test track or in a driving simulator, for example. Secondly, critical situations may influence driving behaviour after the incident. Obviously, after an unexpected event has occurred, that event is not 'unexpected' anymore and could cause the human driver to anticipate a similar event during the rest of the testing procedure. This driver could therefore be more attentive during the remaining parts of a test and may react differently than before. For this reason, critical events can mostly only be presented once to a single driver. Lastly, 'unexpected' situations are by definition unpredictable, not only for the manufacturer, but also for the evaluator. It is difficult if not impossible to overcome this limitation because nobody can expect the unexpected. In practice, this means that there could always be specific situations or circumstances that could not have been tested before allowing a vehicle on the road. It is difficult, if not impossible to overcome this limitation.

6. Other considerations

6.1. Technological developments

Technologies facilitating ADAS and ADS will continue to develop rapidly, new (sub)systems will be developed continuously. This means that, while evaluation procedures for the current technologies are established, new technologies will emerge that might require new evaluation procedures. It is important to acknowledge, but at the same critically consider such developments because they might improve safety even further, while on the other hand instigate unknown side-effects (Dutch Safety Board, 2019). Care should be taken that evaluation procedures keep up with such technologies, while not hampering innovation. Ideally, legislators together with manufacturers should find a workable balance between innovation and safety, and evaluations should be designed in such a way that both factors go hand-in-hand. For example, if new systems incorporate the same interaction processes that are evaluated positively with an earlier system (or vehicle), an entirely new evaluation might perhaps not be necessary.

6.2. Continuous monitoring

Some of the test methods described above lend themselves for continuous monitoring of the system during use. With continuous monitoring data from the vehicle, system and/or driver is collected while using the system, either during development or while driving. This gives a possibility to monitor whether systems are operating safely and to detect potential unsafe situations. In combination with other assessment methods (see also section 5.1), this can give valuable insights into the (ongoing) quality of the system. Manufacturers already gather data in their vehicles about how their ADAS are functioning. From a privacy standpoint, gathering information about users can be (or become) problematic and in the end, some rules about how this data is gathered, stored and used may be necessary.

7. Conclusions

The main question for developing evaluation procedures for interactions between human drivers and in-vehicle automation is whether it is possible to reduce the complex reality of driving with and without ADAS/ADS (or any related system) to a framework that can be practically evaluated and covers as much as possible. The present document provides relevant aspects that may be incorporated in such a framework. Even though this list is not exhaustive, it leads to the conclusion that there are still many questions to consider before a framework can be constructed that is decisive and feasible in practice. Part of these questions are very complex and require more research, in spite of the extensive research that has been performed in the field already. There are also issues, however, which could be resolved by reaching consensus and making decisions in the field. Even though many of those issues are highly complex as well, vigour and decisiveness from all relevant parties can contribute to (further) developing evaluation processes. Based on the explorations described in this document, at least the following questions should be considered:

- The criterion for 'acceptable safety' should be determined, in other words: when would society and science in concordance agree that a system is 'safe enough'? A threshold could be set at 95%, 99%, or 99.9% safety, for example, but this should be discussed, agreed upon, and decided by all involved parties. The decision for this threshold not only provides the framework with concrete starting points to form criteria, but also determines the statistical feasibility of performing test procedures in general. The higher the safety threshold, for example, the more participants should be involved in testing, which could also mean that numbers become so large that testing is practically not feasible.
- Decisions should also be made about whether all possible in-vehicle interactions should be evaluated, or whether it is more feasible to measure smaller (sub)segments of these interactions. Evaluating all possible interactions (e.g., also while driving with different or multiple combinations of support and/or automation systems) may lead to a very large number of potential evaluation points, which will be difficult to implement all in practice. The relevant aspects described in this document are suggestions for starting points, but measuring all these aspects could become too much for practical use as well. Therefore, it is advisable to determine which aspects are the most relevant ones and provide the most reliable insights in safety-effects, both positive and negative. This way, evaluation procedures can be made more succinct, relevant, and therefore more practical to perform.
- One way to further limit the potential number of in-vehicle interaction measurements is to define specific contexts that are most relevant or critical for the safety of a specific ADAS or ADS. Within such contexts, concrete scenarios (resembling real-world driving situations) could be created that enable evaluating whether one or more systems allows people to drive safely in that context. Critically determining the most safety-relevant contexts is crucial, because otherwise also the number of system-context combinations could (again) become enormous. In the end, a limited yet representative set of system-specific contexts should be identified, similar to the current driver licensing procedures.
- Criteria should be set that allow pass/fail ratings for 'safe driving'. The criteria can be used to determine a pass/fail for the vehicle as a whole with all systems included, for a given system, or even for a part of a system in a specific context. The advantage of context-specific criteria is that these are most relevant for real world behaviour (i.e., was a driver able to drive safely in the given circumstances), while the most detailed option (i.e., criteria

