TECHNISCHE UNIVERSITÄT MÜNCHEN
TUM School of Engineering and Design

Effect of Fatigue on Take-Over Performance in Conditionally Automated Driving

Anna Franziska Feldhütter, M.Sc.

Vollständiger Abdruck der von der TUM School of Engineering and Design
der Technischen Universität München
zur Erlangung des akademischen Grades eines

Doktors der Ingenieurwissenschaften (Dr.-Ing.)

genehmigten Dissertation.

Vorsitz: Prof. Dr. Birgit Vogel-Heuser

Prüfer*innen der Dissertation: 1. Prof. Dr. Klaus Bengler
2. Prof. Dr. Angelika Bullinger-Hoffmann

Die Dissertation wurde am 05.01.2021 bei der Technischen Universität München eingereicht und durch die TUM School of Engineering and Design am 13.07.2021 angenommen.

Acknowledgment/Danksagung

Diese Dissertation entstand während meiner Zeit als wissenschaftliche Mitarbeiterin und Doktorandin am Lehrstuhl für Ergonomie der Technischen Universität München. Für diese lehrreiche und prägende Zeit möchte ich mich sehr herzlich bei Herrn Prof. Bengler bedanken. Besonders hervorheben möchte ich die vielfältigen Möglichkeiten, das mir entgegen gebrachte große Vertrauen, die fachlichen Diskussionen und die außerordentliche Freiheit bei Projekten, Lehre und Forschung. Meinen Dank möchte ich außerdem Frau Prof. Vogel-Heuser für die Übernahme des Vorsitzes des Verfahrens und Frau Prof. Bullinger-Hoffmann für die Zweitbegutachtung der Dissertation aussprechen.

Bei Prof. Josef Nassauer möchte ich mich herzlich bedanken, dass er sich als Mentor für diese Dissertation zur Verfügung gestellt und stets mit großem Interesse meine Arbeit verfolgt hat.

Großen Dank spreche ich den vielen Studierenden aus, die mich mit ihrer Motivation und ihrem Einsatz beim Gelingen vieler Projekte und dieser Dissertation unterstützt haben. Besonders möchte ich mich in diesem Zusammenhang bei Maria Lechner, Alexandra Ruhl, Dominik Kroll, Kaiyu Huang, Tobias Hecht, Alexander Feierle, Luis Kalb, Annika Boos und Michael Rettenmaier bedanken. Besonders freut es mich, dass einige der Genannten ebenfalls ihren Weg an den Lehrstuhl gefunden haben und ich sie seitdem meine geschätzten Kollegen nennen durfte. Ganz besonders möchte ich mich bei meinem gesamten Kollegium am Lehrstuhl bedanken, welches die Zeit am Lehrstuhl so einzigartig gemacht hat. Ohne die Hilfsbereitschaft, die Kollegialität, den Zusammenhalt, den fachlichen Austausch und die immerzu gute Atmosphäre wäre das Gelingen dieser Arbeit in dieser Form nicht möglich gewesen. Aus verschiedenen Gründen geht mein besonderer Dank in diesem Zusammenhang an Christian Gold, Moritz Körber, Ingrid Bubb, Philipp Maruhn, Lorenz Prasch, Stephanie Cramer, Verena Knott, Patrick Gontar und Martin Götze.

Dass die Zeit am Lehrstuhl so einzigartig war, habe ich im Besonderen meinen Bürokollegen, Jonas Radlmayr und Yucheng Yang, zu verdanken. Die hitzigen Diskussionen, auch über fachliche Themen hinaus, die gute Stimmung, die Konferenzreisen, die Hilfsbereitschaft für Probleme jede Art, die Freundschaft und zuletzt das kritische Korrekturlesen dieser Arbeit sind nicht selbstverständlich und ich wusste/weiß das sehr zu schätzen.

Ich möchte mich außerdem bei meiner Familie bedanken, die mich immer motiviert hat nicht locker zu lassen und unerschütterlich an einen positiven Abschluss dieser Arbeit geglaubt hat.

Mein größter Dank gebührt meinem Mann, Till, der mich von Anfang an und über all die Jahre bei meinem Vorhaben bedingungslos unterstützt hat und mit viel Verständnis, Geduld und Ansporn einen außerordentlich großen Beitrag zum Gelingen dieser Dissertation beigetragen hat. Ohne dich wäre ich nun nicht an diesem Punkt angekommen!

Anna Feldhütter

Abstract

The release of conditionally automated driving (CAD, SAE level 3) in series vehicles is announced for the first half of the 2020s. At this automation level, drivers are no longer responsible for the driving task and are allowed to engage in non-driving-related activities (NDRAs) as long as they stay receptive for a request to intervene (RtI). CAD does not only offer many benefits, but it poses challenges, as well. Taking drivers out of the driver-vehicle-control loop and releasing them from an active role is assumed to promote the development of passive task-related (TR) fatigue. Fatigue is known from manual driving to have various negative effects on driving performance. Therefore, this thesis addresses different research questions regarding passive TR fatigue in CAD, which were derived from an extensive literature review. Research questions include the temporal development of passive TR fatigue, NDRAs as possible countermeasures, and the effect of passive TR fatigue on take-over performance.

To address the research questions, four driving simulator studies with a total of 191 participants were conducted. Participants were exposed to prolonged monotonous conditions with CAD to manipulate the fatigue state. Each participant experienced at least one take-over situation either when she/he was in a predetermined fatigue state or after a fixed driving duration of CAD. The take-over scenario was the same for all studies (broken-down vehicle in the ego-lane), only with slight variations of time budget and traffic condition. Take-over performance was assessed by temporal (take-over time) and quality metrics (e.g., accelerations, time-to-collision, reaction types). Over all studies, fatigue was assessed by a standardized expert rating based on certain behavioral fatigue indicators, retrospectively or in real-time. An additional subjective fatigue rating was applied in most experiments to further evaluate the success of the fatigue manipulation. With the applied method, a good comparability between studies could be achieved. Furthermore, the PERCLOS metric, promising for fatigue assessment and frequently used in past research on fatigue, was evaluated over all experiments.

Results showed that fatigue development is highly individual regarding whether and when a high level of fatigue is reached. Very short phases of CAD were already sufficient to promote a high level of fatigue (close to falling asleep). Motivating NDRAs could counter the onset of high fatigue levels; however, the success strongly depends on the individual motivation for an activity. Fatigue caused startle reactions and a reduced take-over quality, namely undeliberate maneuvers, such as accelerating, or strong decelerations indicating low situation awareness. Slower take-over times for fatigued drivers were not found.

The study results do not give a clear picture of the safety relevance of fatigue during CAD, which underlines findings from current related research. Nevertheless, the results of this thesis lead to the recommendation to avoid high fatigue levels, since a safe take-over cannot be ensured. For this purpose, a fatigue monitoring system should be implemented in vehicles with CAD. PERCLOS was found to be distinctive for extreme fatigue states (not fatigued at all and close to falling asleep). To achieve more robustness and higher reliability for fatigue monitoring, a high data availability is essential, and fusing multiple metrics for fatigue assessment is beneficial.

Kurzfassung

Die Serieneinführung von hochautomatisierten Fahren (HAF, SAE Level 3) wurde für die erste Hälfte der 2020er angekündigt. In diesem Automationslevel sind Fahrer nicht länger für die Fahrzeugsteuerung verantwortlich und dürfen sich fahrfremden Tätigkeiten (FFT) widmen, solange sie übernahmebereit bleiben. Neben vielen Vorteilen birgt HAF auch einige Herausforderungen. Es ist anzunehmen, dass das Herausnehmen der Fahrer aus dem Fahrer-Fahrzeug-Regelkreis und der Wegfall einer aktiven Rolle die Entwicklung von passiver aufgabenbezogener (AB) Müdigkeit fördert. Vom manuellen Fahren ist bekannt, dass Müdigkeit negative Effekte auf die Fahrleistung hat. Daher werden in dieser Arbeit verschiedene Fragestellungen hinsichtlich passiver AB Müdigkeit beim HAF adressiert, welche mithilfe einer ausführlichen Literaturrecherche hergeleitet wurden. Forschungsfragen umfassen die zeitliche Entwicklung von Müdigkeit, FFT als mögliche Gegenmaßnahme und die Auswirkung von Müdigkeit auf die Übernahmeleistung.

Um die Forschungsfragen zu beantworten, wurden vier Fahrsimulatorversuche mit 191 Probanden durchgeführt. Die Probanden wurden längeren monotonen HAF Fahrten ausgesetzt, um den Müdigkeitszustand zu manipulieren. Jeder Proband erlebte mindestens eine Übernahmesituation entweder bei Erreichen eines vordefinierten Müdigkeitszustands oder nach einer festen Automationsdauer. In allen Studien erlebten die Probanden das gleiche Übernahmeszenario (Pannenfahrzeug im Ego-Fahrstreifen) mit geringer Variation von Zeitbudget und Verkehrsbedingung. Die Übernahmeleistung wurde mithilfe von zeitlichen (Übernahmezeit) und qualitativen Metriken (z. B. Beschleunigungen, Time-to-collision, Reaktionsart) gemessen. Müdigkeit wurde durch eine standardisierte Expertenbewertung basierend auf der Beobachtung bestimmter Müdigkeitsindikatoren erfasst und entweder nachträglich oder in Echtzeit durchgeführt. Zusätzlich wurde in den meisten Studien Müdigkeit subjektiv erfasst, um den Erfolg der Müdigkeitsmanipulation zu überprüfen. Mit der verwendeten Methodik wurde eine gute Vergleichbarkeit zwischen den Studien erreicht. Darüber hinaus wurde PERCLOS evaluiert, eine zur Müdigkeitserfassung vielversprechende und viel verwendete Metrik.

Die Ergebnisse zeigten, dass die Müdigkeitsentwicklung, in Bezug auf wann und ob hohe Müdigkeitslevels auftreten, äußerst individuell ist. Bereits kurze Phasen von HAF reichten aus, um hohe Müdigkeitslevels hervorzurufen. Motivierende FFT konnten dem Beginn von hohen Müdigkeitslevels entgegenwirken, aber der Erfolg hängt stark von der individuellen Motivation für eine Tätigkeit ab. Während der Übernahme verursachte Müdigkeit Schreckreaktionen und eine reduzierte Übernahmequalität, welche sich in unüberlegten Manövern, wie Beschleunigen oder starkem Bremsen, äußerte, was auf ein geringes Situationsbewusstsein schließen lässt. Langsamere Übernahmezeiten wurden bei müden Fahrern nicht festgestellt.

Die Ergebnisse geben kein eindeutiges Bild bezüglich der Sicherheitsrelevanz von Müdigkeit bei HAF, was mit anderen aktuellen Forschungsarbeiten übereinstimmt. Dennoch wird empfohlen hohe Müdigkeitslevels zu vermeiden, da eine sichere Übernahme nicht gewährleistet ist. Daher wird der Einsatz einer Müdigkeitsüberwachung in HAF-Fahrzeugen empfohlen. PERCLOS ist eine geeignete Metrik, um extreme Müdigkeitszustände zu unterscheiden. Um größere Robustheit und Zuverlässigkeit der Müdigkeitsüberwachung zu erzielen, ist eine hohe Datenverfügbarkeit essentiell und die Fusionierung mehrerer Metriken sinnvoll.

Acronyms

AccLat	Maximum Lateral Acceleration
AccLong	Maximum Longitudinal Acceleration
ADS	Automated Driving System
ANOVA	Analysis of Variance
CA	Conditional Automation
CAD	Conditionally Automated Driving
CI	Confidence Interval
DDT	Dynamic Driving Task
DSSQ	Dundee Stress State Questionnaire
ECG	Electrocardiogram
EEG	Electroencephalography, Electroencephalogram
EOG	Electrooculography
Euro NCAP	European New Car Assessment Programme
FC	Fatigued Condition
FinRe	Final Response
FL1, FL2, FL3, FL4	Fatigue Level 1, Fatigue Level 2, Fatigue Level 3, Fatigue Level 4
HMI	Human-Machine Interaction, Human-Machine Interface
HRV	Heart Rate Variability
InRe	Initial Response
IQR	Interquartile Range
KSS	Karolinska Sleepiness Scale
MC	Mirror Check
MLR	Multinomial Logistic Regression
NC	NDRA Condition
NDRA	Non-Driving-Related Activity
NFC	Non-Fatigued Condition
NLC	Natural Load Condition
NMS	Number of Microsleep Events
ODD	Operational Design Domain
OEDR	Object and Event Detection and Response
OEM	Original Equipment Manufacturer
PERCLOS	Percentage of Eyelid Closure
PHM	Head Movement Behavior
Q1, Q3	First quartile, Third quartile
RtI	Request to Intervene
RtM	Request to Monitor
SAE	Society of Automotive Engineers
SDLP	Standard Deviation of Lateral Position
SR	Sleep-Related
SSS	Stanford Sleepiness Scale
STS	Standard Deviation of Steering Wheel Angle
SuRT	Surrogate Reference Task
TOC	Take-Over Controllability
TOT	Take-Over Time
TR	Task-Related
TTC	Time-To-Collision
UC	Underload Condition

Contents

1 Introduction

1.1 Motivation

Since the early 1970s, the number of fatal road accidents in Western countries has dropped considerably. Measures have been taken in road building, legislation, the rescue chain and emergency medicine (Winkle, 2016, p. 337). However, technical achievements with regard to passive vehicle safety, for instance the crash-optimized passenger cabin, airbag systems and seatbelts, which reduce accident severity, as well as achievements concerning active vehicle safety, such as the anti-lock braking system or electronic stability control, have also contributed to this success (Kompass, Gruber, & Domsch, 2010). Despite this great success, the so called "vision zero" of the European Commission—no more road fatalities until 2050—is still a long way away, since in 2018 there were still 3,275 road traffic deaths in Germany (Statistisches Bundesamt, 2019), around 25,100 in Europe (European Commission, 2019), and even 1.35 million worldwide (World Health Organization, 2018). The most promising leverage point to attain "vision zero" is the further development of advanced driver assistant systems and the increasing automatization of more and more parts of the driving task, since 93.5 % of all road accidents are caused by human errors (Winkle, 2016, p. 353). However, research on automated vehicles is not only driven by the enhancement of road safety, but also for economic and user comfort reasons. Supporters expect that the traffic flow will be optimized (Maurer, Gerdes, Lenz, & Winner, 2016, p. 4), gas mileage will be reduced, the user's workload will decrease (de Winter, Happee, Martens, & Stanton, 2014) and user will attain greater freedom of time usage when driving is automated.

The ongoing technical advance in sensor technology, data fusion and computing power in the last few decades has enabled great progress for automated driving. Partly automated systems that perform completely lateral and longitudinal vehicle guidance but still need permanent supervision by the driver were already established in series vehicles in the 2010s. The next step on the way to fully automated driving will be conditional automation (CA), which is announced for the 2020s by the main automotive manufacturers. At this stage, the driver is no longer responsible for the driving task and is absolved from the supervision duty (SAE International, 2018). However, the driver remains the fallback level of the system and must, therefore, stay receptive for a system-initiated request to intervene (RtI) and for taking over vehicle guidance in a proper manner and within a specific time frame. For this reason, the driver still plays an important role for road safety because she/he is still needed to defuse and resolve specific situations once the technical system has reached its limits. The take-over of the vehicle guidance out of a passive role poses new challenges for the design of human-machine interaction (HMI). Past research has already identified a variety of influencing factors on how drivers perform in such take-over situations, for instance, the effect of different scenarios and situational factors on take-over performance. However, there are still many unanswered questions. A large research field in the last years has comprised the influence of the driver's state on take-over performance. Driver state is a complex concept with many aspects that are important for safe driving. In addition to mental workload or attention, for instance, one further important aspect of the driver's state is fatigue.

Many past studies have identified driver fatigue as one of the major contributing factors to accidents in road traffic in manual driving (Brown, 1994; Connor et al., 2002; Knipling & Wang, 1994). Depending on the study, the proportion of accidents mainly caused by fatigue amounts to up to 25–31 % (Jan, Karnahl, Seifert, Hilgenstock, & Zobel, 2005; Klauer, Dingus, Neale, Sudweek, & Ramsey, 2006; Zhao & Rong, 2013, p. 19). These high rates can probably be explained by various effects of fatigue on driving performance. In conditionally automated driving (CAD), fatigue may even be promoted due to the loss of an active task. Emerging monotony and boredom cause constant underload (May & Baldwin, 2009) which, in turn, may lead to the development of fatigue. However, the question of how fatigue affects take-over performance in CAD has not been answered. It is still unclear whether the driver is able to take over vehicle control safely and properly once she/he has reached a high level of fatigue.

Therefore, this dissertation thesis contributes to evaluating and quantifying the effect of fatigue on take-over performance in CAD to define requirements for the design of a suitable user interface and to enable a safe user interaction with an automated driving system.

1.2 Scope and Structure of this Thesis

For decades, fatigue has been a strongly researched topic, which is relevant for many different professions and disciplines. There are many approaches for classifying different types of fatigue (e.g., Matthews, Desmond, & Hitchcock, 2012, p. 141), for instance muscular or physical fatigue, which is particularly relevant for manual work, chronical or pathological fatigue, which often results from or is part of specific diseases, or mental fatigue. This thesis, however, focusses on one specific fatigue type, namely driver fatigue, and here a certain subgroup, passive task-related (TR) fatigue. In **chapter 4**, the aspects of fatigue which are relevant for this thesis are described. From the human factors perspective in CAD, it is particularly important that the driver can take over the driving task when she/he is required to do so. There are mental processes that need to function to access automatisms and to make the right decisions when transitioning from CAD back to manual driving. To understand why fatigue plays an important role for these processes taking place during a take-over, the relevant processes and models that apply to driving and the take-over are described in **chapter 2**. The classification of CAD, the take-over process and the known influencing factors on the take-over performance of drivers are outlined in **chapter 3**. Related research on driver fatigue in CAD is presented and discussed in **chapter 4.9**. Based on the findings from literature review, the four main research questions are stated in **chapter 5**. Since there are many methodological aspects in this thesis that were not varied between experiments, **chapter 6** describes the general method. In **chapter 7**, the four experiments conducted and their methodological features are described, results are presented and discussed. In **chapter 8**, all results are discussed in an overarching context and implications for system design are derived. Furthermore, the limitations are discussed, and recommendations for future research are given. **Chapter 9** summarizes the findings of this thesis.

Some passages of this thesis (in particular in chapter 6 and chapter 7) have been pre-published in Feldhütter, Feierle, Kalb, and Bengler (2018), Feldhütter, Kroll, and Bengler (2018), Feldhütter, Hecht, Kalb, and Bengler (2019) and Feldhütter, Ruhl, Feierle, and Bengler (2019). Some parts of the written text have been literally adopted.

Beside fatigue, there is a range of further automation effects known from other disciplines, which need to be addressed in the context of automated driving and which have not been in the focus of research that much to date. Examples are overtrust in automation (Parasuraman & Riley, 1997), mode confusion/mode errors (Bainbridge, 1983), and impaired manual performance due to automatization, to name only a few. Recent research on vehicle automation showed similar issues regarding mode awareness and overtrust (Feldhütter, Härtwig, Kurpiers, Mejia Hernandez, & Bengler, 2019; Feldhütter, Segler, & Bengler, 2018) accompanying automated driving. One cause for issues described is the strict differentiation in automation levels with significantly differing driver responsibilities, as is done, for example, by the SAE International (2018). This is necessary for the technical development of such systems, and probably also for legal reasons; however, it poses great challenges for the generation of a correct mental model of the users. For this reason, it is important to address all relevant automation effects and find a comprehensive solution to ensure a safe operation of automated vehicles.

2 The Human as Driver

To understand the potential effects of fatigue on driving and ultimately on take-over performance, it is first of all important to comprehend the basic processes when driving.

According to Geiser (1985), the driving task is categorized in three hierarchical levels (see Figure 2-1): the primary, secondary and tertiary task (Bubb, 2015a, pp. 20–24). In principle, the primary driving task comprises all actions that are required to transport persons or goods in a safe manner to a certain destination within a certain time frame. This includes the navigation on a specific route (navigation level), the maintenance of a certain velocity and keeping the course depending on the environmental conditions (road course, weather, etc.) (maneuver level), as well as the continuous correction of the longitudinal and lateral dynamics by operating the required controls, such as steering wheel, pedals, and gear shift (stabilization level). The secondary driving tasks arise depending on the primary task to either indicate the driver's intentions to other road users (such as indicating or honking) or to react to specific environmental circumstances (activating the wipers or turning on the lights). Tertiary driving tasks have nothing to do with the actual driving task, but include actions within the vehicle to increase comfort (e.g., regulation of the air conditioner, adjustment of the seat position, etc.) or entertainment (e.g., radio, phone, etc.). Depending on the level of automation, the system takes over more and more of the primary driving tasks, but also parts of the secondary driving tasks have to be performed (e.g., indicating for lane change).

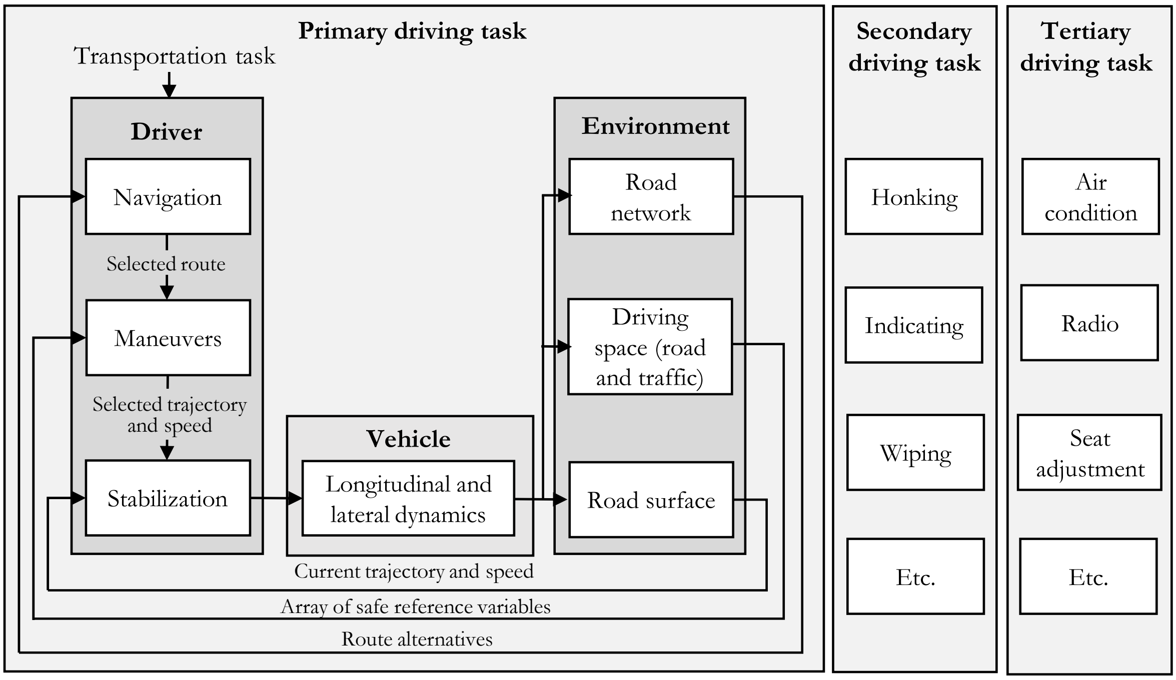


Figure 2-1. Tasks when driving based on Donges (2015, p. 19), Bubb (2015a, p. 21) and Gold (2016).

To be able to fulfill all arising tasks when driving in a safe manner, the driver has to constantly sense and perceive large amounts of information. Of the five senses for absorbing information from the environment, the most important one for driving is optical perception, but the acoustic, kinesthetic and haptic ones are also required (Bubb, 2015b, p. 68). From all the information that is sensed from the environment (e.g., taillights, traffic lights, curve radius, radio music, conversations of other passengers, etc.), the relevant ones for the respective driving situation have to be perceived. This perceived information is processed with the help of the working and long-term memory, which is the basis for making decisions and executing a respective response on the operator controls. In principle, these basic stages of information processing apply to all human interactions with technical systems and are generally modelled by Wickens, Hollands, Banbury, and Parasuraman (2013, pp. 3–6) (see Figure 2-2). All stages require a certain amount of attentional resources, whereas the amount of resources required in the individual stage depends on the respective task (Wickens et al., 2013, p. 5). However, the supply of resources is limited, which is the source of performance limitations and impairments or even errors, once the resource demands of a single task or for multitasking exceed available resources.

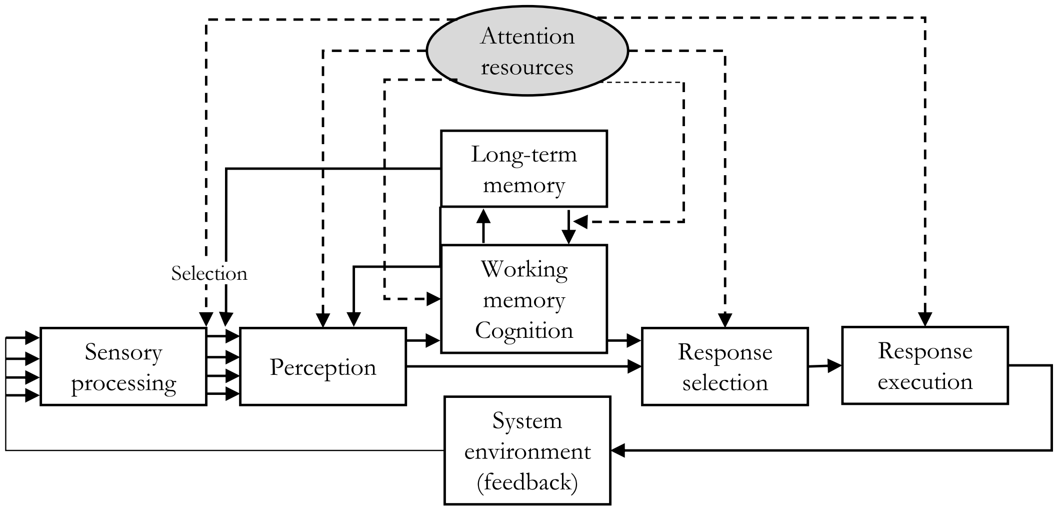


Figure 2-2. Information processing model according to Wickens et al. (2013, p. 4).