for parts of systems) provide the most insights in system-specific functioning and whether that (part of a) system should be improved or not. Either way, further choices should be made about relevant measurement variables (i.e., operationalisation) and acceptable criteria for each scenario (based on the decision about 'acceptable safety' earlier, see e.g., Souman et al., 2021c). For example, acceptable ToC times in general and/or critical situations should be determined in order to evaluate these.

- Long-term effects are also important, as these may differ from short-term effects. Behavioural adaptation, for instance, could play a major role, as has been demonstrated with other systems that may lead to complacency (i.e., unjustified overreliance). What exactly is meant by long-term effects, and thus what comprises a long-term effect evaluation, has to be determined as well. In principle, this could range from hours to months of use. Monitoring the effects on behaviour after introduction of systems is advisable in this regard.

Finally, the present report explored the possibilities and obstacles for developing an assessment framework for ADS and ADAS, focussing on the interaction between driver and system. It is difficult to give a clear-cut answer to the feasibility of such a framework. As this document shows, there is a large body of research about all the different topics related to the interaction and assessment. Very little of the research has however led to standardization or golden rules for a certain, standard approach to ADS/ADAS. The reason, probably, is that it is difficult to formulate such a standard or golden rule. In some cases, it will demand a pragmatic approach to overcome some of the difficulties to develop an assessment protocol (e.g., to reduce the number of factors that need to be considered). In other cases, choices have to be made on what is acceptable or acceptably safe. These choices are more often than not a political or social choice. Important for making these choices is cooperation and agreement between the different stakeholders (government, OEMs, etc).

References

Andrews, E.C., & Westerman, S.J. (2012). Age differences in simulated driving performance: Compensatory processes. *Accident Analysis and Prevention*, 45, 660-668.

Anstey, K.J., & Wood, J. (2011). Chronological Age and Age-Related Cognitive Deficits Are Associated with an Increase in Multiple Types of Driving Errors in Late Life. *Neuropsychology*, 25(5), 613-621.

Automated Vehicle Safety Consortium (2021). AVSC Best Practice for Evaluation of Behavioral Competencies for Automated Driving System Dedicated Vehicles (ADS-DVs). *SAE-ITC Report, AVSC00008202111.*

Beggiato, M., Hartwich, F., Schleinitz, K., Krems, J., Othersen, I., Petermann-Stock, I. (2015). What would drivers like to know during automated driving? Information needs at different levels of automation. *7th conference on driver assistance*, Munich, 25-26, 26.11.2015.

Biernacki, M.P., & Lewkowics, R. (2021). How do older drivers perceive visual information under increasing cognitive load? Significance of personality on-road safety. *Accident Analysis and Prevention*, 157, 106186.

Blomberg, R.D., Peck, R.C., Moskowitz, H., Burns, M., & Fiorentino, D. (2009). The Long Beach/Fort Lauderdale relative risk study. *Journal of Safety Research*, 40(4), 285-292.

Boelhouwer, A. (2021). Exploring, Developing, and Evaluating In-car HMI to Support Appropriate Use of Partially Automated Cars. *PhD Thesis,* University of Twente.

Borkenstein, R.F., Crowther, R.F., Shumate, R.P., Zeil, W.W., & Zylman, R. (1964). The role of the drinking driver in traffic accidents. Bloomington, IN: Department of Police Administration, Indiana University.

Borkenstein R.F., Crowther F.R., Shumate, R.P., Ziel, W.B., Zylman, R. (1974). The role of the drinking driver in traffic accidents (the Grand Rapids Study). *Blutalkohol*, 11, 1-131.

Borowsky, A., Shinar, D., & Oron-Gilad, T. (2010). Age, skill, and hazard perception in driving. *Accident Analysis and Prevention*, 42, 1240-1249.