In a more general context of human work processes, Rasmussen (1983) developed a qualitative model (see Figure 2-3) to display three types of goal-driven sensomotoric activities with varying cognitive demands, which can also be applied to the driving context (Donges, 2015, p. 18). Complex and highly demanding or completely unknown situations, which require untrained actions, lead to the level of knowledge-based behavior. This behavior pattern activates resource-demanding mental processes that enable the human to compare different alternatives based on available knowledge and derive the action which is considered to fit the intended goal best. The next level is rule-based behavior, which is activated when the situational circumstances have already been experienced in the past and certain behavior patterns (rules) for handling such situations have been stored in the memory. The lowest level of skill-based behavior requires fewer mental resources and is characterized by reflexive stimulus-response mechanisms, which are long-term trained and unconsciously recalled in the respective situation. Skill-based behavior typically occurs for routine, recurring course of actions. (Donges, 2015, 18)

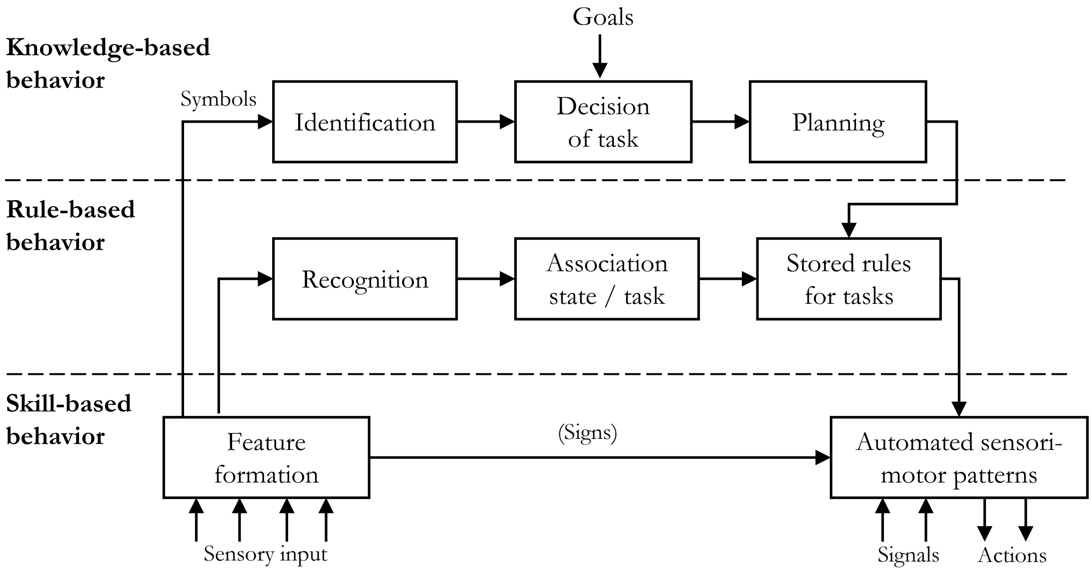


Figure 2-3. Three levels of behavior patterns defined by Rasmussen (1983).

The models described show that key elements for safe driving are visual and cognitive attention and sufficient resources for the single and, partially, also multiple tasks when driving. During manual driving, attentional lapses (for instance, what is referred to as the *look but failed to see* phenomenon) or distraction, for instance, by a non-driving-related activity (NDRA), are frequent causes for accidents (e.g., Phillips, 2014, p. 67). However, also for CAD the models described help to understand differences in performance in the transition from CAD to manual driving and how fatigue affects this process.

3 Automated Driving

3.1 History and Technology of Automated Driving

This chapter only includes a brief summary of the historical development of automated vehicles. A more detailed description is provided, for instance, by Kröger (2016) or Gold (2016).

The vision of self-driving cars began about a century ago with the development of two vehicles in 1921 and 1925 (“American Wonder”), which were remotely controlled via radio from a following vehicle (Kröger, 2016, p. 43). In the mid-1930s, the idea of electromagnetic wires integrated into the road surface to guide a vehicle emerged for the first time. General Motors picked up on this idea, and after a development process of only about five years, they conducted a one-mile test route with a rebuilt Chevrolet in 1958 (Kröger, 2016, p. 53). However, the dependency on the infrastructure was quickly considered expensive and inflexible. Therefore, from the 1970s onwards, research attempted to implement sensors and computers in the vehicles to become independent from infrastructure, which was also empowered by the rise of microelectronics (Kröger, 2016, p. 58). With the development of the anti-lock braking system in 1978, the first driver assistance systems actively intervening in vehicle guidance entered the market (Kröger, 2016, p. 59). Since then, numerous national and international research projects in academia and industry expedite the incremental development of automated driving: the pioneer project PROMETHEUS from 1987 to 1994 (Kröger, 2016, p. 59); DARPA (Defense Advanced Research Projects Agency) Grand Challenge I in 2004 (Gold, 2016), DARPA Grand Challenge II in 2005 (Buehler, Iagnemma, & Singh, 2007) and DARPA Urban Challenge in 2007 (Buehler, Iagnemma, & Singh, 2009); HAVE-it (2008–2011) (Hoeger et al., 2008); AdaptIVe (2014–2017) (Langenberg, Bartels, & Etemand, 2014); KoHAF (2015–2018) (ZENTEC Zentrum für Technologie, Existenzgründung und Cooperation GmbH); PEGASUS (2016–2019) (German Aerospace Center [DLR]); L3Pilot (2017–2021, expected) (L3Pilot consortium); to name only a few. Based on these project efforts and the accompanying ongoing achievements in sensor technology, data processing and computational power, automated driving, which was once only a utopian vision, has become a realistic scenario.

Today’s vehicles that drive in an automated way are equipped with a set of multiple sensors relying on different measurement principles (see Table 1 and Figure 3-1) to reliably capture all relevant information from the environment. Relevant information is, for instance, infrastructure, such as lane markings or traffic signs, other road users or obstacles on the road (Aptiv et al., 2019). Together with information from high definition maps, global navigation satellite systems and all environmental data are fused into an accurate world model (Aptiv et al., 2019). The simultaneous usage of at least two sensors with different measurement principles or information sources at the same time for object detection and localization enables higher reliability and data availability, and therefore increases safety (Aptiv et al., 2019). Based on the current world model, the future behavior of relevant objects is predicted to derive an adequate driving strategy, which includes a collision-free and lawful driving plan (Aptiv et al., 2019; Kämpchen, Aeberhard, Ardelt, & Rauch, 2012).

Table 1. List of on-board sensors for environment perception (adopted from Aptiv et al., 2019; Gold, 2016).

Sensor	Description
Camera	Sensor similar to human perception since capturing visible cues; highest extractable information content; most important sensor for object classification; relatively limited range; high sensitivity to weather conditions.
LIDAR	Optical (laser radiation) sensor with high-precision measurement of structured and unstructured objects; medium sensitivity to environmental conditions (e.g., reflectivity of material).
RADAR	High-precision detection and measurement of moving objects with appropriate reflectivity of radio waves, high robustness against weather conditions.
Ultrasonic	Well-established near-field sensor; detecting objects in closest distances; relying on the reflectivity of sound waves; for instance, deployed in parking assistant systems.
Microphone	Acoustic signals are used in public traffic to prevent crashes or to regulate traffic; Therefore, sensors capable to capture acoustic signals are required for automated driving.

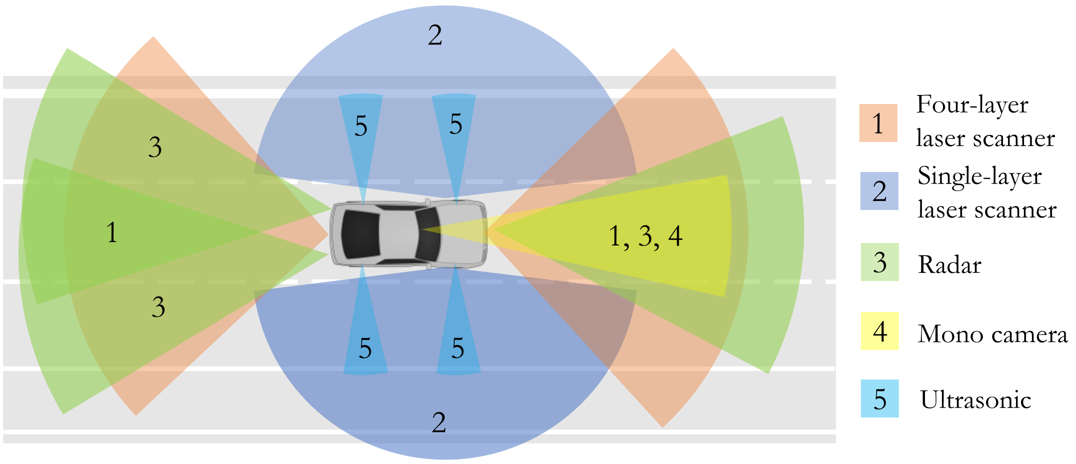


Figure 3-1. Possible sensor cluster for an automated vehicle illustrated by Aeberhard (2017, p. 134).

3.2 Levels of Automation and System Capability

The evolution to autonomous or fully automated driving includes various maturity levels, which are determined by specific system capabilities. There are different taxonomies that assign these system capabilities to different levels of automation and the respective remaining driver responsibility. The latter in particular needs to be clearly elaborated on and communicated to drivers to avoid general mode awareness problems in the driver (Sarter & Woods, 1995). The most noted taxonomy for driving automation is the one of the Society of Automotive Engineers (SAE), which is an international association of engineers and related professions concentrating on the development of standards. The SAE taxonomy (SAE International, 2018) includes six levels distinguishing the system capability in terms of the performance of the dynamic driving task (DDT), the DDT fallback and the operational design domain (ODD). The DDT contains all tactical and operational functions of the primary driving task (see chapter 2), except the strategic functions such as trip scheduling or destination selection. Tactical and operational subtasks include the two lower levels of the primary driving task, including the longitudinal and lateral vehicle motion control (steering and accelerating/decelerating) as well as maneuver planning. Besides, a further subtask of the DDT is object and event detection and response (OEDR), which comprises monitoring the driving environment (detection, recognition, classification of objects and events) and the preparation and execution of the corresponding actions. The ODD determines the conditions under which the automated driving system can operate and reflect the technological capability of the system. Conditions may contain environmental, geographical, time-of-day, traffic or roadway characteristics. The DDT fallback is relevant in case of a system failure or when the ODD limitations are reached. (SAE International, 2018)

An overview of the six SAE levels of driving automation and their properties is depicted in Table 2. Level 0 is manual driving, only supported by active safety systems (e.g., anti-lock braking system or electronic stability control). Level 1 comprises driver assistant systems, that either perform longitudinal or lateral vehicle motion control, for example, (adaptive) cruise control or lane keeping assistant. At level 2 (partial driving automation), the system performs the entire vehicle motion control; however, the full driving responsibility including constantly monitoring the system and the driving environment remains with the driver. If she/he recognizes an inappropriate action taken by the system, the driver must intervene without request. Consequently, in level 0–2, the user is considered the driver at all times. This restriction is removed from level 3 (Conditional Driving Automation or CAD/CA[1]) on. When CA is activated, the driver is released from her/his supervision responsibility and is allowed to avert her/his attention from the driving task. Instead, the driver can engage in NDRAs, for example, reading a book or using a smartphone. Nevertheless, the driver must stay receptive for a system-initiated RtI and must respond to it by promptly—within a few seconds—taking over the DDT. Receptivity is characterized by "a person's ability to reliably and appropriately focus her/his attention in response to a stimulus" (SAE International, 2018, p. 14). A sleeping driver is not considered to be receptive for a RtI. In some corner cases, the driver must intervene without receiving a request from the system, namely if the intended functionality is clearly no longer given, and this is not detected by the system (Deutscher Bundestag, 2017; SAE International, 2018). An example of this corner case of unintended functionality would be when the tie rod breaks and consequently driving behavior changes dramatically (SAE International, 2018). The next levels following CAD are level 4 (High Driving Automation) and level 5 (Full Driving

[1]According to SAE International (2018), level 3 automation is termed conditional driving automation. In this thesis, the term conditionally automated driving (CAD) or conditional automation (CA) is used which is synonymous to conditional driving automation.

Automation). The further development compared to level 3 is that the driver is no longer required as DDT fallback (level 4) and, for level 5, that the ODD is no longer limited.

Table 2. Levels of automation according to SAE International (2018). Focus of this thesis is on level 3 (Conditional Driving Automation).

SAE level	**SAE name**	**SAE narrative definition**	Dynamic driving task **(DDT)**		DDT fallback	Opera-tional design domain **(ODD)**
			Sustained lateral and longitudinal vehicle motion control	Object and event detection and response **(OEDR)**		
Driver performs part or all of the dynamic driving tasks						
0	**No driving automation**	The performance by the driver of the entire DDT, even when enhanced by active safety systems.	Driver	Driver	Driver	n/a
1	**Driver assistance**	The sustained and ODD-specific execution by a driving automation system of either the lateral or the longitudinal vehicle motion control subtask of the DDT (but not both simultaneously) with the expectation that the driver performs the remainder of the DDT.	Driver and system	Driver	Driver	Limited
2	**Partial driving automation**	The sustained and ODD-specific execution by a driving automation system of both the lateral and longitudinal vehicle motion control subtasks of the DDT with the expectation that the driver completes the OEDR subtask and supervises the driving automation system.	**System**	Driver	Driver	Limited
Automated driving system (ADS) (“system”) performs the entire dynamic driving task						
3	**Conditional driving automation**	The sustained and ODD-specific performance by an ADS of the entire DDT with the expectation that the DDT fallback-ready user is receptive to ADS-issued *requests to intervene*, as well as to DDT performance-relevant system failures in other vehicle systems, and will respond appropriately.	System	**System**	Fallback-ready user (becomes the driver during fallback)	Limited
4	**High driving automation**	The sustained and ODD-specific performance by an ADS of the entire DDT and DDT fallback without any expectation that a user will respond to a *request to intervene.*	System	System	**System**	Limited
5	**Full driving automation**	The sustained and unconditional (i.e., not ODD-specific) performance by an ADS of the entire DDT and DDT fallback without any expectation that a user will respond to a request to intervene.	System	System	System	**Unlimited**

Driving automation systems of SAE Level 0–2 are already available on the market. Active safety systems are even required by law for new registrations. Numerous driving assistant systems are in the product portfolio of all automotive manufacturers, and level-2 systems are also offered by most of them. CAD, which is the subject of this thesis, is the next automation level that will enter the market and is announced by the Original Equipment Manufacturers (OEMs) for the 2020s. For the first generation of level-3 driving automation systems, conceivable operating areas are certain freeway sections. Level 4 could follow towards the end 2020s, potentially in the form of robot taxis. It is only the introduction of level-5 driving automation that is still questionable. The additional requirement of an unlimited operating area represents a great if not unaffordable challenge.

3.3 The Transition from Conditionally Automated Driving to Manual Driving

From a user-centered perspective, one key element for a safe and efficient operation of a CA is the appropriate design of the transition phase from CAD to manual driving, during which the user becomes the driver again. Transitions can be driver-initiated—when the user decides to continue with manual driving—or system-initiated—when the system is not able to further perform the DDT. Both transitions must be controllable for the user, but the design requirements are significantly different. The system-initiated take-over issued by a RtI is subject of this thesis and is discussed in the following sections. The corner cases of unintended functionality mentioned in the previous chapter are not considered.

3.3.1 Take-Over Scenarios

There are various reasons why a CA would issue a RtI. As mentioned above, one reason would be exceeding the ODD, another reason could be a technical failure or limitation. Limitations of the ODD could be heavy rain, night time illumination, missing lane markings or exiting the freeway (SAE International, 2018). Depending on the specific cause of exceeding the ODD, the temporal criticality of the take-over ranges from low (the freeway exit is already known since the beginning of the trip) to high (suddenly occurring downpour). Despite being within the ODD, the performance of CA is limited by the performance of its system components or subsystems. Sensors or other elements may degrade due to electric/electronic faults, software shortcomings or sensor problems (e.g., blindness, de-calibration or misalignment). Degradation may also occur due to natural physical (performance) limitations, for instance, in terms of sensor coverage or processing power. CA should be designed to be capable of self-diagnostics, e.g. by redundant sensor sets and reciprocal plausibility checks, and it should detect all failures that represent a system limit and require drivers' intervention or at least drivers' further interpretation to maintain a safe operation.

Many of the abovementioned reasons for a RtI have been tested in human factors studies on CAD. The testing scenarios used for evaluating take-over performance and take-over behavior varied strongly between the various options. Gold, Naujoks, Radlmayr, Bellem, and Jarosch (2018) reviewed numerous existing studies and suggest structuring testing scenarios based on four main factors that determine driver's response and behavior: predictability, urgency, criticality and driver response complexity. Gold et al. (2018) propose a three-level ranking (low, medium, high) for each of the factors (see Table 3). Predictability describes how far in advance a take-over can be detected, for instance, based on available information from the backend. Urgency indicates the required reaction time of the driver and is associated with the time budget that the driver has for taking over vehicle control. The time budget represents the time provided to the driver for the take-over from the moment of the RtI to the system limit (Gold, 2016). The criticality of a scenario "scales the cost when failing to take over vehicle control in time" (Gold et al., 2018, p. 555). The cost when failing may be low, for example when missing a freeway exit, or severe, when colliding with an obstacle. Driver response indicates the complexity of the required action to respond to a RtI and to resolve the situation. Complexity ranges from low, when only a vehicle stabilizing maneuver is required because the freeway ends in a few kilometers, to high, when an evasive maneuver has to be performed due to an obstacle on the road.

Table 3. The four main factors for structuring take-over scenarios and their three increments according to Gold et al. (2018).

Factor	Low	Medium	High
Predictability	Near-term detection of the system limit	Predictable, but occurrence dependent on situational conditions	Known from backend, deposited map, V2V-communication
Urgency	High time budget	Medium time budget	Small time budget
Criticality	Low safety risk	Increased safety risk	High safety risk
Driver response	Low complexity	Medium complexity	High complexity

Human factors research questions on CAD vary strongly. Some studies focus on the dynamic configuration of driving maneuvers to prevent motion sickness and to generate maximum driver comfort. Other studies test the usability of the human-machine interface (HMI), and again others concentrate on evaluating system controllability and examine drivers' maximum possible performance (Gold et al., 2018). Depending on the leading question of the study, the manifestation of the four factors of the testing scenario is adapted. For example, Gold et al. (2018) recommend choosing demanding testing scenarios with high urgency, low predictability, high criticality and a medium to high driver response complexity to evaluate maximum driver performance. A typical example for scenarios that meet these requirements is a danger zone or an obstacle ahead in the driver's own lane, which is detected by on-board sensors and is not known from the backend (Gold et al., 2018).

3.3.2 Take-Over Process and Performance Measures

When driving in a conditionally automated mode, the driver is responsible for being receptive; however, it is to be expected that she/he is out-of-the-loop, with the eyes off the road, cognitive and visual attention completely on a NDRA. In this state, the driver is very unlikely to have an appropriate situation awareness, meaning that she/he has not perceived and comprehended the relevant elements of the environment and the driving situation (Endsley, 1995). These steps are essential to project the driving situation to the near future and to derivate an appropriate decision (Endsley, 1995). Once a RtI is issued, the driver has a situation-specific time budget to resolve the situation. The main goal of the take-over process is to shift the driver's attention to the driving task, to get the driver back in the loop by building up situation awareness and to initiate an appropriate action within a specific time frame (see Figure 3-2).

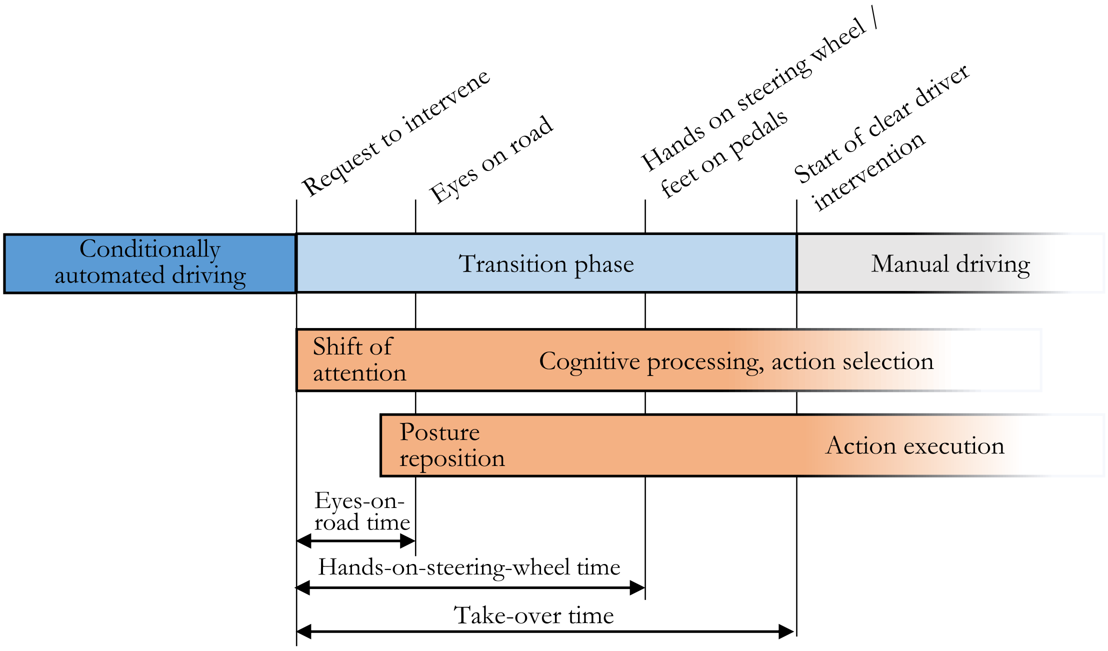


Figure 3-2. Schematic take-over process adapted from Petermeijer, de Winter, and Bengler (2016), Marberger et al. (2018) and Gold, Damböck, Lorenz, and Bengler (2013).

Therefore, multiple steps are required, which can be traced back to the information processing model by Wickens et al. (2013) (Gold, 2016; Körber, 2018; Petermeijer et al., 2016), introduced in chapter 2. The first step is the sensory perception of the RtI. Displaying it on multiple channels (e.g., visual and auditory) is recommended in order to create a rather salient stimulus (Schlake, 2019a). Furthermore, the vestibular (Schlake, 2019b) and tactile (Petermeijer, 2017) sensory channel has been evaluated to further improve the salience and efficacy of a RtI. If the salience is high enough, the driver turns her/his attention automatically towards the road (Zeeb, Buchner, & Schrauf, 2015). With the help of the working and long-term memory (Wickens et al., 2013), all gathered information about the situation are processed and interpreted, which represents a consciously controlled, effortful and capacity-limited process (Körber, 2018). Depending on how frequent the take-over scenario has already been experienced and how demanding it is, the driver must rely on knowledge-based, rule-based or even skill-based behavior and requires less resources (see chapter 2). While the information is being processed and interpreted, drivers automatically regains their normal driving position by adjusting their posture, placing their hands on the steering wheel and their feet on the pedals to be ready for the selected action. At the end of this process, a specific action is executed. Since highly demanding take-overs should be avoided by all means, a beneficial training effect can be expected for the automatic take-over actions rather than for the more complex ones.

All metrics that serve to evaluate the success of a take-over are summarized under the term **take-over performance**, consisting of timing and quality aspects (Gold, 2016). Both dimensions are important for a conclusive and holistic examination (Gold & Bengler, 2014). Timing aspects comprise the duration of both types of actions: the automatic and the conscious ones. Automatic actions are mainly required in the beginning of a take-over to shift (visual) attention and to regain the normal driving position (e.g., **gaze reaction time, eyes-on-road time, hands-on-steering-wheel time, feet-on-pedal time**). Conscious actions aim at resolving the situation (e.g., **time to mirror check, take-over time**). Times are always measured starting from the RtI (see Figure 3-2). The most frequently used timing metric and probably the most important one is the **take-over time** (TOT), because it consolidates the durations of all preceding steps (Körber, 2018). The TOT indicates the first conscious input of the driver aiming at finally taking over the driving task by, for instance, exceeding either a certain steering wheel angle or a certain change of pedal position (Gold, 2016) or by only pushing the deactivation button. This metric is also important, because the time span that the driver needs to take over will determine the system requirements regarding sensor ranges and processing power. Quality aspects comprise all metrics that consider how a take-over has been resolved and how safe it was (Körber, 2018). The metrics to be chosen strongly depend on the situation. For demanding situations, in which an active maneuver is required, the most important metrics are the **maximal occurring accelerations (longitudinal and lateral)**, **the minimal time to collision** (TTC) and the **occurrence of a crash** (Gold, 2016). The TTC is defined as theoretically remaining time until a potential collision with an obstacle, assuming constant speed of the obstacle and the ego-vehicle (Gold, 2016). The TTC is a classic metric to assess the criticality of a situation and, indirectly, to assess traffic safety, because it also reveals near misses (Gold, Happee, & Bengler, 2017; Körber, 2018). In less demanding situations, in which the requirement is only to stabilize the vehicle in its own lane, the take-over quality is usually assessed by e.g. the **standard deviation of the lateral position** or the **steering wheel reversal rate** (Gold, 2016). There are several publications that have reviewed and explicitly described the majority of existing take-over performance parameters (Gold, 2016; Kerschbaum, 2017; Körber, 2018; Radlmayr, Ratter, et al., 2019). Therefore, apart from the abovementioned examples, no further parameters of take-over performance are repeated here. Those used in this thesis are listed and described in more detail in chapter 6.5 (cf. Table 9).

All metrics for assessing take-over performance have in common that they have to be interpreted relative to the specific situation (Gold, 2016). Consequently, an absolute comparison of the metrics between different take-over situations is often not possible or not useful. High accelerations, for example, could in principle indicate a maneuver of low quality. However, if the time budget is rather low, a crash with an obstacle can only be avoided with a dynamic evasive maneuver resulting in high accelerations. In that case, the take-over would be considered successful and of rather high quality. Furthermore, due to this breakdown of take-over performance into multiple metrics, it is often difficult to come to a conclusive judgment, since the metrics may not show consistent findings. Typical cases would be take-overs with a short take-over time but high accelerations or a long take-over time and a very low TTC. An approach to classify a take-over as successful/unsuccessful is to predefine pass and fail criteria for each single situation (Damböck, Farid, Tönert, & Bengler, 2012). However, it is not always possible to determine a normative (ideal) behavior for complex scenarios with various courses of action, and the result of the classification obviously depends strongly on these predefined criteria.

A further approach to compensate this weakness is the Take-Over Controllability (TOC) rating developed by the Würzburg Institute for Traffic Sciences (WIVW) GmbH as part of the KoHAF project funded by the German Federal Ministry of Economics and Technology (Naujoks, Wiedemann, Schömig, Jarosch, & Gold, 2018). Based on a standardized scheme, the take-over behavior of the driver, which is recorded in video sequences, is assigned to one global value between 1 and 10 by trained experts (Naujoks et al., 2018). 10 ("not controllable") would be the loss of vehicle control or a collision and 1 would be perfect take-over performance. The levels in between are classed in categories: 7–9 "dangerous/unacceptable risk", 4–6 "driving errors/acceptable risk" and 2–3 "no driving errors/imprecisions" (Naujoks et al., 2018). Jarosch and Bengler (2019a) and Jarosch and Bengler (2019b) applied the TOC rating in their experiments to evaluate take-over performance. They found that the TOC rating is an adequate tool to rate controllability in addition to classic take-over parameters. However, take-over performance is not completely represented, because situation-specific evaluation criteria are not considered. For instance, if the complexity of the scenario is high and many options to respond to the situation are available to the driver, the number of possible driving errors depends on the specific maneuver, and therefore raises the probability of the driver committing these errors.

A further approach to condense different metrics into a global value was used by Radlmayr, Ratter, et al. (2019). They integrated the most important take-over metrics into three dimensionless parameters: The Vehicle Guidance Parameter (including crash, TTC, maximal lateral and longitudinal acceleration), Mental Processing Parameter (lane check, gaze reaction time, eyes-on-road time, take-over time) and the Subjective Rating Parameter (perceived criticality, perceived complexity of the situation, subjective time budget). Even though a reduction of parameters to three take-over performance parameters was successful and a reasonable and transparent way to describe take-over performance was found, the issue of situation dependency also remains in this approach.

So far, no method has been found in literature to obtain an overall score for take-over performance which is comparable over all different take-over situations and is suitable for all study goals. However, it is questionable whether such a method is affordable or necessarily required, because the various metrics give a detailed insight into many aspects of the take-over, which to some extent is lost when reducing all information to one or only a few metrics. Nevertheless, the overall evaluation and interpretation of take-over performance remains with the expert who analyzes it in relation to the specific situation.

3.3.3 Influencing Factors on Take-Over Performance

There has been a lot of research in the last few years that aims to examine what determines the take-over performance of drivers. Critical and urgent situations in particular were in the focus, because these situations will probably determine the functional characteristics and ODD of level 3. In a novel approach, Gold (2016) analyzed 735 time-critical take-overs (time budget between five and eight seconds), using multiple regression to predict the effect of individual factors on take-over performance. All take-over scenarios were designed with an obstacle blocking the driver's own lane as the reason for the RtI. Gold (2016) identified situational as well as driver-related factors as significantly relevant for take-over performance. The results are in line with the findings of numerous previous studies. The most important influencing factors are briefly summarized in the following. For the detailed modeling approach, see Gold (2016) and Gold et al. (2017).

- **Time budget** determines the urgency of a take-over (as introduced in chapter 3.3.1) and was found to be one of the strongest factors affecting take-over performance. The lower the time budget, the higher the time pressure on the driver and the smaller the TOT. However, a quick take-over leads to a decrease in take-over quality: The reduction of the time budget results in higher accelerations and shorter TTC (less safe take-over), even though a significant increase of crash probability is not evident. These findings are in line with numerous other studies that examined the effect of time budget (e.g., Damböck et al., 2012; Damböck, 2013; Gold et al., 2013; Zeeb et al., 2015).
- **Traffic density** mainly determines the complexity of the required maneuver (see chapter 3.3.1), thus strongly reducing all parameters of take-over performance. The higher the traffic density, the more information about surrounding vehicles needs to be perceived and processed. The worst take-over performance manifests when the traffic density is medium, because the course of action is not evident at first glance. Based on existing gaps between vehicles, the driver has to evaluate whether a lane change or a full-braking maneuver is more sensible. This decision-making process takes time, which is then missing for a successful maneuver. When the traffic density is high and adjacent lanes are blocked by other vehicles, it is more obvious that no evasive maneuver is possible. On the other hand, when traffic density is low and the adjacent lanes are free of traffic, no additional objects have to be perceived and processed. Findings of the modeling approach are in line with other studies that examined the effect of traffic density (e.g., Gold, Körber, Lechner, & Bengler, 2016; Körber, Gold, Lechner, & Bengler, 2016; Radlmayr, Gold, Lorenz, Farid, & Bengler, 2014).
- **Training** the take-over significantly improves drivers' take-over performance. The more frequently a take-over situation is experienced, the lower the TOT, the lower the crash risk and the greater the TTC. Like other learning effects, the improvement curve of drivers' performance by repeating take-overs is capped and can be described by a logarithmic function. Other authors also report a positive effect of the repeated

experience of take-overs on the driver's performance (Hergeth, Lorenz, & Krems, 2017; Körber et al., 2016; Petermann-Stock, Hackenberg, Muhr, & Mergl, 2013).

- **Lane**: Similar to the traffic density, the lane on which the vehicle is located when the RtI is issued affects the complexity of the scenario. The highest complexity is induced when the driver is in the middle lane of three lanes in total. However, unlike traffic density, lane only has a slight negative effect on take-over performance (TOT, lateral acceleration and TTC), which potentially became evident due to the large sample. Other studies did not find this correlation (Gold, Lorenz, & Bengler, 2014; Radlmayr et al., 2014).
- **Age**: Even though age affects almost all take-over performance parameters (TOT, longitudinal and lateral acceleration and TTC), this effect is always relatively small as already reported for lane. While the lowest TOT is achieved by middle-aged drivers, elderly drivers have the greatest TTC. Gold (2016) argues that elderly drivers probably compensate age-related limitations by a strong brake application, which ultimately increased the take-over quality. This argumentation is in line with the results of other studies (Körber et al., 2016).
- **NDRA**: The most inconsistent effect is found for the load induced by different types of NDRAs. Even though an effect of NDRAs was only found for the TTC—and here, as expected from other study findings (Gold, Berisha, & Bengler, 2015; Radlmayr et al., 2014), this effect was rather small—the direction of the effect was unexpected: TTC increases with the load of the NDRA. Gold et al. (2017) argue that this may result from an additional activation of attentional resources due to a higher load or an overcompensation of missing situation awareness through a stronger brake reaction. Other studies also account for the attentional theory: additional activities improved the take-over performance compared to no activities since a specific level of arousal could be maintained (Neubauer, Matthews, & Saxby, 2012) (see chapter 4.3 and 4.6). In a more general sense and as a result of the national research project KoHAF, Naujoks, Befelein, Wiedemann, and Neukum (2018), Marberger et al. (2018) and Jarosch, Gold, et al. (2019) propose that not only the cognitive state needs to be ready for a take-over but also the sensory (gaze back on the road) and the motoric state (freeing the hands). Arousal and motivation may affect all three aspects, and all the named aspects have to be considered when examining the effect on take-over performance.

Some of the suggested factors could explain a large degree of variance in take-over performance, whereas others could only explain less. Gold (2016) proposes that additional variance may be explained by considering other factors: he suggests further examining, for instance, the effect of different driver states prior to the RtI or the effect of individual driver characteristics and driver skills. Furthermore, the design of the HMI and the remaining support of the driving task by the system during the transition phase may significantly affect take-over performance (Gold et al., 2014; Kerschbaum, 2017; Petermeijer, 2017; Yang, Karakaya, Dominioni, Kawabe, & Bengler, 2018).

4 The Concept of Driver Fatigue

While research and findings on technical and human factors aspects of CAD are still emerging, the issue of driver fatigue is almost as old as driving itself. Due to its relevance for road safety in both professional and non-professional drivers, fatigue at the wheel has been a highly researched topic for the past decades—or even for the last century. Research findings assumed to be most relevant for this thesis are compiled in this chapter. Before defining how driver fatigue is understood in this thesis, it is initially explained how driver fatigue is incorporated into the construct of driver state since this is an umbrella term often used in the context of fatigue. After discussing causes, concrete effects and assessment methods for driver fatigue, the chapter concludes with a literature review in the context of fatigue in CAD.

4.1 Incorporation in the Construct of Driver State

As mentioned in chapter 3.3.3, Gold (2016) suggests generating knowledge about the driver state, because it could potentially be an important influencing factor on take-over performance and may explain 30–40 % of variance. But what exactly is the construct of driver state?

According to Langer, Abendroth, and Bruder (2015, p. 688), all time-variant characteristics of the driver that may be relevant for driving behavior or driving performance are summarized under the term driver state. These characteristics may be variable in the short or middle term, for instance, driver's situation awareness, vigilance, fatigue or being under the influence of alcohol (Langer et al., 2015, p. 688). Characteristics that cannot be changed at all, are difficult to change or that are only changeable in the long term, for instance, the risk-taking propensity of a driver or driving skills, are referred to as driver traits. Both—states and traits—affect the driver's behavior, but only the states (and in particular the state fatigue) are considered in this thesis. It needs to be pointed out, however, that the differentiation of states and traits is not always clearly recognizable and the boundaries are often fluid. For example, trust in automation can be a trait on the one hand, meaning that a person can have a very trustful basic attitude towards automation technologies. On the other hand, the same person can develop a great mistrust in automated driving due to her/his recent experience with such a system, for example, because she/he experienced one or multiple critical system boundaries.

Approaches to model the driver state

There are different approaches to consolidate driver state factors into a framework. In a literature review, Hecht et al. (2019) have collected the three most important ones with regard to automated driving, which are outlined in the following.

A rather general, literature-based model was provided by Stanton and Young (2000). They assembled eight psychological factors reported in literature in a unique structure and hypothesized the relationship between them based on previously known individual links.

Factors are situation awareness, mental workload, mental model, feedback, locus of control, stress, task demands and trust. This model was revised and extended by Heikoop, de Winter, van Arem, and Stanton (2015) by adding seven further potentially relevant factors in order to address longer-term effects. These factors are attention, vigilance, satisfaction, acceptance, arousal, complacency and fatigue. With the help of a simplified literature review, Heikoop et al. (2015) selected nine human factors most relevant for research on automated driving and re-assembled them into a model (cf. Figure 4-1). Remaining factors are mental workload, attention, feedback, stress, situation awareness, task demands, fatigue, trust and mental model. According to the model, fatigue is negatively affected by task demands and has a U-shaped causal link to mental workload. Furthermore, fatigue and stress are mutually dependent and can promote each other. Because this approach has several limitations, such as the selection process of the nine most relevant factors or the proliferation of the constructs in the reviewed articles, it represents a good overview of the current research status regarding modelling the driver state rather than a holistic model (Hecht et al., 2019).

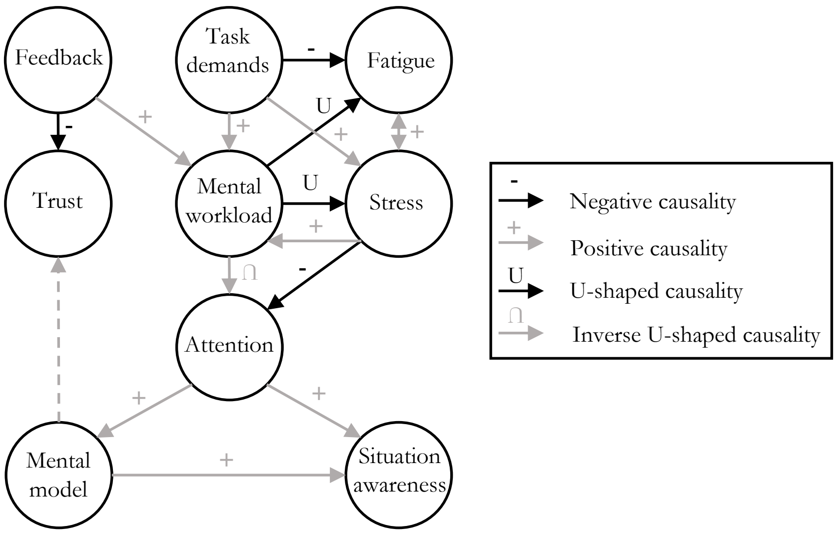


Figure 4-1. Driver states according to Heikoop et al. (2015).

A further, more simplified approach to provide a model for the factors of driver state was initiated by the EU-funded project HAVEit (Rauch, Kaussner, Boverie, & Giralt, 2009). This model is motivated by addressing the main factors that limit driver's performance capabilities. Rauch, Kaussner, Boverie, and Giralt (2009) argue that both the energetic state and the attentional state must be ideal to achieve a driver state optimal for maximum driving performance. In their opinion, the energetic state includes the alertness or arousal level, which represents the individually available capabilities or resources of the driver to manage the respective task demand while driving. The arousal level is negatively affected by vigilance decrements, fatigue and drowsiness/sleepiness. The attentional state is determined by the selective, task-oriented attentiveness, that is directed towards the primary driving task (see chapter 2) and the relevant cues in the environment. Attentiveness is negatively influenced by distraction from the actual driving task.

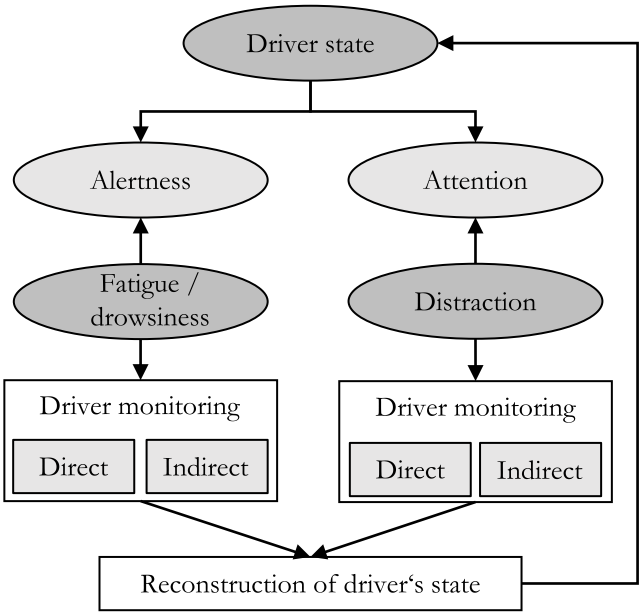


Figure 4-2. Driver state model according to Rauch, Kaussner, Krüger, Boverie, and Flemisch (2009).

The model of Marberger et al. (2018) focusses even more strongly on automated driving, since it deals with the specific requirements of the driver state right before or during the transition from CAD to manual driving, which they refer to as driver availability. The model assumes a target state in which the driver must be right before or during the take-over for a safe and successful take-over (see Figure 4-3). The model takes into account that take-over performance strongly depends on situational and personal factors (see chapter 3.3.3). Therefore, the target state varies for each individual and each situation. To achieve this target state, a driver needs to undergo a transition process, in which the different state components—sensory, motoric and cognitive ones—are converted for the requirements of the take-over situation. For instance, the sensory state might be converted by looking back to the road from any NDRA, the motoric state by re-placing the hands on the steering wheel and the cognitive state by re-configuring mental task sets or response rules. The re-configuration of each single state takes time, and the actual duration depends on—apart from personal skills—the arousal level (including, among others, fatigue) and the motivational condition of the driver. All these three factors also interact with the type of a potential NDRA the driver is engaged in. Depending on whether the transition duration exceeds the specific available time budget of the take-over situation or not, the driver is considered available or unavailable.

As a result of their literature review, Hecht et al. (2019) also selected certain factors to be most relevant for automated driving. These are vigilance, attention, stress, mental workload, and fatigue.

As we can see, due to the complexity of this topic, there is no consensus about what belongs to the driver state construct. It depends on the model approach and the use case how many and which kind of factors are incorporated and how they correlate. However, most of the models have in common that they consider fatigue an important factor for both driving and take-over performance. There are also other factors, which have not been mentioned yet, but which potentially influence take-over performance and may be equally essential as fatigue. Factors may be the general driving skills of a person, the automation experience, the mode awareness or the trust in the automated system, to name only a few. The focus of this thesis is on driver fatigue, as introduced in chapter 1.1, though, so only this factor and some closely related ones are discussed in more detail in the following chapters.

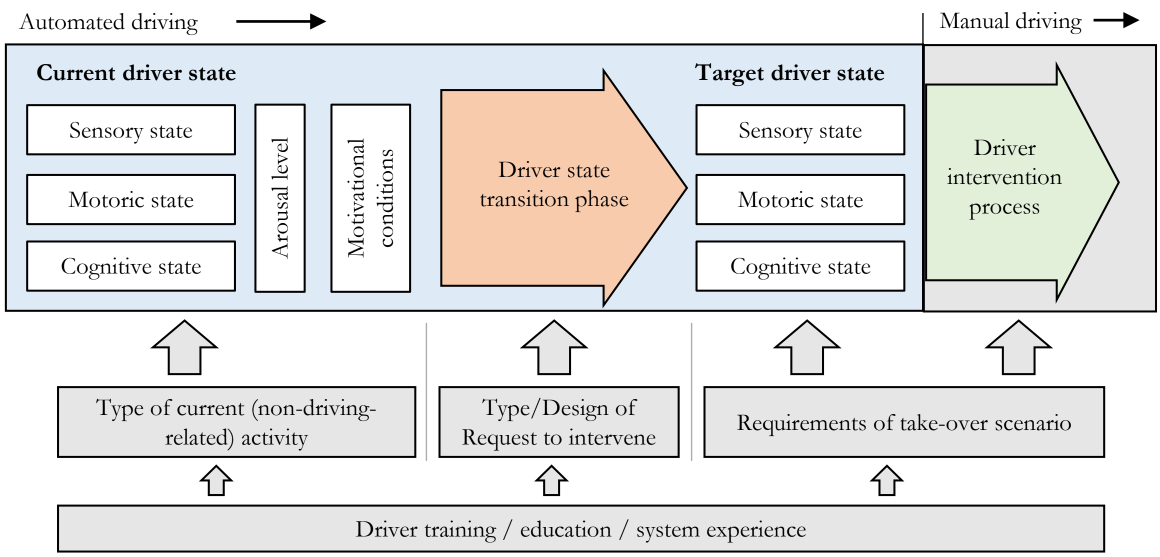


Figure 4-3. Driver state model according to Marberger et al. (2018).

4.2 What is Fatigue?

Before going into detail on causes, measuring possibilities and effects of driver fatigue in manual driving and CAD, it needs to be clarified what fatigue is.

Fatigue is a complex concept that has been difficult for researchers to define in the past, and terms such as fatigue, sleepiness, tiredness or drowsiness seem to hardly differ from each other. Fatigue is often reduced to a subjective feeling. Brown (1994, p. 302), for instance, describes fatigue as “a subjectively experienced disinclination to continue performing the task because of perceived reductions in efficiency”, and also Shen, Barbera, and Shapiro (2006, p. 70) define fatigue as “an overwhelming sense of tiredness, lack of energy and a feeling of exhaustion, associated with impaired physical and/or cognitive functioning”. However, by considering only the manifestation of fatigue in a subjective feeling, other effects of fatigue—namely on physiology and performance—are not taken into account. Therefore, Phillips (2014) suggests that fatigue may best be defined in a more general sense as a multidimensional construct with experiential, physiological and performance aspects.

However, how should fatigue be differentiated from the abovementioned related terms, or are they all identical? Johns (2007) claims that due to English dictionary definitions both the terms fatigue and tiredness can be used synonymously, as can the terms drowsiness and sleepiness. He further states that drowsiness/sleepiness and fatigue/tiredness are different concepts and have to be distinguished. For Johns (2000, p. 242), drowsiness is an "intermediate state between alert wakefulness and sleep", whereas fatigue is "a subjective state of weariness, often with muscle aches or discomfort, emotional irritability and a disinclination to continue activities" (Johns, 2007, p. 3). On the other hand, Desmond and Hancock (2001) use a very similar definition for fatigue to the one that Johns (2000) uses for drowsiness: They take "fatigue to be a transition state between alertness and somnolence" (Desmond & Hancock, 2001, p. 459). And also Lal and Craig (2001, p. 175) assume fatigue to be "the transitory period between awake and sleep". Gimeno, Cerezuela, and Montanes (2006) suggest that drowsiness is an accompanying feature of mental fatigue and the two phenomena often but not always occur simultaneously. Sleepiness is often understood as a "craving or desire for sleep" (Dement & Carskadon, 1982, p. 57) where the pressure to sleep is increased (Mullins, Cortina, Drake, & Dalal, 2014). Sleepiness results from two main factors: the homeostatic factor that involves sleep quality and quantity, and the circadian rhythms, i.e. the time of day (Mullins et al., 2014; Phillips, 2014). Fatigue, by contrast, may appear independently of these factors and have its main origin in exertion and time-on task (Mullins et al., 2014). Like drowsiness, neither is sleepiness necessarily accompanied by fatigue, but it may be. Furthermore, there are numerous reputable publications that use—unlike Johns (2007) and others—the terms fatigue and drowsiness completely synonymously without any explanation (Brown, 1994; Eskandarian, Sayed, Delaigue, Blum, & Mortazavi, 2007; Knipling & Wang, 1994; May & Baldwin, 2009; Rauch, Kaussner, Boverie, & Giralt, 2009). Others declare that the difference between these states is negligible (Schultz & Young, 2007) or the constructs overlap so extensively that the terms can be used interchangeably (Dinges, 1995).

The question of how to differentiate between fatigue and drowsiness/sleepiness cannot be settled in a final way, because there is no consensus on it in the literature. Some researchers make no distinction between the constructs, others advocate a strict distinction, again others consider drowsiness/sleepiness a sub-dimension of fatigue (Matthews, Desmond, & Hitchcock, 2012, p. 141). Based on the literature review conducted, the latter seems to be reasonable. That is why the view is shared in this thesis that sleepiness/drowsiness may but does not have to occur at the same time as fatigue, but if they do, they are hardly distinguishable (Phillips, 2014). However, most of the researchers in this field are agreed that both constructs are equally unfavorable with regard to task performance, in particular for driving performance, which is the emphasis of this thesis. Therefore, in this thesis, sleepiness/drowsiness is seen as a specific case of fatigue, and no explicit distinction is made between the two constructs following the proposal of Dinges (1995) and Radlmayr, Feldhütter, et al. (2019). The focus of this thesis is on the effects of fatigue on different take-over performance parameters. For the sake of simplicity and for readability reasons, both constructs are summarized under the term *fatigue,* which will be used hereinafter. Deviations from this approach only occur if the original term is part of a scale description or part of a method name, or the like. In these cases, the original term is used instead of the summarizing term fatigue.

4.3 Differentiation from Related Constructs

Arousal and Alertness

Fatigue is frequently related with the terms arousal or alertness. According to Cohen (2011a, p. 247) , arousal is a "psychological and physiological state of wakefulness, excitement, and/or activation [that] enables readiness for action". From a neuronal perspective, arousal refers to the activity level of the central nervous system and is responsible, inter alia, for increased alertness and automatic responses (Cohen, 2011a, p. 247; Lal & Craig, 2001). Alertness, in turn, is a "state of being mentally perceptive and responsive to external stimuli" and represents a "readiness to respond" (Loftis, 2011, p. 77). Therefore, arousal can be considered the prerequisite for alertness (Lal & Craig, 2001). The ability to stay receptive to stimuli and to react to them appropriately is essential for safe driving, so both a certain level of arousal and of alertness is required. This also applies to CAD, since a RtI represents such an external stimulus, and an adequate reaction is required. According to the Yerkes-Dodson Law (1908), the best task performance is achieved when arousal is at a medium level (Cohen, 2011b, p. 2737; Wickens et al., 2013, p. 362), where task performance and arousal level follow an inverted U-shaped relation (see Figure 4-4). Both excessively high and excessively low arousal levels lead to a decrease in task performance. Stressors such as anxiety and noise are thought to increase the level of arousal, whereas factors such as sleep loss and fatigue will decrease arousal (Lal & Craig, 2001, 2002; Rauch, Kaussner, Boverie, & Giralt, 2009; Wickens et al., 2013, p. 362).

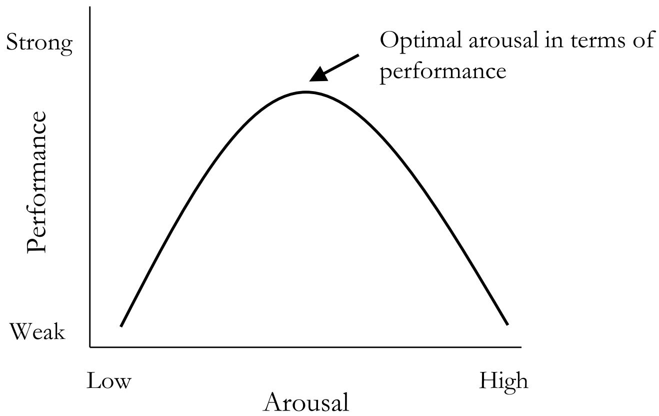


Figure 4-4. Correlation between Arousal and Performance (Yerkes-Dodson Law).

Vigilance

Vigilance is also often associated with fatigue. Vigilance is defined as "the state of readiness to detect and respond to certain specified small changes occurring at random time intervals in the environment" (Mackworth, 1957; as cited in E. A. Schmidt et al., 2009, p. 1087) or, in a more general sense, vigilance is "the ability to sustain attention to a task for a period of time" (Parasuraman, Warm, & See, 1998; as cited in E. A. Schmidt et al., 2009). Comparing the definitions with the ones of arousal and alertness, it appears that there is a close relation of these concepts. Some authors even argue that from a neurobiological perspective, the terms vigilance and arousal/alertness can be used interchangeably because they have similar mechanisms on the central nervous system level (Lal & Craig, 2001). Like arousal and alertness, sustained attention is also essential for safe driving, and it has often been reported that fatigue can produce impairments in vigilance (Lal & Craig, 2001; E. A. Schmidt et al., 2009).

Summarizing, it can be noted that the concepts of vigilance, arousal and alertness are negatively affected by conditions where only few stimuli are present over a prolonged time period, for instance during monotonous drives, resulting in states of hypovigilance (Larue, Rakotonirainy, & Pettitt, 2010) and low arousal (Lal & Craig, 2001). This, in turn, may act as a promoter for fatigue because the effort to maintain a certain level of vigilance and arousal is high. Conversely, it was reported from studies with sleep deprivation that already existing fatigue may have a negative effect on vigilance and arousal/alertness (Lal & Craig, 2001; Phillips, 2014). In this case, the effort to maintain a certain level of arousal/alertness and vigilance apparently cannot be raised, so that stimuli are not detected and performance cannot be delivered.

4.4 Cause of Driver Fatigue: Model by May & Baldwin

In literature, there are mainly two leading causes of driver fatigue, namely sleep-related and task-related ones. In order to depict these two main causes and their interaction, May and Baldwin (2009) developed an important model based on the considerations of Desmond and Hancock (2001) (see Figure 4-5). Sleep-related (SR) fatigue, which is often referred to as sleepiness (see chapter 4.2), has its origin in homeostatic and circadian factors (Phillips, 2014). Homeostatic factors include sleep duration, sleep quality and time spent awake since last sleep, and determine the extent of pressure for sleep an individual has. The less sleep we get at night, if sleep is disturbed or we stay awake until early in the morning, the more sleepiness increases. However, the time of day also influences sleepiness or SR fatigue due to an individual internal clock referred to as circadian rhythm. This circadian factor causes peak alertness towards late afternoon or early evening and peak sleepiness in the early hours of the morning. (Phillips, 2014)

The other type of driver fatigue that is described in the model of May and Baldwin (2009) is TR fatigue. It results from demands of the driving task itself and/or the driving environment. Task demands can be either too high or too low, but both may lead to fatigue when the time on this task is extended. Other authors also mention time-on-task and resulting exertion, among others, as a factor causing fatigue (Karrer-Gauß, 2011; Phillips, 2015; Rauch, Kaussner, Boverie, & Giralt, 2009). May and Baldwin (2009) further distinguish TR fatigue according to too low and too high task demands. Active TR fatigue results from mental overload, which emerges from complex tasks with high demands during driving including, for instance, high traffic density, poor visibility or the need to complete a secondary task. By contrast, if fatigue is induced by permanent mental underload, May and Baldwin (2009) talk about passive TR fatigue. Mental underload may occur when the driver is only exposed to rare stimuli from the environment or from the driving task. This is the case, for instance, when driving in a monotonous environment where input by the driver is rarely required, especially when driving on long straight roads (Matthews & Desmond, 2002; Saxby, Matthews, Warm, Hitchcock, & Neubauer, 2013; Thiffault & Bergeron, 2003b), or when using (partly) automated systems (Desmond & Hancock, 2001, p. 462; May & Baldwin, 2009; Neubauer et al., 2012; Saxby et al., 2013).

Both fatigue types—TR and SR—are often not independent of each other, but they may occur simultaneously or act as an accelerator for each other. For instance, it is conceivable that sleepiness due to sleep deprivation or a circadian low is exacerbated by a monotonous or demanding drive because of an even higher exertion to stay awake (Phillips, 2014; Williamson et al., 2011). This interaction is also depicted by the linking arrow from passive and active TR fatigue to SR fatigue in Figure 4-5.

In the context of automated driving, the increasing automatization of the driving task potentially prevents the development of active TR fatigue because the driver is released from the demanding driving task and workload is reduced (Saxby, Matthews, Hitchcock, & Warm, 2007). However, passive TR fatigue can even be promoted by vehicle automation due to the elimination of an active role. SR fatigue is also likely to occur, since users may activate CA after a night shift or in the circadian low after lunch. Drivers will probably use CA even more frequently if they realize that they are very fatigued, since they do not have to drive themselves.