Brookhuis, K.A., De Waard, D., & Fairclough, S.H. (2003). Criteria for driver impairment, *Ergonomics*, 46(5), 433 - 445.

Campbell, J.L., Brown, J.L., Graving, J.S., Richard, C.M., Lichty, M.G., Sanquist, T., Bacon, L.P., Woods, R., Li, H., Williams, D.N., & Morgan, J.F. (2016). Human Factors Design Guidance for Driver-Vehicle Interfaces (DOT HS 812 360). Washington, DC: National Highway Traffic Safety Administration.

Campbell, J.L., Graving, J.L.B.J.S., Richard, C.M., Lichty, M.G., Bacon, L.P., Morgan, J.F., Li, H., Williams, D.N., & Sanquist, T. (2018). Human Factors Design Guidance for Level 2 and Level 3 Automated Driving Concepts (DOT HS 812 555). Washington, DC: National Highway Traffic Safety Administration.

Champely, S. (2020). pwr: Basic Functions for Power Analysis. R package version 1.3-0. https://CRAN.R-project.org/package=pwr

Cundill, B., & Alexander, N.D.E. (2015). Sample size calculations for skewed distributions. *BMC Medical Research Methodology*, 15:38. doi:10.1186/s12874-015-0023-0

De Winter, J. C. F., Happee, R., Martens, M. H. & Stanton, N. A. (2014). Effects of adaptive cruise control and highly automated driving on workload and situation awareness: A review of the empirical evidence. *Transportation Research Part F: Traffic Psychology and Behaviour*, 27, 196-217.

Depestele, S., Ross, V., Verstraelen, S., Brijs, K., Brijs, T., Van Dun, K., & Meesen, R. (2020). The impact of cognitive functioning on driving performance of older persons in comparison to younger age groups: A systematic review. *Transportation Research Part F: Traffic Psychology and Behaviour*, 73, 433-452.

Dunn, N., Dingus, T. & Soccolich, S. (2019). Understanding the Impact of Technology: Do Advanced Driver Assistance and Semi-Automated Vehicle Systems Lead to Improper Driving Behavior? (Report). Washington, DC: AAA Foundation for Traffic Safety.

Dutch Safety Board (2019). Who is in control? Road safety and automation in road traffic (Report). The Hague: Dutch Safety Board.

Endsley, M.R. (1985). Technological Change and Individual Adjustment. In: Proceedings of the Human Factors Society Annual Meeting, Volume 29(6), 598-602.

Endsley, M.R. (1995). Toward a theory of situation awareness in dynamic systems. *Human Factors*, 37(1), 32-64.

Endsley, M.R. (2000). *Situation Models: An Avenue to the Modeling of Mental Models.* In Proceedings of the XIVth Triennial Congress of the International Ergonomics Association and 44th Annual Meeting of the Human Factors and Ergonomics Association, 'Ergonomics for the New Millennium' (pp. 61-64). Human Factors and Ergonomics Society.

Eriksson, A., & Stanton, N.A. (2017). Takeover Time in Highly Automated Vehicles: Noncritical Transitions to and From Manual Control. *Human Factors*, 59, 689-705.

Ernst, A.F., & Albers, C.J. (2017). Regression assumptions in clinical psychology research practice - a systematic review of common misconceptions. *PeerJ*, 5, e3323. doi:10.7717/peerj.3323

Falkenstein, M., Karthaus, M., & Brüne-Cohrs (2020). Age-Related Diseases and Driving Safety. *Geriatrics*, 5, 80.

Faul, F., Erdfelder, E., Lang, A.G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, 39, 175-191.

Godthelp, J, Käppler, W.D. (1988). Effects of vehicle handling characteristics on driving strategy. *Human Factors*, 30 (2), 219-229.

Gold, C., Happee, R., & Bengler (2018). Modeling take-over performance in level 3 conditionally automated vehicles. *Accident Analysis and Prevention*, 116, 3-13.

Gold, C., Körber, M., Lechner, D., & Bengler, K., (2016). Taking Over Control From Highly Automated Vehicles in Complex Traffic Situations: The Role of Traffic Density. *Human Factors*, 58(4), 642-652.