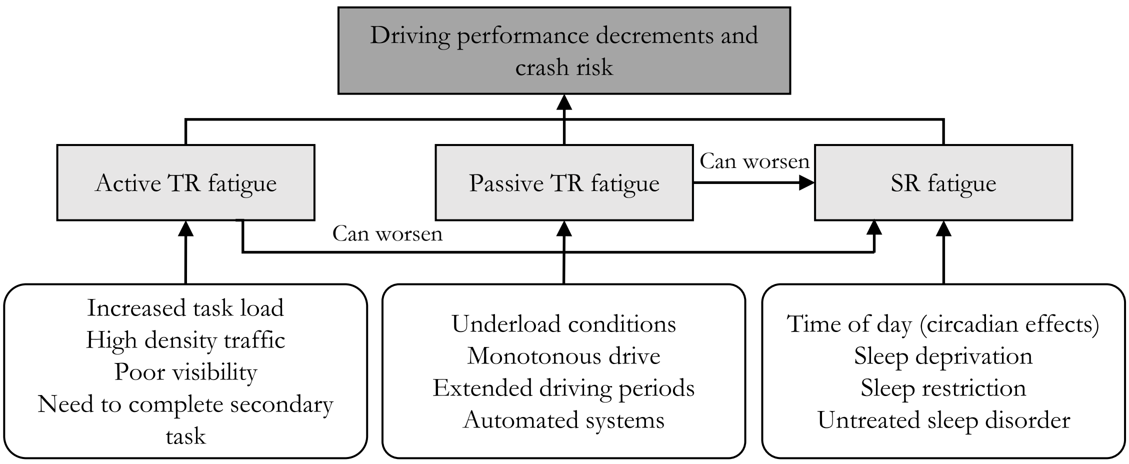


Figure 4-5. Fatigue model based on the underlying cause according to May and Baldwin (2009).

4.5 Effects of Driver Fatigue on Cognitive Functions and Performance

In general, fatigue draws on different mechanism of human information processing and cognitive functions which are described in chapter 2. Even though both types of fatigue, TR and SR, have been examined in the past in manual driving, the effect of SR fatigue on manual driving has been the object of investigation significantly more often. This thesis mainly focusses on TR fatigue. However, as mentioned in the previous chapter, SR and TR fatigue are frequently difficult to clearly separate. For instance, deficits due to SR fatigue (sleepiness) may be accelerated with time on a demanding task (Dinges, 1995). Many studies take advantage of this and mingle both types of fatigue on purpose, for instance, because only the general effect of fatigue was object of the study, or the studies just failed to avoid the respective fatigue cause. This may be because, for example, the circadian rhythm is rather individual and, therefore, its effect is hard to completely rule out in experiments, or drivers have to perform monotonous drives under the influence of sleep deprivation. Therefore, effects of the two fatigue types are often mingled, results are confounded (May & Baldwin, 2009; Phillips, 2014), and often no strict allocation of specific effects can be made.

The effects of TR fatigue on elementary cognitive functions and processes most frequently reported in studies and compiled in already existing literature reviews (see e.g., Brown, 1994; Dinges, 1995; Phillips, 2014; Wylie, Shultz, Miller, Mitler, & Mackie, 1996) include the following:

- Reduced alertness (e.g., readiness) (Phillips, 2014; Wylie et al., 1996), psychophysiological arousal (e.g., brain waves, heart action) and vigilance (e.g., watchfulness) (Wylie et al., 1996)
- Lapses in attention: difficulty to identify and process information from the environment (Brown, 1994; Dinges, 1995; Phillips, 2014; Wylie et al., 1996), mind wandering (Matthews, Desmond, & Hitchcock, 2012, p. 147), *look but failed to see* phenomena (Phillips, 2014) and "driving without awareness" (Brown, 1994, p. 312)
- General cognitive slowing: slowed information processing and increased decision-making time (cognitive throughput) (Brown, 1994; Dinges, 1995; Phillips, 2014; Wylie et al., 1996)
- Reduced accuracy of short-term memory (Dinges, 1995; Phillips, 2014)

Apart from decrements in information and cognitive processing, TR fatigue may also have an impact on the subjective experience level such as negative emotions (Matthews, Desmond, & Hitchcock, 2012, p. 147), decreased motivation to sustain performance (Wylie et al., 1996) and reduced motivation/disinclination to continue a task (Brown, 1994; Matthews, Desmond, & Hitchcock, 2012, p. 147). The described decrements may manifest in general driving performance measures, such as increased reaction times (to critical events) (Dinges, 1995; Phillips, 2014; Wylie et al., 1996) or variation in reaction times (Phillips, 2014), more variable and less effective control responses (Wylie et al., 1996) as well as more frequent errors or skill degradation (e.g., poor lateral and longitudinal control) (Phillips, 2014). Complex tasks are affected more strongly than simple or well-learned ones, where performance can be maintained even for a prolonged period of time and with high levels of fatigue (Phillips, 2014). Phillips (2014) argues that the sensitivity of the cognitive functions and performance decrements due to fatigue varies and that it is often not clear, yet, why precisely these effects occur (for instance memory lapses).

In the context of CAD, critical events may occur in the form of urgent vehicle-initiated RtI. A quick but nevertheless appropriate driver reaction is essential during such a take-over to maintain a general level of road safety independent of the situation's complexity. All abovementioned performance decrements potentially resulting from fatigue, which are extended response times and a degraded response quality, may have an additional negative effect on the general cognitive task switching costs that are associated with the transition from CAD to manual driving (Wickens et al., 2013, described in chapter 3.3.2). Therefore, fatigue can be assumed to be highly relevant for take-over performance and a safe transition from CAD to manual driving.

4.6 Individual Differences in Fatigue Development and Performance Decrements

Even though the performance decrements described in chapter 4.5 have been reported in many studies on driver fatigue, it has to be pointed out that these are potential consequences of fatigue but not obligatory ones when being exposed to a fatiguing task or when suffering from sleep deprivation. There are drivers who are greatly affected by the experiment's condition, while others are almost completely resistant to it, which leads to the conclusion that there are large individual differences in the vulnerability to fatigue (Desmond & Matthews, 2009; Thiffault & Bergeron, 2003a; Verwey & Zaidel, 2000; Wylie et al., 1996). There are even experiments with an ostensibly identical composition of task setting and participant population, which produce a wide range of inconsistent findings regarding fatigue development and performance variation (Ackerman, Calderwood, & Conklin, 2012, p. 92). There are different theories to explain this discrepancy all accounting for individual differences in fatigue susceptibility by applying different coping strategies. The most important ones are the compensatory control model developed by Hockey (1997), the transactional theory (original theory from Lazarus, 1999) and the dynamic model of stress and sustained attention of Hancock and Warm (1989). For a more detailed description of the theories and how they account for individual differences in fatigue in detail see, for example, Szalma (2012) or Phillips (2014). Even though these models focus on distinct aspects of human response to changes in environmental conditions, "they share a common, energetic, resource-based perspective that emphasizes the importance of considering both the specific properties of the environment and how people respond and adapt to the demands imposed on them" (Szalma, 2012, p. 76). Summarizing the key messages of the models, whether driver fatigue occurs or not and whether it affects driving performance or not depends on the particular task/environment characteristic (e.g., understimulating monitoring vs. overstimulating traffic condition or monotonous straight freeway vs. serpentine mountain road), the personality traits of the driver in general (e.g., the general amount of available resources, individual standard arousal level or the individual driving skills) and the transaction between the task and the personality (Ackerman et al., 2012, p. 99; Phillips, 2014; Szalma, 2012, p. 82). Especially the latter—the joint function of personality and task property—is probably the most significant predictor for fatigue vulnerability; however, it is also the most complex one because it has numerous facets and, by far, not all influencing factors are known, yet (Szalma, 2012, p. 88). For instance, there is the hypothesis that extravert individuals may be more resistant to overload conditions and the occurrence of active fatigue, since they are associated with greater resource capacity and higher energetic arousal. Reversely, extraverts would be more susceptible to driving conditions with underload and passive TR fatigue due to the missing stimulation and a greater effort to compensate this compared to introverts (Szalma, 2012, p. 84). A further manifestation of the complex transaction between task and personality causing different fatigue vulnerabilities is the individual motivation or intrinsic interest to perform a specific task (Ackerman et al., 2012, p. 99) or the personal performance goals for a task (Phillips, 2014). Ultimately, the individual motivation and choice of strategy to manage fatigue resulting from the specific interaction of task and personality is assumed to be a key moderator factor for individual differences in fatigue vulnerability (Matthews, Hancock, & Desmond, 2012, p. 166).

4.7 Countermeasures for Fatigue

According to May and Baldwin (2009), knowing the cause of fatigue is crucial for the choice of a suitable countermeasure. In principle, to prevent the occurrence of SR fatigue, the best and most obvious countermeasure as well as prevention is sleep (Williamson, 2012, p. 444). Driving should only be commenced when the driver is sufficiently rested (enough hours of sleep at night, adequate sleep quality) and not after a long time awake or during the individual circadian low (Williamson, 2012, p. 443), which is between midnight and dawn for most people (Horne & Reyner, 1995). Naturally, several factors may easily coincide and create particularly unfavorable conditions in terms of fatigue. Such conditions would be, for instance, driving in the early morning because it falls within the period of circadian low and, simultaneously, of being awake for a long time.

Since the abovementioned countermeasures are known to most drivers, but are often not feasible in real life, there are measures that support drivers during driving, helping them to manage their fatigue. For instance, structural measures were taken in order to either encourage drivers to take breaks (frequent "reviver" service stations) or to raise drivers' awareness that their driving ability has deteriorated (Job & Dalziel, 2001, p. 354; Williamson, 2012, p. 442). The latter aim is also pursued by in-vehicle technologies that either monitor the physiology of the driver or the driving performance, and give some sort of warning to the driver once an adverse condition is detected (see also chapter 4.8.4). Taking breaks is assumed to be the most promising short-term countermeasures for fatigue (Williamson, 2012, p. 447). The composition of the break may afford further opportunities to overcome fatigue for a certain time period. For instance, there is evidence that fatigue can be countered by physical exercise, naps or having caffeinated drinks (Williamson, 2012, 447-449). However, none of these approaches will have a long-lasting effect, but they may only prolong the functional capability of drivers for a certain time span.

Applying the Yerkes-Dodson Law

Another approach to prevent and counter fatigue resulting from enduring overstimulation by a continuously performed task, which is referred to as TR fatigue (May & Baldwin, 2009), is the elimination of the fatigue source, namely the task. In terms of active TR fatigue, automatizing the driving task relieves drivers from workload because they are supported when the drive is long or when there are highly demanding driving situations (May & Baldwin, 2009; Williamson, 2012, p. 447). For instance, de Winter et al. (2014) found that CAD significantly reduces drivers' workload, and Funke, Matthews, Warm, Emo, and Fellner (2016) showed that even simple cruise control usage improves the lane keeping performance of drivers. However, automated driving also entails a downside, since the risk of understimulation and resulting passive TR fatigue is increased. To counter passive TR fatigue—whether evoked by automated driving or by a monotonous road section—non-driving-related sources, such as additional tasks, may apply stimulation to the driver to maintain an ideal level of alertness (Yerkes-Dodson Law, see chapter 4.3). For instance, Baulk, Reyner, and Horne (2001) found that the stimulation provided by a simple auditory reaction time test resulted in less subjective and objective fatigue in a driving simulator study. In other driving simulator studies, demanding games were used to provide stimulation under monotonous driving conditions (Gershon, Ronen, Oron-Gilad, & Shinar, 2009; Verwey & Zaidel, 1999). Results demonstrated that participants playing the game were significantly less subjectively fatigued and experienced fewer sleeping episodes, which manifests in fewer safety-relevant driving errors and a better driving performance. A positive effect of a well-timed secondary verbal task on driving performance and neurophysiological alertness (assessed by brain activity) could also be demonstrated during fatiguing drives (Atchley, Chan, & Gregersen, 2014). However, Verwey and Zaidel (2000) argue that artificial secondary tasks do not necessarily have this positive effect, but it depends on the interaction between task and driver (does the individual find the task challenging, motivating or interesting; see also chapter 4.6) and on the voluntary choice of the task. Compulsory tasks may be fatiguing, whereas an activity chosen at will may have a "strong restorative effect", emphasizing the important role of autonomy on motivation and alertness (Szalma, 2012, p. 82). According to Ackerman et al. (2012, p. 99), if "there is a good match between the task demands and the individual's intrinsic interests", fatigue will be less likely to occur.

Especially in the context of CAD, where NDRAs are explicitly permitted, the driver may benefit from their stimulating effect in terms of countering fatigue. Stimulation can also be provided by aborting the CA and giving the driving task back to the driver. However, this approach is only partially effective for those cases in which CAD evoked fatigue.

4.8 Assessment of Driver Fatigue

Fatigue cannot be measured directly; it is only the effects which fatigue has on an individual and which result in specific indicators which can be operationalized by different methods or measures. As suggested by Phillips (2014), fatigue is a multidimensional construct that can manifest itself in experiential/subjective, physiological and performance characteristics. Phillips (2014) recommends that all three aspects should be measured in order to obtain a comprehensive picture of the status and the effect of fatigue. The subjective manifestation refers to the individually experienced or perceived feeling of fatigue which is assessed by means of questionnaires or scales. The physiological manifestation as well as task performance can be measured objectively with the help of sensors, camera-based technology, cognitive tasks or behavior observation.

The following methods are not necessarily only applied to driver fatigue, but are also suitable to operationalize fatigue in general. Only methods were selected that are also established in research on driver fatigue. It has to be anticipated that a gold standard for a completely reliable and accurate fatigue indicator measurement method is still missing, and just upcoming fatigue or low levels of fatigue, in particular, are extremely difficult to detect (Platho, Pietrek, & Kolrep-Rometsch, 2013).

4.8.1 Subjective Measures

There is a great selection of scales developed for a self-report on fatigue, which comprise different aspects that typically correlate with fatigue, e.g., cognitive, motivational and physical ones. Please see Phillips (2014) for a detailed overview of existing scales.

The **Fatigue Severity Scale** is probably the best known and most frequently used questionnaire with a large application range (Phillips, 2014), although the focus is more on clinical research and diagnostics. It is a unidimensional scale with nine generally formulated items including, among others, motivational and functional outcomes related to fatigue. Responses to the statements can be given on a seven-point Likert scale ranging from completely disagree to completely agree. To reflect the multidimensionality of fatigue and closely related states that affect performance, such as stress and arousal, Matthews et al. (2002) proposed the **Dundee Stress State Questionnaire** (DSSQ), which is completed prior to and after the task. The questionnaire consists of 96-item measures designed to assess transient states associated with stress, arousal and fatigue. Items are grouped into three higher-order factors associated with task engagement (energetic arousal, motivation, and concentration), distress (tense arousal, hedonic tone, and confidence and control) and worry (self-focused attention, self-esteem, and both cognitive interference scales) (Matthews et al., 2002). Both laboratory and field studies have shown the DSSQ to be sensitive to driver stress factors (e.g., Matthews & Desmond, 2002). To assess situational sleepiness and not overall general fatigue, there are two popular one-item scales that are very similar to each other and easy to apply. One is the **Karolinska Sleepiness Scale** (KSS) with a nine-point verbally anchored scale with the following steps: “extremely alert” (score = l), “alert” (3), “neither alert nor sleepy” (5), “sleepy—but no difficulty remaining awake” (7), “Extremely sleepy—fighting sleep” (9). The steps in between have a scale value but no verbal label in its original version (Åkerstedt & Gillberg, 1990). In an adapted version of Baulk et al. (2001), the intermediate steps were also provided with labels (for a comparison see Table 4). Miley, Kecklund, and Åkerstedt (2016) showed that the two versions do not lead to different results. The other one is the **Stanford Sleepiness Scale** (SSS) (Hoddes, Zarcone, Smythe, Phillips, & Dement, 1973). The SSS only has seven scores all anchored to detailed descriptions. The statements and their scale values are (1) "feeling active and vital; alert; wide awake", (2) "Functioning at a high level, but not at peak; able to concentrate", (3) Relaxed; awake; not at full alertness; responsive", (4) A little foggy; not at peak; able to concentrate" (5) "Fogginess; beginning to lose interest in remaining awake", (6) "Sleepiness; prefer to be laying down; fighting sleep; woozy", (7) "Almost in reverie; sleep onset soon; lost struggle to remain awake". Hoddes et al. (1973) recommend that the rating should be completed repeatedly (they used 15-minute intervals) to obtain the progression of fatigue over a specific time interval.

Both scales—the KSS and SSS— were validated against physiological or performance fatigue indicators (Åkerstedt & Gillberg, 1990; Hoddes et al., 1973; Otmani, Pebayle, Roge, & Muzet, 2005) and have been frequently used in the context of manual driving (Phillips, 2014) and CAD (Gonçalves, Happee, & Bengler, 2016; Jarosch, Kuhnt, Paradies, & Bengler, 2017; J. Schmidt, Braunagel, Stolzmann, & Karrer-Gauss, 2016; J. Schmidt, Stolzmann, & Karrer-Gauss, 2016).

Table 4. Comparison of the two versions of the Karolinska Sleepiness Scale (KSS).

Level	**Original KSS (Åkerstedt & Gillberg, 1990)**	**Adapted KSS (Baulk et al., 2001)**
1	Extremely alert	Extremely alert
2	–	Very alert
3	Alert	Alert
4	–	Rather alert
5	Neither alert nor sleepy	Neither alert nor sleepy
6	–	Some signs of sleepiness
7	Sleepy—but no difficulty remaining awake	Sleepy, no effort to stay awake
8	–	Sleepy, some effort to stay awake
9	Extremely sleepy—fighting sleep	Very sleepy, great effort to keep awake; fighting sleep

There are many studies that provide evidence for the sensitivity of subjective fatigue ratings when validating them with other measurement methods, such as performance or physiological metrics (e.g., Horne & Baulk, 2004; Ingre, Akerstedt, Peters, Anund, & Kecklund, 2006; Kaida, Akerstedt, Kecklund, Nilsson, & Axelsson, 2007; Otmani et al., 2005). However, it has often been criticized that the ability of a person to self-assess their fatigue level is limited (Platho et al., 2013) or may even be impaired by fatigue itself (Brown, 1994). Some authors argue that the same state of fatigue may be experienced or perceived differently due to personality and temperament (Lal & Craig, 2001). For those reasons, some researchers argue for a mere objective assessment of fatigue (Gimeno et al., 2006). From a study planning perspective, it has to be considered that asking participants to fill in questionnaires—whether only once or repeatedly—can have an activating effect on the test persons (E. A. Schmidt, Schrauf, Simon, Buchner, & Kincses, 2011). This might be undesirable when the study aims at provoking a higher fatigue level for testing reasons. Furthermore, questionnaires cannot provide real-time or similar to real-time information on the current fatigue status due to the limited asking intervals (Lal & Craig, 2002).

4.8.2 Measures of Physiology

A great strength of physiological measures is that they cannot be influenced deliberately or only to a limited extent. There were many approaches to assessing fatigue by various physiological factors. The most relevant ones are introduced in the following. It has to be noted that there are further physiological methods that have been related to fatigue, for instance, the assessment of voice characteristics (Krajewski, Batliner, & Golz, 2009; Krajewski, Trutschel, Golz, Sommer, & Edwards, 2017), skin potential level (Bittner et al., 2000) or muscle tone (Karrer-Gauß, 2011). However, these methods have been less researched so far, or the validity is questionable. That is why they will not be further considered in the following.

4.8.2.1 Eyelid Closure Behavior

Methods for assessing fatigue based on eye activity are probably the most popular ones. Basically, eye activity methods can be classified into two types: the assessment of the eyelid closure behavior and the assessment of gaze behavior (Platho et al., 2013). Due to the low validity of gaze behavior metrics for the assessment of fatigue, e.g., number and duration of saccades and fixations (Platho et al., 2013), only the metrics determined by eyelid closure behavior are presented in the following.

Metrics based on eyelid closure behavior are often referred to as the most valid ones for assessing fatigue (Platho et al., 2013). Eyelid closure behavior can be registered either by an electrooculography (EOG), where electrodes attached close to the eyes measure the change in the electric potential when blinking, or by head-mounted or contactless camera-based systems. While the contactless assessment is beneficial in terms of not biasing the participants, the sampling frequency of the EOG is many times higher than the one of commercial eye-tracking systems and can, thus, provide a higher resolution (Feierle, 2017).

One metric of eyelid closure behavior is the **blink frequency/rate**, which is the number of blinks within a determined time interval (e.g., 60 seconds). It was found in several studies that blink frequency increases with upcoming or moderate fatigue (Galley, Schleicher, & Galley, 2003; Gimeno et al., 2006; Papadelis et al., 2009; Schleicher, Galley, Briest, & Galley, 2008) and decreases again with severe fatigue (Platho et al., 2013). However, the blink frequency also correlates with other factors, for instance, with workload or time-on-task, which may occur with fatigue, but does not have to (Stern, Boyer, & Schroeder, 1994). Therefore, this metric alone is not a distinct indicator for fatigue. A further indicator for high fatigue levels is the **blink duration** (Gimeno et al., 2006). Studies proved a strong correlation between an increase of the blink duration and both subjective fatigue (Galley et al., 2003; Ingre et al., 2006) and fatigue assessed by observer ratings (Galley et al., 2003). A further significant metric for fatigue based on eyelid closure is **microsleep**. Literature does not agree on the exact value of blink duration from which on it has to be considered as microsleep. Durations from 500 ms (Friedrichs & Yang, 2010; Schleicher et al., 2008) to 1000 ms (Ebrahim, 2016; Friedrichs & Yang, 2010) are discussed. There are publications that speak for durations longer than three seconds based on results from brain activity measurements (Wilkinson et al., 2013). The indicator **PERCLOS** (PERcentage eyelid CLOSure), which is closely related to blink frequency, blink duration and microsleep, is probably the most popular and most promising one for predicting higher levels of fatigue. PERCLOS indicates the percentage of time within a predefined time interval, during which the eyes are more than 80 % closed, referring to a previously measured reference value of an individual's normally opened eyes (Wierwille, Wreggit, Kirn, Ellsworth, & Fairbanks, 1994). Greater PERCLOS values indicate increased fatigue. It is often argued that PERCLOS represents a measure for relatively high fatigue levels. However, many studies attempted to define thresholds for distinct levels of fatigue, also lower ones (see Table 5 for an overview). It is notable that values for attributing drivers to a higher fatigue level range from 1.2 % to 45 %. Measurement intervals vary between 20 seconds (Rosario, Solaz, Rodriguez, & Bergasa, 2010), 30 seconds (Bergasa, Nuevo, Sotelo, Barea, & Lopez, 2008, p. 36; J. Schmidt, Braunagel, et al., 2016), 60 seconds (Boverie, Rodriguez, & Bande, D. Saccagno, A., 2013; Wierwille et al., 1994), which was the assessment interval of the originally developed PERCLOS, and several minutes, e.g., 3 minutes in Hanowski, Bowman, Alden, Wierwille, and Carroll (2008).

Table 5. Fatigue levels and corresponding PERCLOS values based on literature; adapted from Feldhütter et al. (2019).

PERCLOS value range [%]	**Fatigue level**	**Source**
< 7.5	Awake	Wierwille et al. (1994)
7.5–15	Questionable	
> 15	**Drowsy**	
< 12.5	Low Drowsiness	Hanowski et al. (2008)
12.5–25	Moderate drowsiness	
> 25	**Severe Drowsiness**	
< 24	Attentive	Rosario et al. (2010)
24–45	Fatigued	
> 45	**Drowsy**	
< 6	Awake	G. Wu et al. (2018)
6–19.8	Mild fatigue	
>19.8	**Drowsy**	
=100	Sleep	
> 1.2	**Drowsy**	Tijerina et al. (1999)
> 40	**Fatigued**	Lin et al. (2015)
> 30	**Drowsy**	Selvakumar, Jerome, Rajamani, and Shankar (2016)

4.8.2.2 Pupillography

Further eye-based methods are the assessment of the spontaneous variation of the **pupil diameter** (pupillometry) and of the **pupil light reflex** (pupillography) (Cluydts, Valck, Verstraeten, & Theys, 2002; Karrer-Gauß, 2011). There is a standardized technique for assessing daytime sleepiness via pupillography (pupillographic sleepiness test, PST), which was proven to have a high validity (Platho et al., 2013). However, the partially strong interaction of the pupil diameter and light reflex reaction and other factors, such as lighting conditions, workload or boredom, demands strict test settings, in which all these factors are eliminated, to achieve a high validity of the test (Karrer-Gauß, 2011; Platho et al., 2013). Therefore, the application in driving experiments is not suitable.

4.8.2.3 Electroencephalography

As mentioned in chapter 4.3, there is a strong correlation between fatigue and arousal of the central nervous system, mainly concerning the brain activity level. This variation of brain activity can be measured by means of electrodes, which are attached to the scalp, and is amplified and evaluated using an electroencephalogram (EEG). Brainwaves related to fatigue can be classified in frequency bands. **Alpha waves** (8–13 Hz) are mainly present in relation to relaxing (Lal & Craig, 2001); however, they may also be an early indicator for fatigue (C. C. Liu, Hosking, & Lenné, 2009; as cited in Platho et al., 2013). **Beta waves** (13–30 Hz) are correlated with increased alertness, arousal and excitement. With the occurrence of fatigue, beta waves are reduced (Lal & Craig, 2001). Lower frequencies, such as **theta waves** (4–7 Hz) and **delta waves** (0.5–4 Hz), emerge more frequently due to severe fatigue (Lal & Craig, 2001; C. C. Liu et al., 2009; as cited in Platho et al., 2013). Apart from the sole consideration of frequency band activities, what is referred to as the **alpha spindle rate** is another indicator for vigilance decrements and fatigue (Papadelis et al., 2009; E. A. Schmidt et al., 2011). Spindles are temporal (0.5–3 seconds) bursts in the alpha band and can be determined by peak frequency, amplitude and duration. The rate is the incidents of spindles within a certain time interval (Platho et al., 2013). The alpha spindle rate is much more robust against confounding effects (E. A. Schmidt et al., 2011), which are usually a significant issue when using EEG.

Electroencephalography (also abbreviated as EEG) is often declared to be the gold standard method for assessing fatigue in a valid way. However, the method is very limited because it is highly invasive and resource intensive. That is why the application beyond laboratory conditions, e.g., in a vehicle, is limited. Furthermore, the method is very sensitive to errors and tends to produce artifacts. Thus, users/researchers need a high degree of expertise in terms of applying the method, processing the data and interpreting the results.

4.8.2.4 Heart Rate

Reduced activation, which goes along with fatigue, affects not only brain activity but may also have an impact on heart activity, which can be assessed by means of an electrocardiography. The electrical activity of the heart is measured using electrodes placed on the chest skin and recorded, resulting in an electrocardiogram (ECG). Two metrics are of interest in terms of assessing fatigue: **heart rate**—the number of heart beats per a defined time interval (most frequently per minutes)—and **heart rate variability** (HRV)—the variation in the time interval between consecutive heartbeats. Heart rate was found to be significantly decreased during fatigue (Lal & Craig, 2001, 2002; E. A. Schmidt et al., 2011; Wierwille et al., 1994), and different components of HRV proved to be sensitive in predicting fatigue (Kaida et al., 2007). However, both metrics appear not to be distinct, since they have also been correlated with a variation in other factors, e.g., in workload (Lal & Craig, 2001; Platho et al., 2013).

4.8.2.5 Motor Activity: Head and Body Movements

According to Friedrichs and Yang (2010), an often observed indicator for fatigue is a variation in head movements. This may have two different sources: the first one is attributed to self-activating or fatigue compensatory behavior (also called mannerism), when drivers move their body in the seat and move their head to fight sleep. The second reason is the typical **head nodding,** that comes along with microsleep events. Popieul, Simon, and Loslever (2002) found an increase of the **variability of the head position** along with an increase of fatigue. Furthermore, J. Schmidt, Braunagel, et al. (2016) stated that fatigued or sleeping drivers do not perform large or intended head movements, which are related to normal gaze shifts. J. Schmidt, Braunagel, et al. (2016) defined a threshold for driver's minimum **head movement velocity** at 25 °/s to distinguish intended and unintended head movements. Apart from this, **leaning the head back against the headrest** has also been found to be an indicator for severe fatigue (Tsuchida, Bhuiyan, & Oguri, 2009; Wierwille & Ellsworth, 1994).

While many authors report an increase of different aspects of body movement as part of the mannerism during the onset of fatigue (see chapter 4.8.5), a different result was found by Zilberg, Xu, Burton, Karrar, and Lal (2009). They showed a significant reduction of seat movement magnitude objectively measured by piezo sensors integrated in the driver's seat. It is assumed that this discrepancy may result from the evaluation of different fatigue levels in the studies.

4.8.3 Measures of Driving Performance

Many studies were able to show that fatigue leads to a deterioration of manual driving performance, which may lead to minor and major incidents (Reyner & Horne, 1998). Driving performance is assessed by means of various metrics of lateral or longitudinal guidance. The lateral guidance, and here in particular the steering behavior and the lane keeping quality, was examined comprehensively in the past. Highly valid metrics are, for instance, the **standard deviation of steering wheel angle** (STS) (Eskandarian et al., 2007; Thiffault & Bergeron, 2003b), the s**tandard deviation of lateral position** (SDLP) (Anund, Fors, Hallvig, Åkerstedt, & Kecklund, 2013) or **touching lane markings/partially leaving the lane** (Otmani et al., 2005). All three metrics increase significantly for fatigued drivers. STS and SDLP are not independent of each other, so using only one of them is recommended (Eskandarian et al., 2007) when applying them in studies. In terms of the longitudinal guidance, an increase of the speed variability (Wierwille et al., 1994) or of the average speed (Platho et al., 2013) and the inability to maintain a constant speed (Lal & Craig, 2001) can be an indicator for fatigue.

It is criticized that driving performance metrics are not unique, because many of them also indicate inattention or alcohol ingestion (Karrer-Gauß, 2011; Platho et al., 2013). The main issue of manual driving performance assessment is, however, that it cannot be applied in automated driving since vehicle guidance is controlled by the automated function.

4.8.4 Existing Systems for Assessing Driver Fatigue

Research on in-vehicle driver fatigue assessment systems has had a long history because the avoidance of fatigue in road traffic promises to reduce accidents and to increase road safety. There are many challenges and demanding requirements when integrating systems to detect driver fatigue in vehicles. To achieve user acceptance and user value the system has to detect fatigue accurately (Bowman et al., 2012). This means that the sensitivity (true positive rate) and specificity (true negative rate) has to be nearly 100 %. Furthermore, the system should not negatively affect the driver's comfort (Karrer-Gauß, 2011). Therefore, the system has to be unobtrusive and non-intrusive (Larue et al., 2010). Moreover, the fatigue assessment system has to be real-time capable (Bowman et al., 2012) to allow fatigue evaluation during the ongoing drive and to deliver its benefits for road safety.

For manual driving, several in-vehicle fatigue detection systems have been on the market for many years. They have in common that they mainly use measures of driving performance to predict the fatigue state. For instance, the *Driver Alert System* developed by the Volkswagen AG continuously evaluates the driver's steering behavior by analyzing the steering wheel angle and compares it with a driver's individual baseline taken at the beginning of each trip. In case of a deviation, an optical and acoustic alarm is prompted to recommend taking a break to the driver (Volkswagen AG). The *Driver Alert Control* by the Volvo Group analyses the lateral position of the car within the lane via a camera system, which assesses the lane marking. In case of an unsafe steering behavior, e.g., swerving, an alert is initiated containing a coffee cup symbol, an acoustic warning and a text message in the instrument cluster, which recommends taking a break (Volvo Group, 2020). Other OEMs and suppliers, such as Daimler AG (*Attention Assist*), BMW Group (*Attention Assistant*), Ford Motor Company (*Driver Alert System*) or Bosch Group (*Driver Drowsiness Detection*), also provide systems that function in a very similar way based on steering behavior or other driving-related inputs. With this, carmakers and suppliers followed a recommendation of the European New Car Assessment Programme (Euro NCAP), which is a coalition of European departments of transport, insurance associations and automobile associations with a consumer protection-oriented program to evaluate the safety of cars. In its roadmap for a further improvement of road safety for the period between 2020–2025, it targeted the implementation of driver monitoring in each car by 2022 to mitigate risks resulting from driver distraction, fatigue and alcohol consumption (Euro NCAP, 2017).

As already mentioned, metrics of driving performance cannot be utilized in automated driving, since the lateral and longitudinal vehicle control is performed by the automated system. Instead, physiological measures (see chapter 4.8.2) gain in importance to predict fatigue. OEMs concentrate on eyelid closure behavior because these measures appear to be the most valid ones (Platho et al., 2013, cf. chapter 4.8.2.1), and the assessment by a built-in camera system meets the above-mentioned requirements of unobtrusiveness and non-intrusiveness. In 2006, Toyota Motor Company was the first automotive manufacturer, which integrated a driver monitoring system in the vehicle including a camera (*Driver Attention Monitor)* to advance their pre-crash safety system. The system initiates an early warning in case of an imminent collision and a visually inattentive driver (not looking straight ahead). With respect to automated driving, Bosch Group introduced an interior monitoring system, which consists of a driver and an interior monitoring camera and aims at detecting driver distraction and drowsiness. There are also many approaches in research to detect fatigue by onboard cameras in real-time. Some use only one metric, preferably PERCLOS (Garcia et al., 2010), or fuse various metrics, such as PERCLOS, microsleep or head movements (Bowman et al., 2012; Cabrall et al., 2016; J. Schmidt, Braunagel, et al., 2016). Some of the approaches are based on machine vision and machine learning algorithms. No approach was found in literature, however, which achieves a reliability of approximately 100 %. According to Bowman et al. (2012), a multiple sensor approach achieves significantly better results and should therefore be preferred over single sensor approaches. However, the reliability of the published fatigue detection systems is not yet sufficient for series employment or, at least, it is too low to be part of a safety concept.

4.8.5 Behavioral Observation

With the onset of fatigue, many observable behavioral changes related to motor activity emerge. Behavioral indicators include, for example, altered eye activity, head and body movement dynamics, facial expression and mannerisms.

Based on the observation of the drivers' faces and upper torsos, the behavioral fatigue indicators are assessed, and the drivers are assigned to a certain level of fatigue with the aid of rating scales. The most popular rating scale is the five-level Wierwille scale of Wierwille and Ellsworth (1994), terming the method **Observer Rating of Drowsiness**. Drivers are attributed to one of the five main levels “not drowsy”, “slightly drowsy”, “moderately drowsy”, “very drowsy” or “extremely drowsy”, whereby also unlabeled intermediate stages between the single main levels are possible. Even though the scale has five main levels, Wierwille and Ellsworth (1994) do not differentiate between “slightly drowsy” and “moderately drowsy” when describing the indicators for these two levels but summarize them under the level of “drowsy”. All indicators corresponding to a certain level of drowsiness according to Wierwille and Ellsworth (1994) are listed in Table 6. Beside the mannerisms that are enumerated in Table 6, there are also other self-activating behaviors, which are related to fatigue. These include **hard and forced eye blinks** (Eskandarian et al., 2007), **slouching** and subsequent **postural adjustments** by lifting the upper body in an upright posture (Senaratne, Hardy, Vanderaa, & Halgamuge, 2007) and **stretching** (Anund et al., 2013).

Derived from the Wierwille scale, Wiegand, McClafferty, McDonald, and Hanowski (2009) advanced the observer rating by developing a comprehensive training protocol for raters. This training includes—apart from the solely written descriptions of fatigue levels—videos of drivers in a naturalistic driving environment showing examples of the different fatigue levels. The evaluation of the training protocol revealed satisfactory intra-rater reliability (r=0.88), inter-rater reliability (r=0.81) and indicators of validity (Wiegand et al., 2009). There are further publications that used rating scales similar to or adapted from the abovementioned ones (Anund et al., 2013; Galley et al., 2003; Hirose, Kitabayashi, & Kubota, 2015; Karrer-Gauß, 2011; Weinbeer et al., 2018; Wylie et al., 1996).

Table 6. Scale and corresponding indicators according to Wierwille and Ellsworth (1994).

Level of drowsiness[a]	General description	Exemplary indicator
Not drowsy	Behaviors typical when alertness is present	- normal facial tone - normal fast eye blinks - short ordinary glances - occasional body movements or gestures
Drowsy (including slightly and moderately drowsy)	Mannerisms or self-activating behaviors as countermeasures to drowsiness (fatigue) *or* More subdued behaviors	- rubbing/touching the face or eyes - scratching - facial contortions - moving restlessly in the seat - yawning *or* - slower eyelid closures - decreased facial tone - glassy-eyed appearance - starring at a fixed position
Very drowsy	None	- prolonged eyelid closures (2 to 3 seconds or longer) - rolling upward or a sideways movement of the eyes - eyes do not properly focus - cross-eyed (lack of proper vergence) look; further decreased facial tone - lack of apparent activity - large isolated (or punctuating) movements (such as a large correction to steering or reorienting the head from a leaning or tilting position)
Extremely drowsy	None	- falling asleep - prolonged eyelid closures (4 seconds or more) - prolonged periods of lack of activity - large punctuated movements due to transition in and out of intervals of dozing

[a] The term drowsiness is used in this table for compliance reasons with the original source. To be consistent within this thesis, the term *level of fatigue* is used, hereinafter.

The observer rating method via behavioral indicators is classified as a highly sensitive method (Platho et al., 2013). A reason for the sensitivity could be that the method not only relies on one single assessment type (e.g., on eye activity) but also adds many other indicators that were correlated to fatigue. Thus, some studies found this method to be significantly more sensitive than physiological measures, for instance, than EEG (Wylie et al., 1996). Furthermore, it has a high face validity (Platho et al., 2013). It has to be pointed out, though, that the reliability and validity of the method strongly depends on the expertise of the rater. Therefore, a previous, extensive training is required like the one suggested by Wiegand and colleagues, to ensure the quality of the fatigue evaluation. That is why this method is often called expert assessment. Furthermore, it is recommended that the evaluation be conducted by at least two independent raters. (Platho et al., 2013)

Precisely because of the loss of driving behavior data as an indicator for fatigue assessment in CAD, physiological and behavioral fatigue assessments gain in importance. Of course, all abovementioned indicators are also suitable for usage in automatic fatigue assessment systems. There have been some approaches to fuse some of them in one system (e.g., Senaratne et al., 2007). However, the technical implementation is challenging, and to date no validated and reliable system has been found in literature that fuses multiple indicators of fatigue and meets all challenging requirements for a series implementation.

4.9 Driver Fatigue in Conditionally Automated Driving

Research on driver fatigue in automated driving initially started with the closer examination of passive and active TR fatigue to test the theory of Desmond and Hancock (2001) which differentiates between the two fatigue types introduced in chapter 4.4. In a series of driving simulator experiments, Saxby and colleagues could prove that passive TR fatigue—assessed by subjective measures mainly consisting of the DSSQ (see chapter 4.8.1)—is induced by prolonged durations of automated driving (Saxby et al., 2007; Saxby et al., 2008; Saxby et al., 2013). Different time periods (ten, 30 or 50 minutes) of automated driving were examined to induce passive TR fatigue. Results showed that even ten minutes of automated driving were enough to induce passive TR fatigue (Saxby et al., 2007). A subsequent manual drive including an emergency event requiring an evasive maneuver showed that prolonged automated driving phases aiming at inducing passive TR fatigue led to impaired response times and an increased crash probability (Saxby et al., 2008; Saxby et al., 2013). In two follow-up studies, Neubauer et al. (2012, 2014) focused on the effect of secondary tasks on the development of passive TR fatigue (like in the studies of Saxby and colleagues subjectively assessed by the higher-order factor task engagement of the DSSQ) induced by prolonged phases of automated driving. In the first study, participants were assigned to three different secondary task conditions: all of them were asked to use a cellphone during an automated drive of 25 minutes either by texting or by phoning or by having the free choice between these two options. Like in the preceding studies of Saxby et al. (2008) and Saxby et al. (2013), participants were confronted with an emergency situation during a manual drive following the automated one. Results of the DSSQ yielded that none of the cellphone conditions could prevent the development of subjective fatigue during automated driving. However, participants in all three cellphone conditions had significantly faster response times to the emergency event compared to the control condition without a secondary task, indicating that cellphone usage maintains alertness during automated driving (Neubauer et al., 2012). In the second study (Neubauer et al., 2014), secondary tasks included a hands-free cellphone conversation and a quiz game of "trivia" and were performed for two 10-minute phases during 40 minutes of automated driving, followed by the manual drive with an emergency situation. Results showed that the media usage led to less passive TR fatigue when driving in automated mode, compared to the control group with manual driving. However, braking response times during the emergency event were significantly slower after driving in automated mode regardless of the secondary task, which contradicts the findings of the first study (Neubauer et al., 2012). Neubauer et al. (2014) argue that the cellphone usage in the second study was less frequent and for a longer period, and therefore, the alerting effect was less strong.

Another early study on automated driving (Omae, Fujioka, Hashimoto, & Shimizu, 2006) examined the effect of the automation duration on the response times when the system malfunctions. In this study, a system similar to a partial automation was implemented, meaning that the failures occurred without a RtI or warning. Results showed that both the mean automatic reaction times (first move of hands and feet) and the mean intervention time (conscious steering or braking) after automation durations of five, ten and 30 minutes did not differ from each other, even though the individual variation increased with automation duration. Only after 60 minutes of driving with partial driving automation, did both mean response times increase (doubled compared to a duration of 5–30 minutes). Besides, more than 25 % of the participants fell asleep even though they were instructed about the possible malfunctions and about their duty to monitor the system.

Subsequent and more recent studies focused on the effect of a prolonged duration of CAD on take-over performance. Feldhütter, Gold, Schneider, and Bengler (2017) compared the effect of quite a short duration of CAD (five minutes) prior to a RtI to a prolonged one of 20 minutes in a driving simulator study. Additionally, the impact of a standardized NDRA—the Surrogate Reference Task (SuRT) (International Organization for Standardization, 2012)—was evaluated. Results showed an increase in the gaze reaction time (the period from the RtI to the point in time when participants averted their eyes off the NDRA back to the road) after 20 minutes of CAD compared to five minutes. However, it remained unclear whether this finding can be attributed to the prolonged automation duration or to the prolonged engagement in the rather monotonous SuRT. The remaining take-over performance parameters (TOT, TTC, accelerations) did not differ between the short and prolonged automation duration. Fatigue was not assessed by specific measures in this study. In a more recent simulator study, Bourrelly et al. (2019) compared take-over performance after ten and 60 minutes of CAD. They found that the response times to bring the hands back on the wheel, the feet back to the pedals and to initiate a manual driving maneuver were each significantly longer after 60 minutes than after ten minutes of CAD. The authors argue that the prolonged response times may result from a deteriorated cognitive functional capability induced by fatigue. However, it cannot be ensured that the effect can be attributed to fatigue in this study, because fatigue was not objectively assessed, but only by means of a retrospective self-appraisal of the fatigue state after completing the take-over event.

Instead of triggering the RtI after a specific fixed duration of automated driving regardless of the prevailing fatigue state, there are a few studies that take a different approach: initiating the RtI depending on a specific predefined fatigue state. In a driving simulator study, Hirose et al. (2015) evaluated participants' alertness level by observing behavioral indicators, analyzing EEG and heart rate. A take-over situation with a RtI was issued once participants reached a state of low alertness. Braking response times were compared between participants with a low and normal alertness level. Results indicate that low alertness resulted in significantly slower response times. In a different driving simulator study, Schömig, Hargutt, Neukum, Petermann-Stock, and Othersen (2015) continuously assessed and classified drivers' fatigue level by eyelid closure measurements during manual driving. Once a certain fatigue level was reached through manual driving, the subsequent 15-minute test phase comprised either driving with CA without any NDRA, driving with CA and engaging in a quiz game or driving further on manual, depending on the experimental condition. Regardless of the condition, the phase ended with a take-over situation with a time budget of 12 seconds. Due to the study design, results could only be analyzed descriptively. The analysis yielded that the fatigue level in the manual driving and CAD condition increased remarkably, whereas in the condition with the quiz game the fatigue level remained on a low level. All drivers—also the ones at a high fatigue level—resolved the take-over situation successfully (Schömig et al., 2015). In a driving simulator experiment of Gonçalves et al. (2016), participants drove in a conditionally automated way on a monotonous freeway without engaging in a NDRA in order to induce passive TR fatigue. The participants were asked to repeatedly rate their self-perceived fatigue on the SSS. Once participants reached level 5 on the SSS, a take-over situation was triggered, which was the case after less than 15 minutes of CAD on average. When analyzing the take-over performance, results showed that only lateral acceleration was significantly higher in the fatigued group compared to a non-fatigued reference group. TOT, TTC and longitudinal acceleration as well as the number of mirror checks did not differ between the two groups. In a real traffic study of Weinbeer et al. (2018) on a German Autobahn, CAD was simulated by an adapted right-hand-drive vehicle (referred to as the *wizard-of-oz* method). Passive TR fatigue was induced by the prolonged conditionally automated drive and relaxing background music. Fatigue was assessed through behavioral observation by two experts assigning participants to a six-level scale. The take-over event was simulated using a response-time task that was specifically developed and which could only display timing aspects of the take-over. Participants had to input a steering action after the RtI sound was triggered. A RtI was triggered several times for each participant depending on different fatigue levels. Hands-on time and the intervention time similar to the TOT were analyzed. Results showed no significant deterioration of both response times due to higher levels of fatigue compared to response times taken from the same participants under a non-fatigued condition. Weinbeer et al. (2018) argue that, despite a previous training of the response-time task, a training effect due to the repeated performance of the task over the multiple fatigue levels cannot be excluded. Vogelpohl, Kühn, Hummel, and Vollrath (2018) examined the development and effect of SR and passive TR fatigue. For this purpose, one group of participants with a lack of sleep (less than five hours) were tested during a circadian low phase (SR fatigue condition). Fatigue was assessed during CAD by observer rating using the method developed by Wiegand et al. (2009). The take-over situation was triggered once participants reached a predefined level of fatigue. Passive TR fatigue was induced by an extended monotonous drive of a fixed 60 minutes of CAD ending with the same take-over situation. For both fatigue conditions, there was a reference group driving manually. Results showed that CAD led to a faster and stronger evolvement of fatigue compared to manual driving. A lack of sleep accelerated this effect. A significant deterioration of take-over performance compared to manual driving was not found even though there were 30 % of participants with a lack of sleep that showed extended reaction times. Vogelpohl et al. (2018) argue that an intervention maneuver after the transition from CAD to manual driving is difficult to compare to an intervention maneuver after manual driving only, since additional processes have to be carried out to regain vehicle control and built up situation awareness.

A study combining a fixed duration with a fatigue-dependent design was conducted by J. Schmidt, Dreißig, Stolzmann, and Rötting (2017). Participants experienced four different take-over situations after a fixed duration of between eleven and 108 minutes of CAD. A fifth take-over situation was triggered once participants missed to respond to a specific alertness request. Hands-on-steering-wheel time was compared between the situations and between fatigue levels that were repeatedly assessed by the KSS. Results revealed no deterioration of the reaction time due to fatigue. Instead, an effect of situational factors was found. Furthermore, J. Schmidt et al. (2017) suppose a stronger dependency between individual ability and reaction time than between fatigue and reaction time. In a different analysis, J. Schmidt, Braunagel, et al. (2016) evaluated different behavioral indicators for fatigue, which were PERCLOS, microsleep and head movements during CAD and manual driving. They found that the indicators differed between the CAD and the manual driving phase, and concluded that findings for these metrics from manual driving cannot be transferred to CAD easily.

Since in CAD NDRAs gain in importance, some studies focused on the effect of different types of NDRAs on driver fatigue and take-over performance. For instance, Jarosch et al. (2017) found in a driving simulator study that a monotonous monitoring/vigilance task (monitoring a display and giving input once a certain sign occurs, called Pqpd task) may fatigue drivers during CAD, whereas a quiz game countered fatigue significantly. The tasks were performed for 25 minutes each and presented in a within-subject design. Fatigue was assessed by the KSS subjectively and by PERCLOS objectively. The take-over performance did not differ between the two task conditions. In a second study, Jarosch, Bellem, and Bengler (2019) applied identical tasks (Pqpd and quiz), however, with a CA duration of 50 minutes and applied in a between-subject design. Again, a significantly higher fatigue level according to the PERCLOS value was found for the monitoring task, whereas this effect was not reflected in the subjective fatigue evaluation (KSS). Regarding take-over performance, significantly slower reaction times for the rather automatic actions (gaze back to the road and foot back to the braking pedal) were found for the Pqpd condition. Jarosch, Bellem, and Bengler (2019) argue that the different result compared to their first study (Jarosch et al., 2017) may be attributed to the extended task exposure duration. A direct comparison of the data of both studies revealed that the worst take-over performance emerged for the 50-minute exposure to the fatiguing monitoring task (Jarosch & Bengler, 2019a). In a consequent real-traffic driving study with a wizard-of-oz vehicle, in which CAD was simulated, Jarosch, Paradies, Feiner, and Bengler (2019) confirmed the findings of their previous driving simulator studies. Participants with the Pqpd task had a significantly higher fatigue level compared to participants with a free choice of NDRAs assessed by PERCLOS and KSS after 60 minutes of CAD. In another real-traffic driving study with a wizard-of-oz vehicle, Weinbeer, Muhr, and Bengler (2019) tested three activity conditions in terms of their reactivating effect following a fatigue generation phase: dictation, light sport activity for the arms and hands (Handytrim fitness device) and no NDRA (control group). Results indicate that both activities had a descriptively reactivating effect on the participants compared to no activity (subjectively assessed by the KSS), regardless of the type of the activity.

Weinbeer et al. (2019) concluded that NDRAs may have a reactivating potential, but more naturalistic activities have to be tested in terms of their effectiveness.

Discussion of previous related research

Only a few studies were found in literature that have researched fatigue in CAD so far. The abovementioned studies were able to verify a slight to strong increase of passive TR fatigue during automated driving, especially if no NDRA is present (e.g., Saxby et al., 2007; Vogelpohl et al., 2018; Weinbeer, Bill, Baur, & Bengler, 2018). Results regarding the impact of fatigue on the take-over performance, however, are contractionary. Hirose et al. (2015) found prolonged response times for participants in a low alertness level which is also expected from literature review on the impairments caused by fatigue when driving manually (cf. chapter 4.5). Bourrelly et al. (2019) found prolonged response times due to prolonged automation duration making increased fatigue responsible for this finding. Gonçalves et al. (2016) found no slower take-over times but higher lateral accelerations for fatigued participants. On the other hand, there are studies that cannot confirm these findings and expectations. For instance, J. Schmidt et al. (2017), Weinbeer et al. (2018) and Vogelpohl et al. (2018) did find neither slower take-over times nor impairments of other take-over performance parameters. The discrepancies among these studies' findings might be attributable to widely divergent methods for assessing fatigue and significantly different study designs. Some studies used a fixed automation duration varying between 10 to 60 minutes of automated driving. Indeed, time-on-task may be an effective tool to induce passive TR fatigue (see chapter 4.4 and Phillips, 2014), however, the duration that is required to induce passive TR fatigue is not known and individual differences are not taken into account (see chapter 4.6). Other studies used a fatigue-state dependent design and triggered the RtI depending on a certain fatigue level. Regardless of the general study design in terms of a fixed or variable automation duration, the methods to assess fatigue during the experiment varied widely from self-rating methods, such as KSS or DSSQ, to objective methods such as eye-tracking metrics and observer rating. Many studies relied only on one method to assess fatigue (subjective or objective) and did not consider the multidimensionality of fatigue (see chapter 4.2 and Phillips, 2014).

The engagement in a NDRA could frequently prevent the development of fatigue and act as a countermeasure (e.g., Jarosch et al., 2017; Neubauer et al., 2014; Schömig et al., 2015; Weinbeer et al., 2019), which is in accordance with past research and theories of manual driving (see chapter 4.6). However, this effect could not always be achieved (Neubauer et al., 2012). Interestingly, despite an activating effect, NDRAs did not necessarily lead to an improvement of take-over performance or of performance during emergency events. While some authors found a variation of the performance according to the targeted manipulation by the NDRA (Jarosch, Bellem, and Bengler, 2019: a fatiguing task led to slower reaction times; Neubauer et al., 2012: an interesting task led to faster reaction times), other studies cannot confirm these findings (Jarosch et al., 2017; Neubauer et al., 2014). Even though the type of the NDRA differed widely for most of the studies—reaching from standardized tasks, such as the SuRT, to more natural tasks like cellphone usage and quiz games—most studies have in common that the engagement in the activities was mandatory, and the activities had to be performed uninterruptedly over prolonged periods of time. Whether an activity can prevent fatigue or activate drivers depends on many factors such as motivation, intrinsic interest, stimulation and voluntary performance (see chapter 4.6). In particular, activities with a high workload (e.g., a quiz) that are mandatory may have the opposite effect and may evoke active TR fatigue over time instead of activating drivers.

5 Research Questions

Driving is a complex task requiring various skills and abilities of humans. Fatigue may negatively affect basic cognitive processes and functions such as attention, information processing and knowledge-based responses resulting in decrements of the manual driving performance and, ultimately, also in accidents (cf. chapter 4.5).

With the introduction of CAD, the driver is relieved from the complex driving task. However, fatigue issues stay relevant since cognitive processes required during a take-over are affected by fatigue. Consequently, this thesis contributes to clarifying whether performance decrements occur in the take-over process due to fatigue. Already existing studies that have examined this subject or related questions have revealed contractionary results (see chapter 4.9) due to strongly varying methods for fatigue assessment and study designs. In this thesis, a coherent method is used to make study results comparable and to derive a comprehensive conclusion.

> RQ1: How does fatigue affect the take-over performance in conditionally automated driving?

Taking the driver out of the driver-vehicle-control loop and releasing him/her from an active role is assumed to promote the development of passive TR fatigue (see chapter 4.4). However, the individual differences in fatigue proneness are significant due to different coping strategies (see chapter 4.6). Therefore, in this thesis, the general progress of fatigue in CAD is investigated to determine the influence of automation duration, time-on-task respectively (short vs. prolonged CAD) and to derive thresholds for a maximum uninterrupted duration of CAD.

> RQ2: How does the duration of the conditionally automated drive affect fatigue?

One characteristic of CAD is that the driver is allowed to engage in NDRAs. Research results have shown that the stimulation by a voluntary interesting or motivational activity may counteract the development of fatigue (cf. chapter 4.6 and chapter 4.9). It is hypothesized that naturalistic activities (instead of artificial standardized ones which were applied frequently in previous studies) chosen at will provide sufficient stimulation and motivation. It is examined in this thesis whether naturalistic NDRAs have an activating effect. This kind of NDRA usage would also represent a more realistic use case for the future scenario of CAD than without any activity or a standardized one.