Harms, I.M., Bingen, L., & Steffens, J. (2020). Addressing the awareness gap: A combined survey and vehicle registration analysis to assess car owners' usage of ADAS in fleets. *Transportation Research Part A: Policy and Practice*, 134, 65-77.

International Organization for Standardization (2020). Road Vehicles: Human Performance and State in the Context of Automated Driving: Part 2 – Considerations in designing experiments to investigate transition processes (ISO/TR 21959-2).

Johnson-Laird, P.N. (1994). Mental models and probabilistic thinking. *Cognition*, 50 (1-3), 189-209.

Khan, M.Q., & Lee, S. (2019). Gaze and Eye Tracking: Techniques and Applications in ADAS. *Sensors*, 19(24), 5540.

Kircher, K., Larsson, A. & Hultgren, J. A. (2014). Tactical driving behavior with different levels of automation. *IEEE Transactions on Intelligent Transportation Systems*, 15(1), 158-167.

Kroon, E.C.M., Martens, M.H., Brookhuis, K.A., De Waard, D., Stuiver, A, Westerhuis, F., de Angelis, M., Hagenzieker, M.P., Alferdinck, J.W.A.M., Harms, I.M., & Hof, T. (2019). Human factor guidelines for the design of safe in-car traffic information services (3rd edition). Groningen: University of Groningen.

Lakens, D., Adolfi, F.G., Albers, C.J., et al. (2018). Justify your alpha. *Nature Human Behaviour*, 2, 168–171. doi:10.1038/s41562-018-0311-x

Larsson, A.F., Kircher, K., & Hultgren, J.A. (2014). Learning from experience: Familiarity with ACC and responding to a cut-in situation in automated driving. *Transportation Research Part F: Traffic Psychology and Behaviour,* 27, 229-237.

Lohani, M., Payne, B.R., & Strayer, D.L. (2019). A Review of Psychophysiological Measures to Assess Cognitive States in Real-World Driving. *Frontiers in Human Neuroscience*, 13:57.

Louwerens, J.W., Gloerich, A.B.M., De Vries, G., Brookhuis, K.A., & O'Hanlon, J.F. (1987). The relationship between drivers' blood alcohol concentration (BAC) and actual driving performance during high speed travel. In P.C. Noordzij & R. Roszbach (Eds.), Alcohol, Drugs and Traffic Safety-T86 (pp. 183-186). Amsterdam: Excerpta Medica.

Lu Z., Happee, R., Cabral, C.D.D., Kyriakidis, M. & De Winter, J.C.F. (2016). Human factors of transitions in automated driving: A general framework and literature survey. *Transportation Research Part F: Traffic Psychology and Behaviour*, 43, 183-198.

Melnicuk, V., Thompson, S., Jennings, P., & Birrell, S. (2021). Effect of cognitive load on drivers' State and task performance during automated driving: Introducing a novel method for determining stabilisation time following take-over of control. *Accident Analysis and Prevention*, 151, 105967.

Merat, N., Jamson, A.H., Lai, F.C.H., Daly, M., & Carsten, O. (2014). Transition to manual: Driver behaviour when resuming control from a highly automated vehicle. *Transportation Research Part F: Traffic Psychology and Behaviour*, 27, 274-282.

Naujoks, F., Höfling, S., Purucker, C., & Zeeb, K. (2018). From partial and high automation to manual driving: Relationship between non-driving related tasks, drowsiness and take-over performance. *Accident Analysis and Prevention*, 121, 28-42.

Norman, D.A. (1983). Some observations on mental models. In: Gentner,D., Stevens, A.L., Mental Models. Psychology Press, New York, pp. 7- 14.

O'Hanlon, J.F., Haak, T.W., Blaauw, G.J., Riemersma, J.B. (1982). Diazepam impairs lateral position control in highway driving. *Science*, 217(4554), 79-81.

O'Hanlon, J.F. (1984). Driving performance under the influence of drugs: rationale for, and application of, a new test. *British Journal of Clinical Pharmacology*, 18(1), 121-129.

Orlovska, J., Novakazi, F., Lars-Ola, B., Karlsson, M., Wickman, C., & Söderberg, R. (2020). Effects of the driving context on the usage of Automated Driver Assistance Systems (ADAS) Naturalistic Driving Study for ADAS evaluation. *Transportation Research Interdisciplinary Perspectives,* 4, 671-683.