> RQ3: Do naturalistic and motivating non-driving-related activities counteract the development of fatigue?

An adequate method needs to be established for the purpose of answering the stated research questions. The method to assess fatigue is probably the major challenge since it must fulfill multiple requirements. Beside the three main quality criteria for a measurement system—validity, objectivity and reliability (Bubb, Bengler, Lange, Aringer, & Trübswetter, 2015, p. 618)—it is essential to assess fatigue in real-time to be able to trigger the take-over situation once the driver reaches a predefined level of fatigue, which is especially important to answer RQ1. Furthermore, with respect to a later application in series vehicles and to avoid biasing the participants, the fatigue assessment system should be unobtrusive and preferably non-intrusive (Larue et al., 2010). In past research, PERCLOS was often declared to be a promising metric to assess fatigue and to fulfill the above-mentioned requirements (cf. chapter 4.8.2.1). However, PERCLOS was mostly evaluated in the context of manual driving, where the focus of the driver's eyes is on the road. J. Schmidt, Braunagel, et al. (2016) suggest that findings cannot be transferred to CAD easily, even if no NDRA is present. Up to now, little is known about the behavior of PERCLOS in the context of the realistic use case of CAD when drivers are engaged in NDRAs. Moreover, head-mounted eye-tracking systems were used in the past, which do not fulfill the requirement of non-intrusiveness and unobtrusiveness. Therefore, this metric is examined in more detail in the context of CAD-relevant use cases and under most realistic conditions using a commercially available remote eye-tracking system.

RQ4: How valid is the PERCLOS metric for fatigue assessment with respect to CAD?

6 General Method

Parts of this chapter have already been pre-published in Feldhütter et al. (2019), Feldhütter et al. (2018), Feldhütter et al. (2018), and Feldhütter et al. (2019), from which some parts of the written text were adopted literally.

6.1 Overview of the Experiments

To answer the research questions defined in chapter 5, four main experiments were conducted. An overview of the experiments is given in Table 7. For each experiment, the research questions considered (RQ1, RQ2, RQ3 or RQ4) were refined by more detailed research questions (see the first chapter on each experiment). Experiment 1 and experiment 2 address RQ2 and RQ3 (see chapter 7.1 and chapter 7.2). For this purpose, a study design with a prolonged fixed duration of CAD is used to examine the development of passive TR fatigue. The exact automation duration varied between the experiments. Participants were offered different types of NDRAs to evaluate their suitability for counteracting fatigue. The situational parameters of time budget and traffic condition are kept constant in these experiments. Experiment 3 and experiment 4 are described in chapter 7.3 and chapter 7.4 and address the question how fatigue affects the take-over performance (RQ1). Over all experiments, the validity and data availability of PERCLOS is evaluated under different experimental conditions (with and without NDRAs, fixed experimental duration and variable duration) to address RQ4.

Table 7. Overview of all experiments conducted.

Experiment	Study-design	Participants	NDRA	Automation duration [min]	Time budget [s]	Traffic condition [-]
Experiment 1	Fixed duration	n=21	None	60	6	No traffic
		n=21	Free choice	60	6	No traffic
Experiment 2	Fixed duration	n=20	None	4	6	Large gap
		n=20	Tetris	35	6	Large gap
Experiment 3	State-dependent	n=25	None	5	6	Medium gap
		n=30	None	18–90	6	Medium gap
Experiment 4	State-dependent	n=25	None	5	5	Medium gap
		n=22	None	11–90	5	Medium gap

The following subchapters give an overview of those experimental components which were mainly constant over all conducted experiments. All experiments took place in a fixed-base high-fidelity driving simulator and were conducted with the identical CA including HMI, system behavior and performance design, all described in subchapter 6.2. The driving simulator was equipped with a commercial camera system for driver observation (see subchapter 6.3). All experiments were conducted in German. In all experiments, the take-over performance was assessed by the same cluster of variables (see subchapter 6.5) to compare results also between the experiments and to be able to draw a more general conclusion. For the same reason, the same take-over scenario was used for all experiments with a slight variation of traffic condition and time budget within the experiments (see subchapter 6.4) Those two parameters were chosen for variation, since the isolated effect on take-over performance has already been identified (see chapter 3.3.3) and a separation from a potential fatigue effect can be achieved.

6.2 Driving Simulator and Driving Automation

All driving experiments took place in a high-fidelity, fixed-base driving simulator of the Chair of Ergonomics. The simulator consisted of a full-vehicle BMW series 6 mock-up (E64). Six projectors created a field of view greater than 180 degrees and allowed the participants the normal use of side mirrors and rearview mirrors (see Figure 6-1). An in-vehicle audio system generated engine sounds and road noise. The simulation software SILAB developed by the WIVW GmbH was used for running the driving simulation. All driving-related data (steering wheel angle, vehicle's position, acceleration, speed, etc.) were recorded at a frequency of 60 Hz.

The CA was implemented in the driving simulation in accordance with the taxonomy of the SAE International (2018), meaning that the automation performed the entire DDT. This means that the automation took over longitudinal and lateral vehicle motion control and conducted passing maneuvers whenever necessary. Furthermore, the CA determined whether there was a DDT performance-relevant system failure and, if so, issued a timely RtI to the participant. The RtI was provided by a salient, acoustic alert in the form of a doubled beep and a notice on the instrument panel (see Figure 6-2). The HMI was designed in a rather basic way, since the effect of HMI was not the focus of this thesis. To activate/deactivate CA, participants had to press a button on the active steering wheel. In addition, CA could be deactivated by exceeding a specific threshold of manual steering (producing a certain steering torque) or manual braking (giving a certain pedal input).

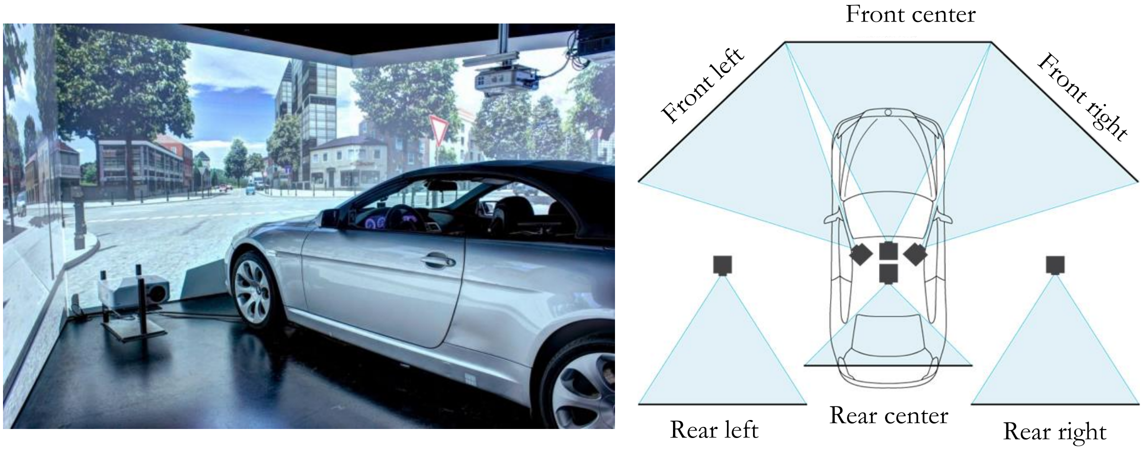


Figure 6-1. Fixed-base high-fidelity driving simulator used for all experiments.

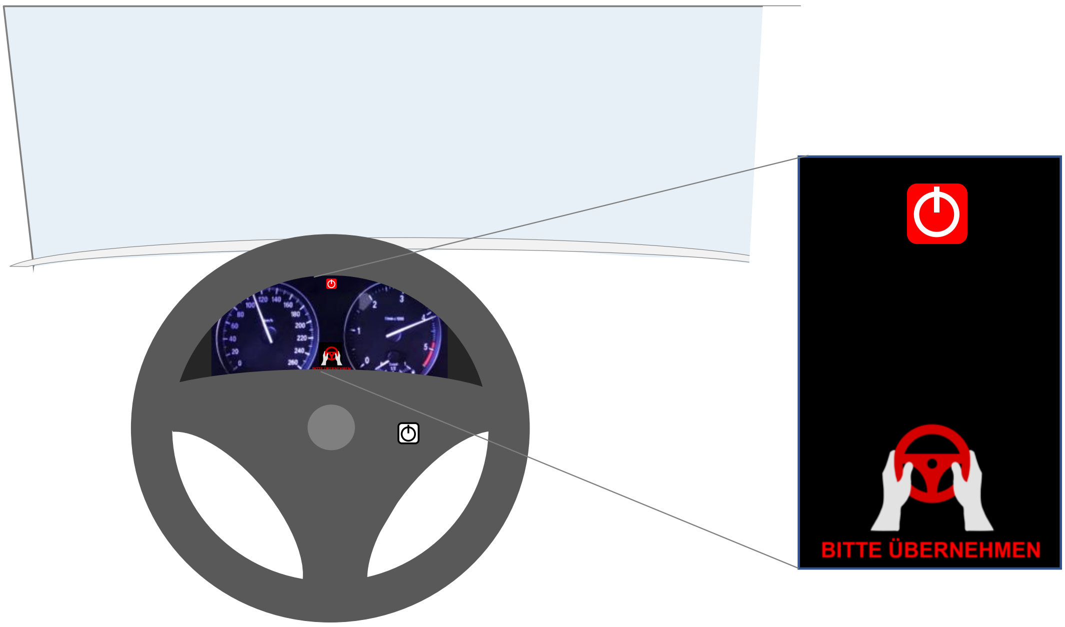


Figure 6-2. Schematic representation of the HMI of the RtI. The German text says "Please take over".

6.3 Camera-Based Driver Observation

The participants' eyes, head and upper torso were recorded at a frequency of 60 Hz by three remote infrared cameras, which were mounted on the top side of the dashboard and in the center console (see Figure 6-3). The cameras are part of the commercial eye tracking system Smart Eye Pro developed by the Smart Eye AB company. Smart Eye also provides a software that processes the camera signals in close to real-time (delays less than 100 msec) and yields numerous variables such as eyelid opening, gaze orientation or head position.

Another wide-angle camera, which was located opposite the front passenger seat, provided another view of the participants' upper body and head, and recorded the driver's behavior at a frequency of 30 Hz.

Figure 6-3. Position of the Smart Eye cameras in the interior of the driving simulator mock-up (Feierle, 2017).

6.4 Test Track and Take-Over Situation

The test track of all experiments contained a simulated drive on a German Autobahn with three lanes. Once CA was activated by the participant, the automation maintained a constant speed of 120 km/h (33.33 m/s), driving constantly on the right lane of the freeway. To promote the development of passive TR fatigue, the test track was designed to be as monotonous as possible (Thiffault & Bergeron, 2003b), meaning that the road was uniformly straight (Larue et al., 2010), the surrounding traffic unvaried and sparse, no lane-changing maneuvers and no diversifying elements such as different landscape features, buildings or tunnels (Saxby et al., 2007) were implemented.

Based on the recommendation of Gold et al. (2018) (cf. chapter 3.3.1), a take-over scenario with high urgency, low predictability, high criticality and a medium driver response complexity was selected to evaluate the controllability of the take-over and maximum driver performance. Concretely, these requirements were realized as follows in all experiments (cf. Figure 6-4): A broken-down vehicle in the ego-vehicle's lane represented an unknown obstacle and caused a RtI. Since the experiments varied regarding the NDRAs provided, participants had a different visual monitoring behavior of the driving scenario prior to the RtI. To ensure that each participant still had the same designated time budget within one experiment for taking over control and responding to the RtI, the obstacle suddenly appeared a certain distance d_3 ahead of the ego-vehicle at the same time as the RtI sounded. The distance d_3 varied between 167 meters and 200 meters among the experiments, resulting in a time budget of five or six seconds. During the RtI, the ego-vehicle was located in the right lane. Depending on the experiment, the adjacent middle lane underlay different traffic conditions, which are depicted in Table 8. The left lane was free of any other traffic in all four experiments.

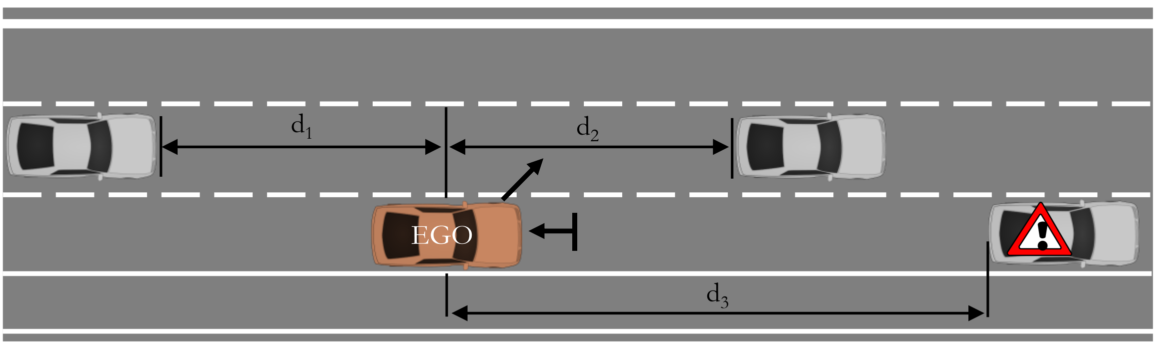


Figure 6-4. The take-over situation for all experiments; d_1, d_2 and d_3 varied among the experiments to realize different time budgets and traffic conditions.

Table 8. The different traffic conditions for the middle lane among the experiments.

Traffic condition	d_1	d_2	Experiment
No traffic	∞	∞	1
Large gap	234 m	166 m	2
Medium gap	50 m	50 m	3, 4

6.5 Assessment of Take-Over Performance

Table 9 lists the metrics used consistently to assess take-over performance in all experiments of this thesis. The evaluation period for all metrics lasts from the moment of the RtI until the ego-vehicle reaches the system limit (broken-down vehicle, see Figure 6-4).

Table 9. Dependent variables for assessing take-over performance in this thesis (timing and quality aspects).

Metric [Unit]	Abbr.	Description
Take-over time [s]	TOT	First conscious maneuver after the TOR (absolute steering wheel angle > 2°, change of braking/gas pedal position > 10 %)
Time-to-collision [s]	TTC	Minimum time theoretically remaining until a potential collision with an obstacle, assuming constant speed of the ego-vehicle
Accelerations [m/s^2]	AccLong	Maximum longitudinal accelerations occurring during the take-over
	AccLat	Maximum lateral accelerations occurring during the take-over
Initial Response [-]	InRe	Type of first reaction to the TOR (steer, brake, accelerate)
Final Response [-]	FinRe	Type of final reaction to ultimately resolve the take-over situation (full braking or evasive maneuver)
MirrorCheck [-]	MC	Whether the driver checked the side-view mirror prior to a lane change maneuver (yes, no)
Crash [-]	-	Whether a crash occurred during the take-over (yes, no)

As described in chapter 3.3.2, the TOT is the most important timing aspect of take-over performance, because it is the time participants need for the first conscious intervention maneuver as a reaction to the RtI. Thresholds for determining a conscious action were adopted from Gold (2016), who evaluated hundreds of take-over situations: Exceeding an absolute steering wheel angle of two degrees for steering or changing the braking pedal or gas pedal position more than 10 % for braking or accelerating. Furthermore, the maximum longitudinal (AccLong) and lateral acceleration (AccLat), and the minimum TTC were analyzed in the experiments. Generally, greater accelerations and smaller TTCs indicate worse take-over quality. Furthermore, based on Gold (2016), analyses were conducted with regard to which action participants consciously performed initially (steering, braking, accelerating) (initial response, abbr. InRe), compared to which maneuver participants ultimately chose to resolve the take-over situation (final response, abbr. FinRe). Maneuver options were full braking in the ego-lane (v<25 km/h) or changing the lane to avoid the obstacle. Rearward and sideward traffic needed to be checked by the participants to determine whether an evasive maneuver towards the adjacent lane or a full-braking maneuver is safe. Thus, it was analyzed whether participants looked into the side-view or rearview mirror consciously before conducting a maneuver (mirror check, abbr. MC) (Gold et al., 2013). Furthermore, the occurrence of collisions with other objects (crash), especially the broken-down vehicle, was examined.

It would have made sense to examine timing metrics mainly assessing automatic responses, since they are also assumed to be affected by fatigue (see chapter 4.5 and chapter 4.9). However, in the majority of the experiments, the participants were not engaged in a NDRA due to the research questions and therefore, do not necessarily have their eyes averted from the road. That is why the gaze reaction time was not a feasible metric. Neither could hands-on-steering-wheel time and feet-on-pedal time be evaluated, because data are not reliably assessable by means of the available procedure. However, the TOT consolidates the durations of these previous steps, which is why it is assumed that strong deteriorations of these response times would be displayed in the TOT. Together with quality metrics, a comprehensive picture of the take-over performance can be provided.

6.6 Data Processing and Statistical Analysis

The eye tracking data recorded by Smart Eye were logged together with the driving and environmental data by the driving simulator software SILAB. The data relevant for the take-over process and for evaluating fatigue were all processed using MATLAB R2017a by MathWorks, Inc. Illustrations were generated using Microsoft Excel for Office 365.

Statistical analysis was performed using the JASP computer software (Version 0.14) by JASP Team (2020) or IBM SPSS Statistics version 25. A significance level of $\alpha=0.05$ was assumed throughout all analyses. The respective statistical analysis (e.g., t-test or analysis of variance, abbr. ANOVA) and corresponding assumption checks are reported in the results sections of the individual experiments because it strongly depends on the study design of the experiment and the data properties. In case of need for the statistical test, normal distribution of data was always tested using the Shapiro-Wilk test; equal variances between groups were tested using the Levene's test and Hartley's Fmax test (in case of need). Results of the Shapiro-Wilk tests and the Levene's tests are reported qualitatively (significant/not significant) in the results sections; exact test results with all relevant values are reported in the Appendix.

All boxplots in this thesis display the median *Mdn*, first quartile (Q1) and third quartile (Q3) as well as the mean value (depicted by a cross) of the data. Whiskers represent the minimum and maximum values of the data, dots represent outliers, which are defined as Q1 minus 1.5 times the interquartile range (IQR) or Q3 plus 1.5 times the IQR.

7 Experiments

7.1 Experiment 1: Effect of Prolonged Periods of Conditionally Automated Driving on the Development of Fatigue—With and Without Non-Driving-Related Activities

This study[2] and its results have been pre-published in Feldhütter et al. (2019). Some parts of the written text were adopted literally from the paper. Figures, tables, and data analyses were adapted for a consistent representation in this thesis.

7.1.1 Research Questions and Purpose of this Study

As outlined in chapter 4.6 and chapter 4.9, NDRAs could be an effective countermeasure for the development of fatigue, even though the results are not consistent, especially the effectiveness for take-over performance. It is assumed that, in past studies, the mandatory NDRA may not have had the same intrinsic motivational character for each participant, and, therefore, did not have the same stimulating effect on each participant. It is even possible that prolonged mandatory engagement in these tasks could have caused active TR fatigue due to their high workload. According to Richter and Hacker (1998), an active organism is constantly willing to maintain the load level—i.e. her/his state—at an optimum (as cited in Hargutt, 2003). Therefore, it is assumed that a driver's voluntary engagement in optional NDRAs might lead to the absence of fatigue thanks to their intrinsic motivation and corresponding self-regulating behavior. This means that participants will find the activity that prevents boredom and fatigue for them personally—even for prolonged phases—since these states are undesirable and uncomfortable. As proposed by Saxby et al. (2007) and Saxby et al. (2013), an automation duration of 60 minutes is defined to be prolonged in this study, since this duration is longer than the durations of most former studies and would suffice evoke high levels of passive TR fatigue and performance decrements. Furthermore, this represents a realistic scenario for the future application of CAD in real road transport and follows the suggestion of J. Schmidt et al. (2017) and Weinbeer et al. (2019) to examine the effect of free and naturalistic engagement in NDRAs.

The aim of experiment 1, therefore, is to investigate in an exploratory way

1) how naturalistic and voluntary NDRAs affect the driver's fatigue state during prolonged monotonous periods of CAD.

For this purpose, two activity conditions during a prolonged automated drive were defined for comparison: the natural load condition (NLC) and the underload condition (UC) as baseline. In

[2] The study was conducted with the assistance of Tobias Hecht as part of his Master's thesis (Hecht, 2016).

the UC, participants were not allowed to engage in any activity to encourage passive TR fatigue. It is assumed that participants in the UC show an increase of fatigue over time. In the NLC, participants were allowed to engage in arbitrary activities to prevent them from becoming bored and fatigued. It is assumed that active participants will not show the same development of fatigue as participants in the UC. Consequently, it is assumed that participants in the NLC show better take-over performances because of the absence of fatigue than participants in the UC, who are assumed to be fatigued. As described in chapter 4.5 fatigue has numerous and variable effects on elementary cognitive functions, which is expected to affect take-over performance. How the effect of fatigue will manifest itself in detail in the individual take-over performance metrics is not clear, since there is no consistent picture from previous research. Prolonged response times have often been reported as a result of fatigue in manual driving (see chapter 4.5).

The second aim of experiment 1 is, therefore, to examine in an exploratory way

2) how fatigue resulting from the manipulation from the two activity conditions affects the take-over performance.

In this study, fatigue is assessed by a retrospective expert rating based on the observation of fatigue-specific behavior described in chapter 4.8.5, which represents the ground truth for fatigue. This method was selected due to its high validity, reliability and low intrusiveness (Platho et al., 2013). As suggested by Platho et al. (2013), fatigue is assessed by means of an additional method, which is the promising and intensively researched eye-based PERCLOS metric (cf. chapter 4.8.2.1). Literature review, however, has also revealed that the classification in different fatigue states or levels based on PERCLOS is not distinct. Literature only suggests that a multi-level classification makes sense. Moreover, only little is known about PERCLOS when it is applied in CAD and drivers engage in NDRAs. Data availability and quality of PERCLOS is examined depending on the activity condition and compared to the expert rating of fatigue.

Therefore, a third aim of this study is

3) the methodological evaluation of the PERCLOS metric for both activity conditions and an examination of the thresholds proposed in literature.

7.1.2 Method

7.1.2.1 Participants

A total of 42 drivers took part in the experiment of which 20 participants were in the UC and 22 participants in the NLC. Nine participants (45 %) in the UC and six participants (27 %) in the NLC were female. The age of the participants ranged in the UC between 23 and 70 years with a mean age of M=46.35 years (SD=19.42) and in the NLC between 23 and 77 years, with a mean age of *M*=45.59 years (*SD*=20.92). Participants in UC had an average driving experience of *M*=26.05 years (*SD*=18.67; Min=5 years, Max=52 years) and in the NLC of M=27.45 years (*SD*=20.66; Min=4 years, Max=59 years). Twenty-two participants (52 %) had participated in earlier driving simulator studies (UC: 30 %; NLC: 64 %).

7.1.2.2 Study Design

As already mentioned, there were two activity conditions (see Figure 7-1). In the UC, any possibilities to engage in NDRAs were removed from the interior of the vehicle. Furthermore, the participants were not allowed to take any items (e.g., smartphone or watch) into the vehicle. Participants in the NLC had been instructed beforehand in their invitation email to bring personal items (e.g., a laptop for working or a book) that they would like to use for passing extended waiting periods. In addition, the vehicle was equipped with several possibilities for engaging in NDRAs for NLC: for instance, drivers could play games on a computer tablet, listen to podcasts or watch videos, as well as read current newspapers/journals or use the radio device that was provided. The options were oriented towards the findings of Petermann-Stock et al. (2013), who analyzed the activities of drivers and co-drivers in traffic jams.

The automation duration was set to 60 minutes. The time budget in the take-over situation was set to six seconds to have a medium to high urgency and the traffic condition was set to no traffic for a medium complexity (see Figure 7-2).

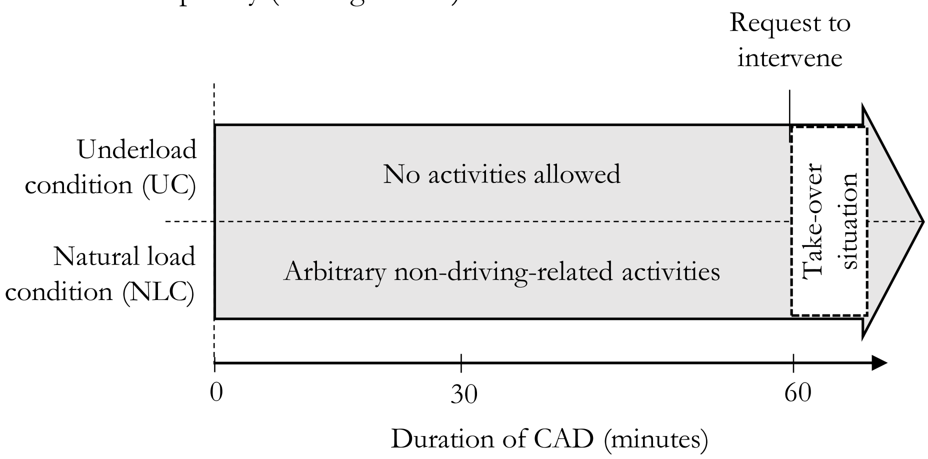


Figure 7-1. Study design of experiment 1 with an automation duration of 60 minutes and two activity conditions.

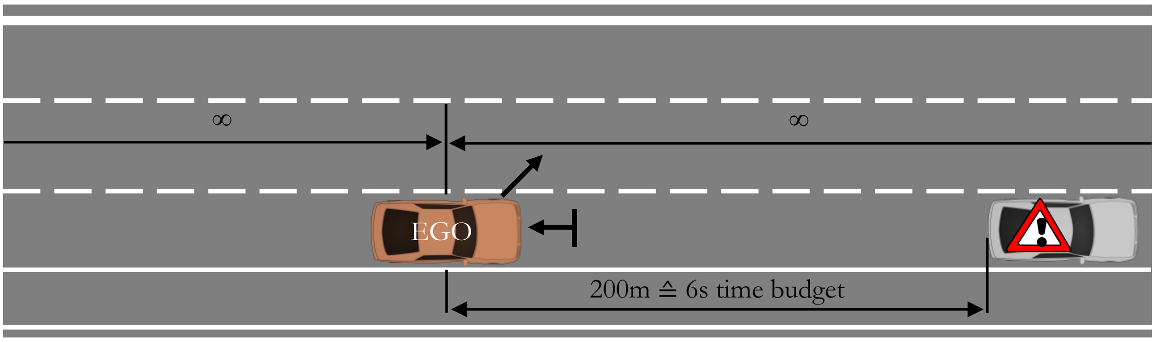


Figure 7-2. Take-over situation of experiment 1. Time budget=6 s, traffic condition=no traffic.

7.1.2.3 Fatigue Assessment

The fatigue rating was retrospectively conducted by two experts that had been intensely trained in advance in a way similar to the protocol proposed by Wiegand et al. (2009) (see also chapter 4.8.5). Video data of each participant were randomly assigned to one of the experts who subsequently annotated participants' behavior for the entire experimental drive according to an annotation guideline (see Appendix) watching recorded video material of the face and upper torso. The guideline is based on the four-level fatigue scales developed by Wierwille and Ellsworth (1994) and Karrer-Gauß (2011), and was supplemented with further fatigue indicators listed in chapter 4.8.5. Because there was only one single rating per participant, an analysis of the inter-rater reliability is not possible. However, a good conformity of the rating behavior of the two raters was ascertained by the previously conducted collective intense training and the clear annotation guideline according to which participants were classified. Based on the annotated fatigue indicators, each participant was continuously assigned to one of the four fatigue levels: fatigue level 1 (FL1) *No fatigue*, fatigue level 2 (FL2) *Questionable state/beginning fatigue*, fatigue level 3 (FL3) *Strong fatigue*, fatigue level 4 (FL4) *Extreme fatigue*. FL3 and FL4 are assumed to be critical in terms of performance decrements and, therefore, are assumed to have a negative effect on take-over performance, since the drivers have strong to extreme difficulties to stay awake, show microsleep events and are on the cusp of falling asleep on these levels.

Data for PERCLOS calculation were delivered by the eye tracking system Smart Eye (see chapter 6.3). An evaluation time frame of 60 seconds was selected during the 60 minutes of CAD according to the definition of Wierwille and Ellsworth (1994) (see chapter 4.8.2.1). Using 60 seconds for calculating PERCLOS instead of 30 seconds PERCLOS is more robust against signal quality variation and short-lasting or erroneously detected fatigue states. The reference value for the 80 % eyelid closure was the individual maximum eyelid opening value of each participant within the first minute of the experimental drive, in which the participants started the drive from standstill on the right lane of the freeway and were still driving manually. Each eyelid opening data point was checked for adequate quality before using it for the PERCLOS calculation. For this purpose, Smart Eye delivers a signal quality estimation for each value ranging between 0 (worst quality) to 1 (best quality). For integrating the data point into the calculation, the quality value for eyelid opening must be greater than 0.5. Furthermore, each PERCLOS value was only valid when it was calculated with at least 70 % of valid data points within the evaluation time frame of 60 seconds. Otherwise, the PERCLOS value was considered not available (N/A) and was discarded. This approach was derived from the ISO/TS 15007:2 (International Organization for Standardization, 2014) according to which data availability lower than 70 % is unacceptable for using eye tracking data. Furthermore, if a calculated PERCLOS value was zero, this minute was treated as a missing data point, since a PERCLOS value of exactly zero is considered implausible and can probably be attributed to internal technical problems of the eye tracker or the circumstance that the eyelid of the participant could not be tracked during the calculation period. The latter eventuates when the head of the participant is in an adverse position or is adversely turned so that the eyes are not properly visible for the cameras.

7.1.2.4 Procedure

Participants were recruited for this experiment via a data base of the Chair of Ergonomics and with the help of notifications at the university and on other public boards. After being welcomed by the experimenter, the participants completed a demographic questionnaire on age, gender, and driving experience. The experimenter instructed the participants in how to use the driving simulator, the capabilities of the CA, and the possible activities (only in NLC) that the participants can engage in while the CA is activated. During a 15-minute period, the participants familiarized themselves with the driving characteristics of the driving simulator and with the CA. During this phase, the participants experienced one RtI, but without a concrete reason for the take-over, so participants only had to regain vehicle motion control and stabilize the vehicle in the ego-lane to become acquainted with the sound and the corresponding required actions (hands back on the steering wheel and feet back on the pedals). Afterwards, the simulation was stopped, and the eye-tracking system was calibrated. Once the participants had no more questions, the experimental track started on the right lane of the three-lane freeway from standstill. Participants were asked by the experimenter to stay on the right lane, to accelerate to 120 km/h and then to activate the CA. After finishing the 60-minute experimental track with the take-over situation in the end, participants were asked to stop on the right hard shoulder and the simulation was stopped. After leaving the vehicle mock-up, participants received their compensation and were free to leave.

7.1.2.5 Data Analysis

The annotation of behavioral indicators for the purpose of fatigue assessment based on video data of face and torso captured by the camera systems was conducted with the Interact annotation software by Mangold International.

Due to quality or technical problems with the camera systems, data of some participants are partly (some minutes during the drive) or completely missing for the analysis of PERCLOS and for the expert fatigue rating, which results in smaller sample sizes for each metric. The exact sample sizes can be abstracted from the tables and illustrations.

7.1.3 Results

7.1.3.1 Fatigue Development Depending on the Activity Condition According to the Expert Rating of Fatigue

Video data of one participant in each condition are completely missing due to technical problems. The fatigue state of one participant in the NLC could not be rated by the experts for several minutes of the drive for reasons of insufficient visibility for the camera. That is why this participant had to be excluded from the data analysis. The fatigue state of one participant in each condition could not be rated during several minutes of the drive, however, the fatigue state of these participants was very stable in the remaining rating intervals, and the periods of missing rating were so short that the missing data points could be disregarded, and the data set was considered for data analysis. This results in a remaining sample size of N=19 participants for UC and N=20 participants for NLC.

All participants except two engaged intensely in NDRAs during the 60 minutes of CAD, meaning that they engaged in the NDRA uninterruptedly or took breaks from their activities with a cumulated length of less than four minutes. The two participants who did not engage in NDRAs intensely stated after the experiment that they were very interested in the technology behind automated driving and wanted to watch how the vehicle performed. One of the two participants remained at FL1 during the entire drive, the other one remained at FL1 for the first 40 minutes of the drive and fluctuated between FL2, FL3 and FL4 for the remaining 20 minutes of the drive.

Figure 7-3 (left) shows the development of fatigue according to the expert rating conducted during the 60 minutes of CAD by displaying the means and 95 % confidence interval (CI) for each minute of the drive depending on the activity condition. The graphical analysis of Figure 7-3 suggests that participants in the UC were on average at a higher fatigue level over the entire drive than participants in the NLC beginning from minute five of the drive. For the statistical analysis of the fatigue level in terms of the activity condition averaged over the entire drive of 60 minutes (cf. Figure 7-3, right), the Mann-Whitney U-test was used, since the assumption of normality was violated. The result reveals that participants in the UC were at a significantly higher fatigue level ($M_{\text{fatigue level,UC}}$=1.90, SD=0.72) when averaging over the entire drive than participants in the NLC ($M_{\text{fatigue level,NLC}}$ =1.44, SD=0.45) with a large effect size (U=280.5, p=0.028, r_B=0.406).

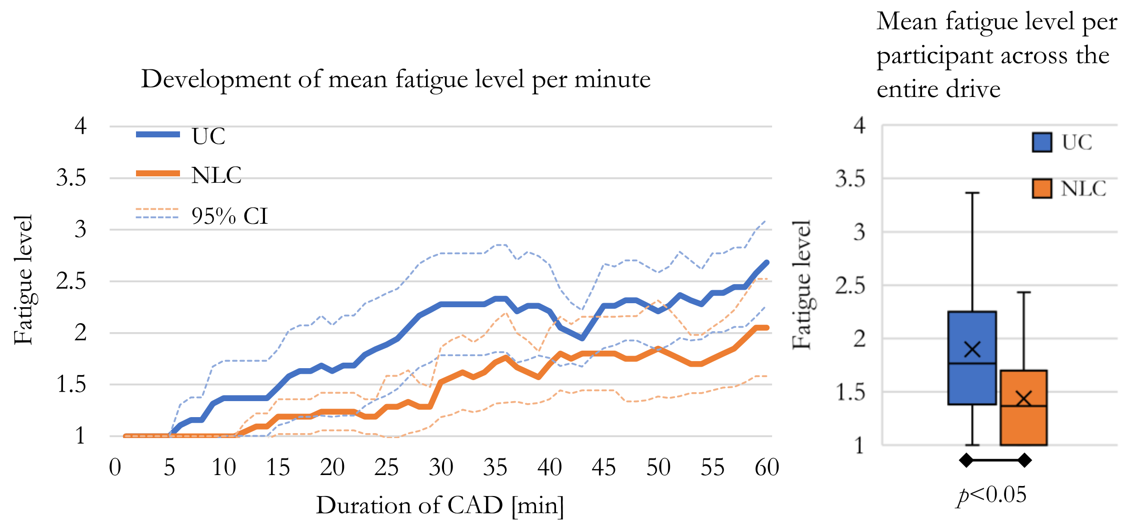


Figure 7-3. Left: Mean fatigue level and corresponding upper/lower 95 % CI of each minute of the drive depending on the activity condition. Right: Boxplot of mean fatigue level per participant averaged over the entire drive depending on the activity condition.

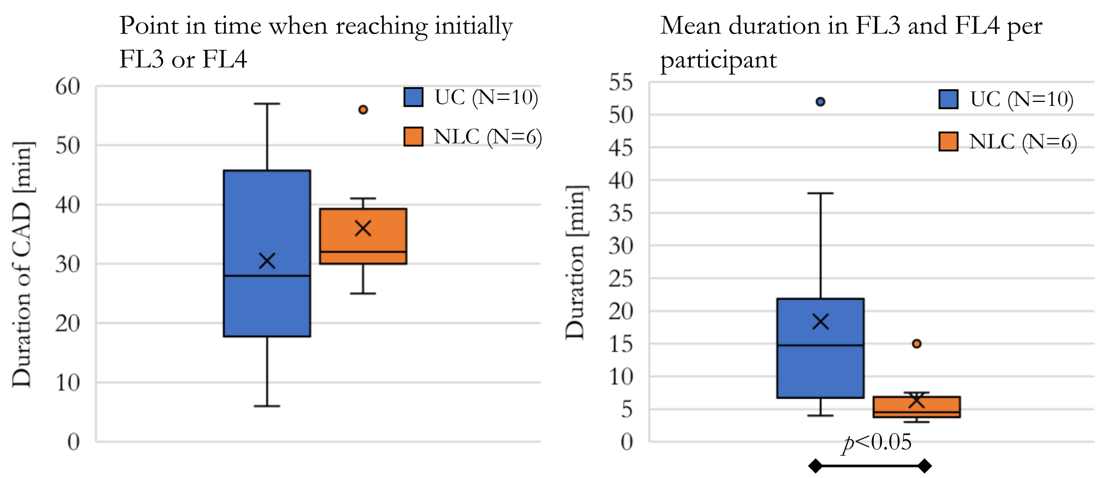


Figure 7-4. Left: Boxplot of the point in time during 60 minutes of CAD when participants initially reached FL3 or FL4 depending on the activity condition. Right: Boxplot of mean duration per participant spent at FL3 and FL4.

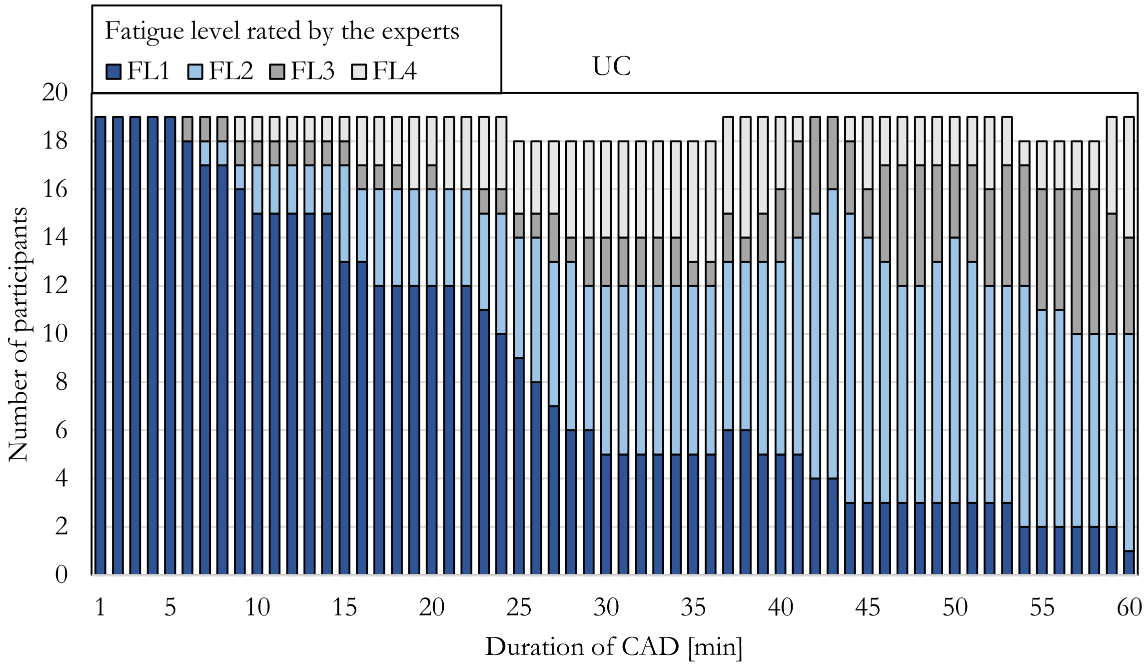


Figure 7-5. Number of participants being in one of the four fatigue levels per minute of the drive for the UC.

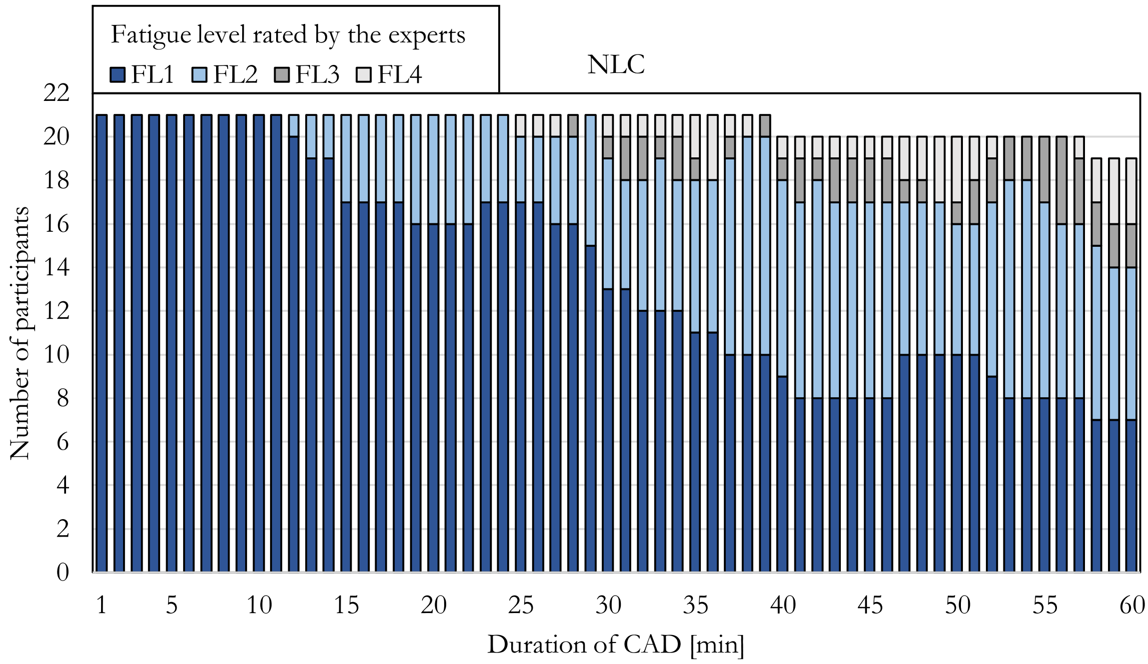


Figure 7-6. Number of participants being in one of the four fatigue levels per minute of the drive for the NLC.

Figure 7-5 (only for NLC) and Figure 7-6 (only for UC) give a more detailed insight into the development of fatigue and display the number of participants at each fatigue level for each minute of the drive. In general, it is noticeable that for both activity conditions, the more time elapses, the more the proportion of participants at FL3 and FL4 increases, even though the individual point in time when FL3 or FL4 was reached varied strongly. This effect is postponed for the NLC: While in the UC, participants initially reached FL3 or FL4, on average, after 32 minutes of CAD (SD=17, Min=6 min, Max=57 min) in the NLC, this was the case after 36 minutes of CAD (SD=10, Min=25 min, Max=56 min) (see also Figure 7-4, left). Student's t-test indicates no significant difference between activity conditions, $t(14)=-0.684$, $p=0.505$, $d=-0.353$. Furthermore, the proportion of participants who reached FL3 and FL4 at all is higher in the UC than in NLC: In total, six of 20 participants (30 %) in the NLC and ten of 19 participants (53 %) in the UC reached FL3 or FL4 at least once during 60 minutes of CAD. A statistical correlation between activity condition and the number of participants reaching FL3 or FL4 was tested with a Chi-squared test. The strength of association (i.e. effect size) is given by the phi coefficient φ. The result was not significant ($\chi^2(1, N=39)=2.063$, $p=0.151$, $\varphi=0.230$). Participants who reached FL3 or FL4 remained there, on average, for 18 minutes (SD=15) when being in the UC and six minutes (SD=5) when being in the NLC (cf. Figure 7-4, right). The Mann-Whitney U-test is used instead of a Student's t-test to test for statistical difference between the activity conditions, since the assumption of normality is violated. The result indicates that participants in UC remained significantly longer at FL3 and FL4 than participants in the NLC with a large effect size ($U=49$, $p=0.044$, $r_B=0.633$). Moreover, seven of 22 participants (32 %) in the NLC and one of 20 participants (5 %) in the UC stayed at FL1 during the entire drive. The Chi-squared test revealed a significant correlation with a medium effect size given by the phi coefficient φ ($\chi^2(1, N=42)=4.886$, $p=0.047$, $\varphi=0.341$).

Furthermore, it is remarkable in the development of fatigue in both conditions that the proportion of participants at FL3 and FL4 fluctuated over time and the point in time when participants reached FL3 or FL4 was not necessarily at the end of the drive but was very individual. A look at the high variance of the individual beginning of FL3 or FL4 of each participant emphasizes this finding (see Figure 7-4, left). For instance, in the last minute of the drive right before the RtI, only nine of 19 participants (47 %) in the UC and five of 20 participants (25 %) in the NLC were still or again at FL3 or FL4 (see Figure 7-5 and Figure 7-6). To examine whether there is an effect of the activity condition on the fatigue level in the minute before the RtI after 60 minutes of CAD, a multinomial logistic regression (MLR) was performed. Results show that the full model is no significant improvement in fit over a null model ($\chi^2(3)=6.498$, $p=0.090$), even though a tendency can be assumed. Looking at the parameter estimates (cf. Table 10), activity condition is not a significant predictor for the fatigue level in the minute before the RtI for all three sub-models, even though p-values are only slightly greater than 0.05. B-coefficients (all negative) and odds ratio Exp(B) indicate that participants of the NLC (coded with 1) tended to be less likely to be rated at a fatigue level higher than FL1 (reference category) after 60 minutes of CAD compared to participants of the UC (coded with 0).

Table 10. Parameter estimates of the MLR with activity condition as independent variable and fatigue level in the minute before the RtI as dependent variable.

Parameter estimates							
Fatigue level before the RtI		***B***	**Std. error**	**Wald**	***df***	***p***	**Exp(B)**
FL2	Intercept	2.197	1.054	4.345	1	0.037	
	Condition	−2.197	1.182	3.456	1	0.063	0.111
FL3	Intercept	1.386	1.118	1.537	1	0.215	
	Condition	−2.639	1.376	3.679	1	0.055	0.071
FL4	Intercept	1.609	1.095	2.159	1	0.142	
	Condition	−2.457	1.295	3.601	1	0.058	0.086

Note: The reference category is FL1.

7.1.3.2 Effect of the Activity Condition on Take-over Performance

Incidentally, five of 21 (24 %) participants were not engaged in an activity when the RtI sounded. However, three of these five participants were engaged in an activity right before the RtI, so that the time to build up sufficient situation awareness most likely did not suffice. This is why they were treated as if they had been engaged in a NDRA. One of the five participants had her/his eyes closed during the RtI, probably because she/he was sleeping.

Two-tailed Student's t-tests for independent samples were conducted to compare means of TOT, AccLong, AccLat and TTC in the two activity conditions (UC and NLC). In case of a violation of the normality assumption, the non-parametric Mann-Whitney U-test was performed. In case of a violation of the variance homogeneity, the Welch's t-test was performed. When both assumptions were violated, the Welch's t-test was chosen over the Mann-Whitney U-test as proposed by Rasch, Kubinger, and Moder (2011) and Ruxton (2006). Effect size is given by the rank biserial correlation r_B for the Mann-Whitney U-test; for the other tests by Cohen's *d*. A Chi-square test was conducted to investigate the existence of a statistical association between the activity conditions and InRe, FinRe, crash and MC (bi- or multinomial data). For expected frequencies smaller than five, Fisher's exact test was used instead. Effect size was either calculated by the phi coefficient (φ) (2x2 contingency table) or by Cramér's *V* (variables other than dichotomous).

Figure 7-7 shows the boxplots of the take-over performance variables TOT, TTC, AccLong and AccLat depending on the activity condition, and Table 11 the corresponding descriptive data. Figure 7-8 and Figure 7-9 display the bi- and multinomial take-over performance variables InRe, FinRe, crash and MC depending on the activity condition. Results of the statistical analysis of all metrical take-over performance data in terms of the activity condition (see Table 12 and Figure 7-7) show that participants in the NLC decelerated significantly more strongly than participants in the UC ($\Delta M_{\mathrm{AccLong,NLC\text{-}UC}}$=2.29 m/s^2) with a medium effect size (t(37.756)=2.362,

p<0.05, d=0.724). No further metrical take-over performance variable was significantly affected by activity condition; however, there was a tendency toward participants in the UC having greater AccLat than participants in the NLC ($\Delta M_{AccLat,NLC-UC}$=1.21 m/s²) with a medium effect size (U=294.000, p=0.064, r_B=0.336). Furthermore, there was no significant association between activity condition and any of the bi- or multinomial take-over parameters (InRe, FinRe, crash and MC) (see Table 13).

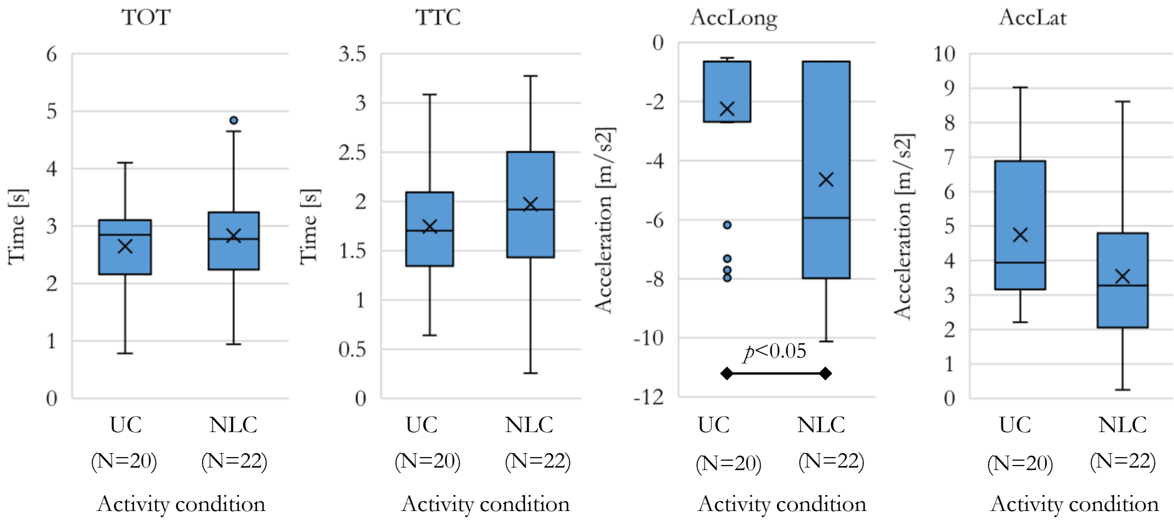


Figure 7-7. Boxplots of all metrical dependent variables of take-over performance depending on the activity condition.

Table 11. Descriptive values of all metrical dependent variables of take-over performance depending on the activity condition.

	Activity condition	*N*	Mean	*SD*	Min	Max
TOT [s]	UC	20	2.65	0.81	0.78	4.10
	NLC	22	2.83	0.95	0.94	4.84
TTC [s]	UC	20	1.75	0.61	0.64	3.08
	NLC	21	1.97	0.76	0.26	3.27
AccLong [m/s²]	UC	20	−2.24	2.69	−7.97	−0.51
	NLC	22	−4.63	3.82	−10.13	−0.64
AccLat [m/s²]	UC	20	4.75	2.19	2.21	9.03
	NLC	22	3.54	2.07	0.25	8.62

Note: In case of a collision with the broken-down vehicle, TTC is 0, hence, it was not considered for further analyses of TTC and sample size was reduced respectively.

Table 12. Results of mean comparisons of the metrical dependent variables between the activity conditions UC and NLC.

	Test	Statistic t (U)	*df*	p	Effect size d (r_B)
TOT	Student	−0.667	40	0.509	−0.206
TTC	Student	−1.034	39	0.308	−0.323
AccLong	Welch	2.362	37.756	**0.023**	0.724
AccLat	Mann-Whitney	294.000	-	0.064	0.336

Note: For the Mann-Whitney U-test, test statistic is given by U-value and effect size by the rank biserial correlation r_B.

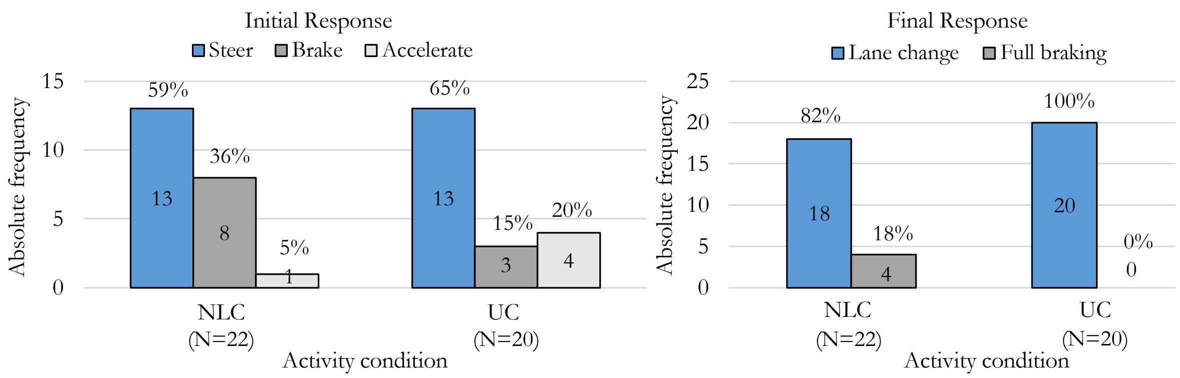


Figure 7-8. Frequency of initial response types (steer, brake, accelerate) and final response types (lane change, full braking) depending on the activity condition.

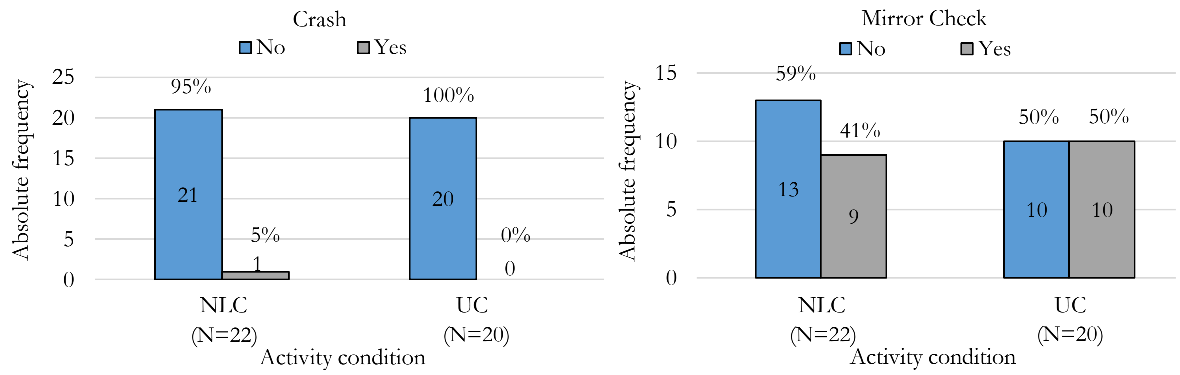


Figure 7-9. Frequency of crash and mirror check depending on the activity condition.