Petermeijer, B., Tinga, A., & de Reus, A. (2021a). Verkenning Kwaliteit Gedeelde Rijtaak: Eindrapport. URL: https://puc.overheid.nl/rijkswaterstaat/doc/PUC_637424_31/

Petermeijer, S.M., Tinga, A.M., de Reus, A., Jansen, R.J., & van Waterschoot, B.M. (2021b). What Makes a Good Team? - Towards the Assessment of Driver-Vehicle Cooperation. In 13th International Conference on Automotive User Interfaces and Interactive Vehicular Applications (AutomotiveUI '21), September 09-14, 2021, Leeds, United Kingdom. ACM, New York, NY, USA.

Ramaekers, J.G. (2017). Drugs and driving research in medicinal drug development. *Trends in Pharmacological Sciences*, 38(4), 319-321.

Rogers, J.L., Howard, K.I., Vessey, J.T. (1993). Using significance tests to evaluate equivalence between two experimental groups. *Psychological Bulletin*, 113(3), 553-565. doi:10.1037/0033-2909.113.3.553

Schömig, N., Wiedemann, K., Hergeth, S., Forster, Y., Muttart, J., Eriksson, A., Mitropoulos-Rundus, D., Grove, K., Krems, J., Keinath, A., Neukum, A., & Naujoks, F. (2020). Checklist for expert evaluation of HMIs of automated vehicles-discussions on its value and adaptions of the method within an expert workshop. *Information (Switzerland)*, 11(4), 360.

Simons-Morton, B.G., Klauer, S.G., Ouimet, M.C., Guo, F., Albert, P.S., Lee, S.E., Ehsani, J.P., Pradhan, A.K., & Dingus, T.A. (2015). Naturalistic teenage driving study: Findings and lessons learned. *Journal of Safety Research*, 54, 41-48.

Snapinn, S.M. (2000). Noninferiority trials. *Current Controlled Trials in Cardiovascular Medicine*, 1(1), 19-21.

Souman, J., Van Weperen, M., Hogema, J., Hoedemaeker, M., Westerhuis, F., Stuiver, A., & De Waard, D. (2021a). Human Factors Guidelines Report 3: Use and Mental Models. Soesterberg: TNO 2020 R12165, Final Report.

Souman, J., Van Weperen, M., Hogema, J., Hoedemaeker, M., Westerhuis, F., Stuiver, A., & De Waard, D. (2021b). Human Factors Guidelines Report 4: Human Factors Guidelines for Advanced Driver Assistance Systems and Automated Driving Systems. Soesterberg: TNO 2020 R12164, Final Report.

Souman, J., Van Weperen, M., Hogema, J., Hoedemaeker, M., Westerhuis, F., Stuiver, A., & De Waard, D. (2021c). Human Factors Guidelines Report 5: Test Criteria. Soesterberg: TNO 2020 R12162, Final Report.

Sussman, E.D., Bishop, H., Madnick, B., & Walter, R. (1985). Driver inattention and highway safety. *Transportation Research Record*, 1047, 40-48.

Starkey, N.J., & Isler, R.B. (2016). The role of executive function, personality, and attitudes to risks in explaining self-reported driving behaviour in adolescent and adult male drivers. *Transportation Research Part F: Traffic Psychology and Behaviour*, 38, 127-136.

Van der Laan, J.D., Heino, A., & De Waard, D. (1997). A Simple Procedure for the Assessment of Acceptance of Advanced Transport Telematics. *Transportation Research Part C: Emerging Technologies*, 5(1), 1-10.

Veldstra, J.L., Brookhuis, K.A., De Waard, D., Molmans, B.H.W., Verstraete, A.G., Skop, G., & Jantos, R. (2011). Effects of alcohol (BAC 0.5‰) and ecstasy (MDMA 100 mg) on simulated driving performance and traffic safety. *Psychopharmacology*, 222, 377-390.

Yoon, S.H., Lee, S.C., & Ji, Y.G. (2021). Modeling takeover time based on non-driving-related task attributes in highly automated driving. *Applied Ergonomics*, 92, 103343.

Zijlstra, F. (1993). *Efficiency in work behavior. A design approach for modern tools.* Delft University of Technology. Delft: Delft University Press.