Table 13. Results of Chi-squared tests to compare the frequencies of the bi- or multinomial dependent variables of take-over performance (InRe, FinRe, crash and MC) between the activity conditions NLC and UC.

	Test	Statistic χ^2	N	df	p	Effect size φ/V
InRe	Fisher's exact	3.758	42	2	0.142	0.308
FinRe	Fisher's exact	4.019	42	1	0.109	0.309
Crash	Fisher's exact	0.931	42	1	1.000	0.149
MC	Pearson's Chi-square	0.349	42	1	0.554	0.091

Note: In case that two cells have an expected frequency smaller than five, Fisher's exact test was conducted. For 2x2 contingency tables, effect size is given by phi coefficient φ. For other contingency tables by Cramér's *V*.

7.1.3.3 Effect of Fatigue Level on Take-over Performance

To evaluate the effect of the fatigue level in the minute before the RtI on the take-over performance, an independent one-way ANOVA with the four-level factor fatigue level (FL1, FL2, FL3, FL4) was conducted for each metrical take-over performance parameter (TOT, AccLong, AccLat and TTC). Even though the sample sizes for the four fatigue levels were different (cf. Table 14), the ANOVA was the adequate method for statistical analysis, since the non-significant Levene's test indicates equal variance of the groups. Due to the small and different sample sizes, ω^2 was used instead of η_p^2 to indicate effect size. For AccLong the assumption of normality was violated, which is why the non-parametric Kruskal-Wallis H-test was conducted instead of the ANOVA. To examine the association of fatigue level in the minute before the RtI and the bi- and multinomial dependent variables of the take-over performance (InRe, FinRe, Crash, MC), Chi-squared tests were performed. However, some of the sample sizes of the fatigue level groups are very small (cf. Figure 7-11 and Figure 7-12), which is why the meaningfulness of the statistical examination may be questionable.

Figure 7-10 shows the boxplots of the take-over performance variables TOT, TTC, AccLong and AccLat depending on the fatigue level in the minute before the RtI, and Table 14 the corresponding descriptive data. Results of the statistical analysis are displayed in Table 15. The ANOVAs show that fatigue level before the RtI has a significant effect on the TOT with a large effect size ($F(3, 34)=3.862$, $p=0.018$, $\omega^2=0.184$). Post-hoc testing using Bonferroni correction reveals that the TOT of participants at FL3 was significantly smaller than the TOT of participants at FL4 ($\Delta M_{\mathrm{TOT,FL3\text{-}FL4}}=1.42\,\mathrm{s}$, $p=0.017$). Furthermore, the TOT of participants at FL3 tended to be smaller than the TOT of participants at FL1 ($\Delta M_{\mathrm{TOT,FL1\text{-}FL3}}=1.21\,\mathrm{s}$, $p=0.059$). Besides, results of the ANOVA indicate a significant effect of the fatigue level before the RtI on AccLat with a large effect size ($F(3, 34)=3.615$, $p=0.023$, $\omega^2=0.171$). Post-hoc testing using Bonferroni correction shows that the AccLat of participants at FL4 was significantly higher than the AccLat of participants at FL1 ($\Delta M_{\mathrm{AccLat,FL1\text{-}FL4}}=3.32\,\mathrm{m/s^2}$, $p=0.015$). The chi-squared tests show no significant association of bi- and multinomial take-over performance metrics with the fatigue level before the RtI (cf. Table 16).

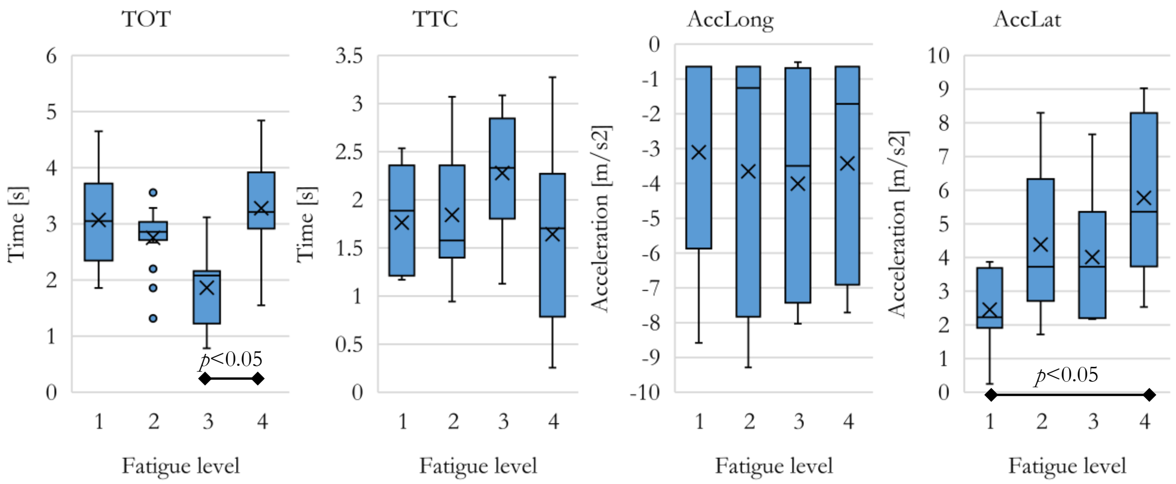


Figure 7-10. The metrical dependent variables of take-over performance depending on the fatigue level in the minute before the RtI.

Table 14. Descriptive data of the metrical dependent variables depending on the fatigue level in the minute before the RtI.

	Fatigue level before the RtI	***N***	**Mean**	***SD***	**Min**	**Max**
TOT [s]	1	8	3.07	0.93	1.86	4.65
	2	16	2.75	0.53	1.32	3.56
	3	6	1.86	0.79	0.78	3.12
	4	8	3.28	0.96	1.55	4.84
TTC [s]	1	7	1.76	0.52	1.17	2.54
	2	16	1.84	0.64	0.94	3.07
	3	6	2.28	0.63	1.13	3.08
	4	8	1.64	0.90	0.26	3.27
AccLong [m/s²]	1	8	−3.11	3.30	−8.58	−0.64
	2	16	−3.65	3.52	−9.28	−0.64
	3	6	−4.00	3.40	−8.03	−0.51
	4	8	−3.43	3.12	−7.71	−0.64
AccLat [m/s²]	1	8	2.45	1.13	0.25	3.87
	2	16	4.39	2.09	1.72	8.29
	3	6	4.01	1.85	2.17	7.66
	4	8	5.77	2.24	2.53	9.03

Note: In case of a collision with the broken-down vehicle, TTC is 0, hence, it was not considered for further analyses of TTC and sample size was reduced respectively.

Table 15. Results of the mean comparisons of the four metrical dependent variables of take-over performance between the fatigue level in the minute before the RtI.

	Test	df_{Within}	$df_{Residual}$	Statistic $F\,(H)^{a}$	p	Effect size ω^2
TOT	ANOVA	3	34	3.862	**0.018**	0.184
TTC	ANOVA	3	33	0.954	0.429	0.000
AccLong	Kruskal-Wallis	3	-	0.655	0.884	-
AccLat	ANOVA	3	34	3.615	**0.023**	0.171

[a] For Kruskal-Wallis test, test statistic is indicated by H-value instead of F-value.

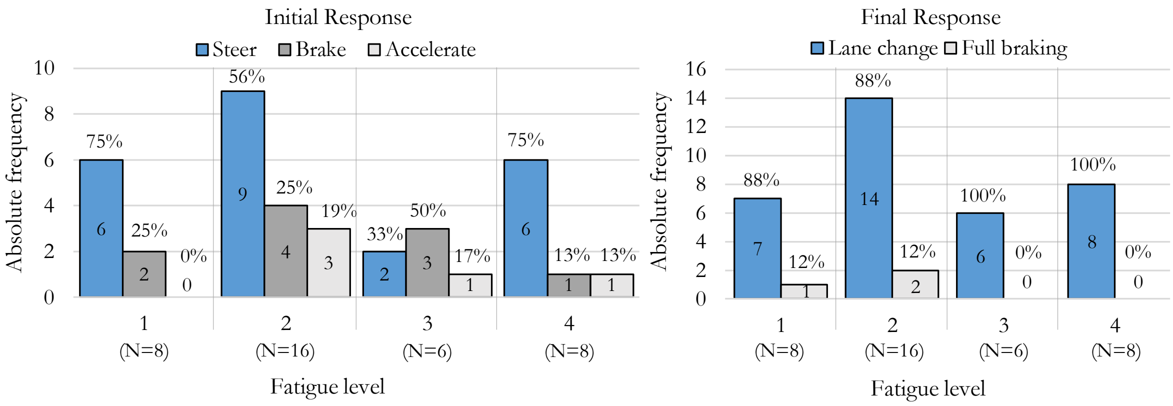


Figure 7-11. Frequency of initial response types (steer, brake, accelerate) and final response types (lane change, full braking) depending on the fatigue level. Note: For rounding reasons the percentage value might not sum up to 100 %.

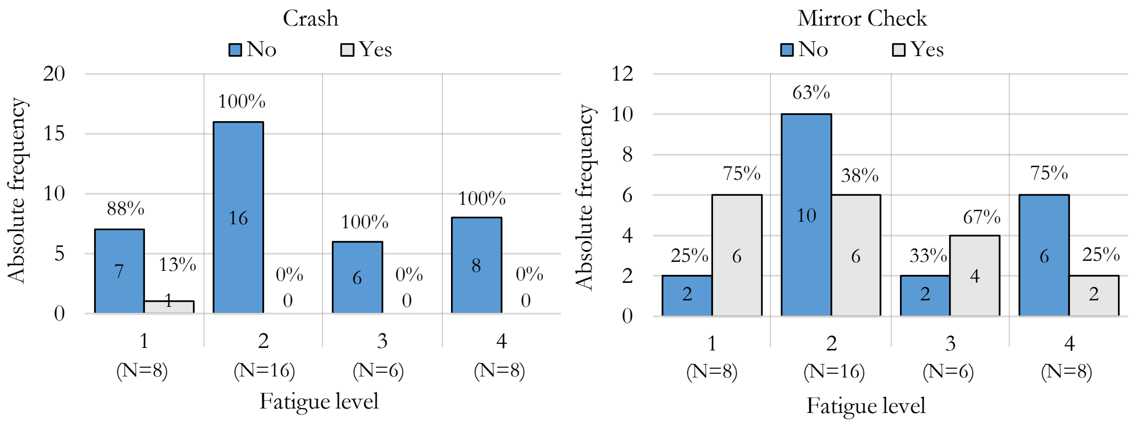


Figure 7-12. Frequency of crash and mirror check depending on the fatigue level. Note: For rounding reasons the percentage value might not sum up to 100 %.

Table 16. Results of Chi-square tests to compare the frequencies of the bi- and multinomial dependent variables of take-over performance (InRe, FinRe, crash and MC) between the fatigue levels in the minute before the RtI.

	Test	Statistic χ^2	N	df	p	Effect size φ/V
InRe	Fisher's exact	4.632	38	6	0.627	0.249
FinRe	Fisher's exact	1.611	38	3	0.879	0.224
Crash	Fisher's exact	3.510	38	3	0.579	0.318
MC	Fisher's exact	5.321	38	3	0.151	0.383

Note: In case that two cells have an expected frequency smaller than five, Fisher's exact test was conducted. For 2x2 contingency tables, effect size is given by phi coefficient φ. For other contingency tables by Cramér's *V*.

7.1.3.4 Evaluation of PERCLOS and Comparison to Fatigue Expert Rating

PERCLOS was calculated using 60-second intervals, which theoretically led to 1200 PERCLOS values for the UC (20 participants x 60 minutes of CAD), and 1320 in the NLC (22 participants x 60 minutes of CAD). For quality reasons, the data processing described in chapter 7.1.2 resulted in a reduced PERCLOS availability for both activity conditions: in the UC, 1068 of 1200 possible PERCLOS values (89 %) were available and in the NLC, 748 of 1320 possible PERCLOS values (57 %). The reduced availability of PERCLOS values led to a strongly varying sample size when calculating the mean PERCLOS per minute across all participants of the respective activity condition (see Figure 7-15): the sample size in the UC is between 14 and 20 (of a maximum of 20 available PERCLOS values) and between eight and 18 (of a maximum of 22 available PERCLOS values) in the NLC (see Figure 7-13). Figure 7-14 shows the frequency distribution of PERCLOS availability per minute: for the UC, the PERCLOS availability was 80 % or higher in 98 % of the minutes (equal to 59 minutes), whereas for the NLC, in only 2 % (equal to 1 minute) of the evaluated 60 minutes the PERCLOS availability was 80 % or higher. Furthermore, for the NLC, the PERCLOS availability was 60 % or less in the majority of the minutes (65 % or 39 minutes) of the drive, and in 20 % of the minutes (equal to 12 minutes) the availability was less than 50 %.

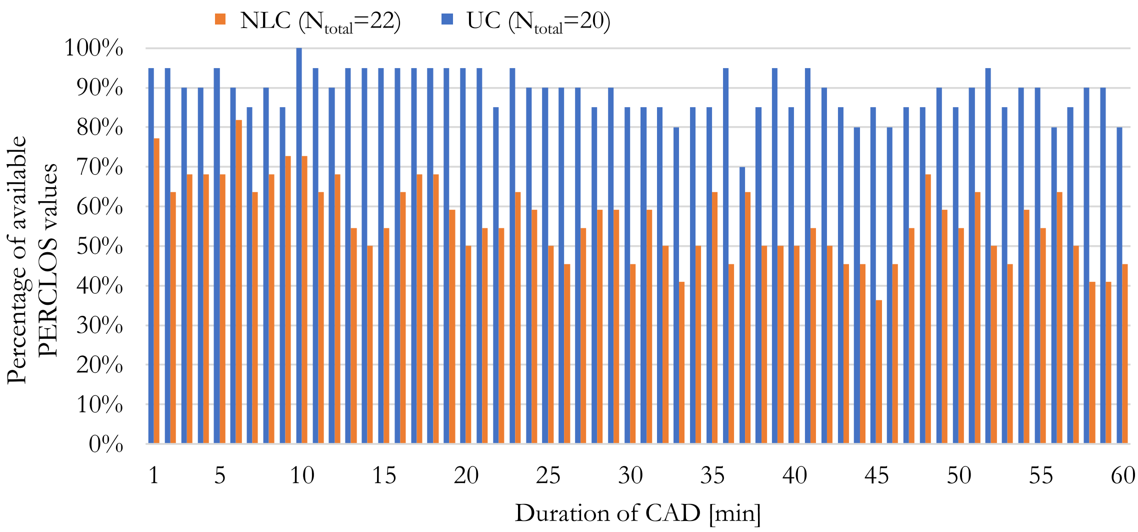


Figure 7-13. Percentage of available PERCLOS values for each minute of the drive relative to the maximum available PERCLOS values depending on the activity condition.

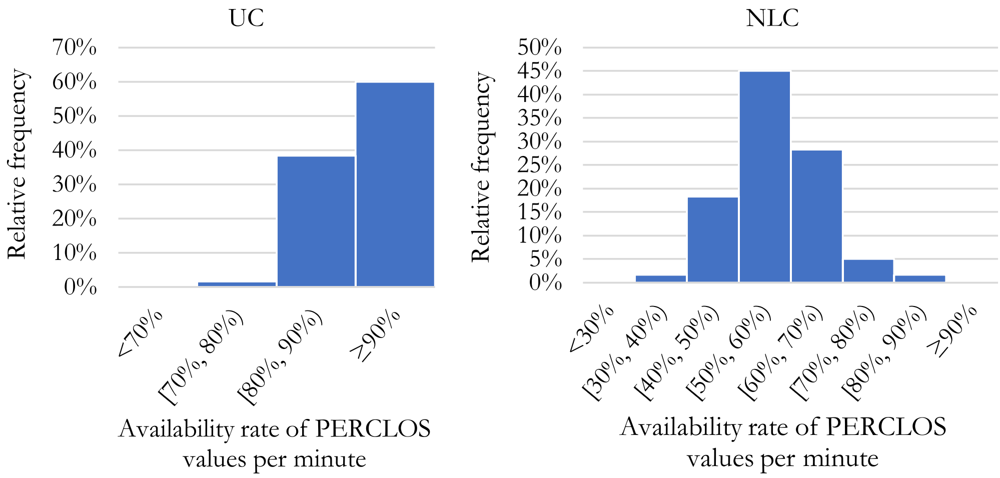


Figure 7-14. Frequency distribution of PERCLOS availability rate per minute depending on the activity condition.

Figure 7-15 (left) displays the mean PERCLOS and 95 % CI using the available PERCLOS values for each minute of the drive depending on the activity condition. It is noticeable that for some periods, the mean PERCLOS of the NLC is greater (indicating higher fatigue) than the mean PERCLOS value of the UC (e.g., mainly in minute 1–11 or minute 45–60), which does not correspond to the fatigue level rated by the experts (cf. Figure 7-3 and subchapter 7.1.3.1). According to that, participants in the UC were in a significantly higher fatigue level when averaging over the entire drive. The statistical analysis of PERCLOS averaged over all minutes of the drive and all participants using Welch's t-test reveals no significant difference between the activity conditions (see Figure 7-15, right; $M_{\mathrm{PERCLOS,UC}}$=5.11 %, SD=7.80; $M_{\mathrm{PERCLOS,NLC}}$=3.09 %, SD=3.54; $t(26.197)=1.035$, $p=0.310$, $d=0.326$). Since both assumptions, of normality and of variances equality, are violated, Welch's t-test was chosen over the Mann-Whitney U-test as proposed by Rasch et al. (2011) and Ruxton (2006).

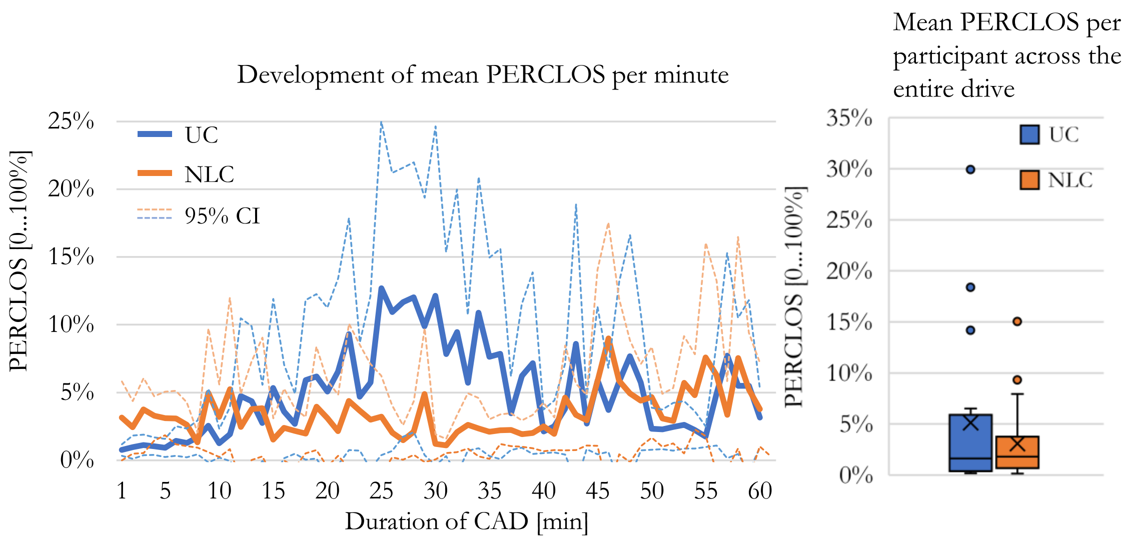


Figure 7-15. Left: Development of mean PERCLOS and 95% confidence interval for each minute of the drive depending on the activity condition. Right: Boxplot of mean PERCLOS averaged over the entire drive per participant depending on the activity condition.

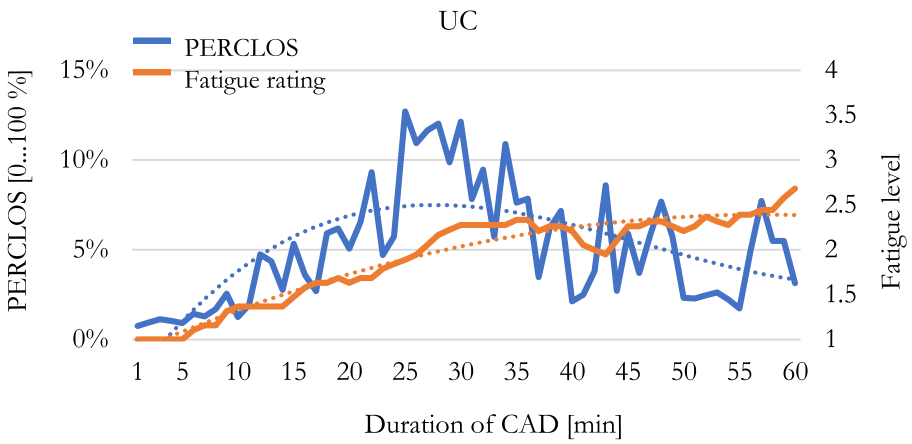


Figure 7-16. Mean PERCLOS and mean fatigue level averaged over all participants of the UC per minute of the drive with a polynomial trend line (each of degree three).

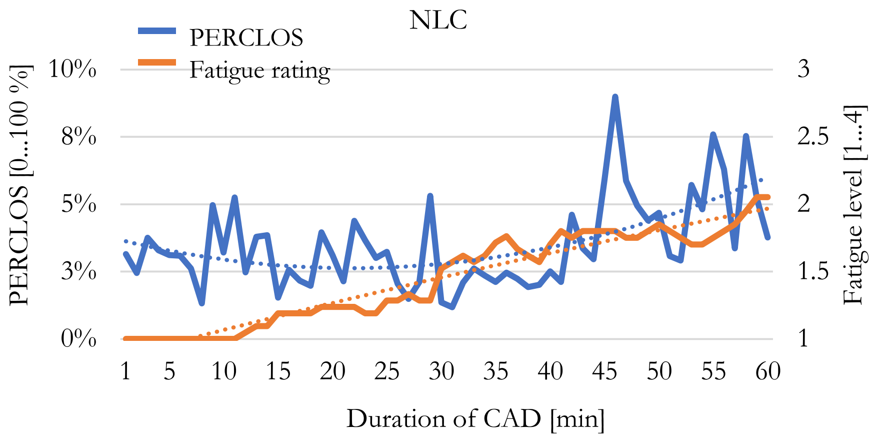


Figure 7-17. Mean PERCLOS and mean fatigue level averaged over all participants of the NLC per minute of the drive with a polynomial trend line (each of degree two).

When comparing the fatigue level and PERCLOS, averaged per minute over all participants only for the UC, the progression of both metrics during the drive seem to coincide for the first 35 minutes of the drive, since both trend lines have a positive gradient (cf. Figure 7-16). For minute 35 to 60 mean PERCLOS tends to develop downwards, which does not correspond to the development of the mean fatigue level rated by the experts. When making the same comparison for the NLC (cf. Figure 7-17), the progression behaves the other way around: for the first 35 minutes of the drive mean PERCLOS does not coincide with the mean fatigue level, since the trend lines are reverse. From minute 35 on, the trend line of PERCLOS turns and evolves as a positive gradient like the fatigue level rated by the experts.

More insights on the evaluation of validity of PERCLOS is provided by Figure 7-18, which shows the boxplot of all available PERCLOS values depending on the fatigue level and the activity condition. Table 17 provides selected descriptive data corresponding to the boxplot; these are the number of participants $N_{\text{participants}}$ reaching the respective fatigue level, number of available PERCLOS values N_{PERCLOS}, mean PERCLOS over all available PERCLOS values M_{PERCLOS} in the respective fatigue level, the PERCLOS median of all available PERCLOS values Mdn_{PERCLOS} in the respective fatigue level, the IQR of all PERCLOS values IQR_{PERCLOS} in the respective fatigue level as a measure of variance and the relative number of normal upper outliers h_{outliers} in the respective fatigue level defined as being 1.5 times the IQR above Q3. Furthermore, the relative number of extreme outliers $h_{\text{outliers7\%}}$ is evaluated. Extreme outliers are defined here by a fixed PERCLOS value of 7 %. This threshold is derived from literature and is correlated with a change from an awake/non-fatigued state to a questionable state (Wierwille & Ellsworth, 1994) or a state of beginning fatigue (G. Wu et al., 2018) (cf. also subchapter 4.8.2.1). When occurring at FL1 and FL2, PERCLOS values of 7 % or greater can justifiably be assumed to be a false positive in terms of increased fatigue.

In the UC, a clear increase of M_{PERCLOS} and Mdn_{PERCLOS} is noted for increasing FLs: M_{PERCLOS} of FL2 is 2.3 times greater than M_{PERCLOS} of FL1 ($Mdn_{\text{PERCLOS,FL2}}=3.2\text{x}Mdn_{\text{PERCLOS,FL1}}$), M_{PERCLOS} of FL3 is 6.1 times greater than M_{PERCLOS} of FL2 ($Mdn_{\text{PERCLOS,FL3}}=5.8\text{x}Mdn_{\text{PERCLOS,FL2}}$) and M_{PERCLOS} of FL4 is 2.5 times greater than M_{PERCLOS} of FL3 ($Mdn_{\text{PERCLOS,FL4}}=2.6\text{x}Mdn_{\text{PERCLOS,FL3}}$). In the NLC, there is not such a consistent evolvement of M_{PERCLOS} and Mdn_{PERCLOS} with the four FLs ($M_{\text{PERCLOS,FL2}}=1.4\text{x}M_{\text{PERCLOS,FL1}}$; $M_{\text{PERCLOS,FL3}}=1.05\text{x}M_{\text{PERCLOS,FL2}}$; $M_{\text{PERCLOS,FL4}}=1.1\text{x}M_{\text{PERCLOS,FL3}}$; $Mdn_{\text{PERCLOS,FL2}}=1.6\text{x}Mdn_{\text{PERCLOS,FL1}}$; $Mdn_{\text{PERCLOS,FL3}}=0.7\text{x}Mdn_{\text{PERCLOS,FL2}}$; $Mdn_{\text{PERCLOS,FL4}}=0.7\text{x}Mdn_{\text{PERCLOS,FL3}}$). Moreover, even though h_{outliers} at FL1 and FL2 is relatively high in both activity conditions and at a similar level around 20 %, the IQR_{PERCLOS} in the NLC is almost three times greater than in the UC for FL1 and 2.5 times greater at FL2. $h_{\text{outliers7\%}}$ is distinctly smaller for the UC ($h_{\text{outliers7\%,FL1}}=1\,\%$, $h_{\text{outliers7\%,FL2}}=6\,\%$) than for NLC ($h_{\text{outliers7\%,FL1}}=12\,\%$, $h_{\text{outliers7\%,FL2}}=15\,\%$).

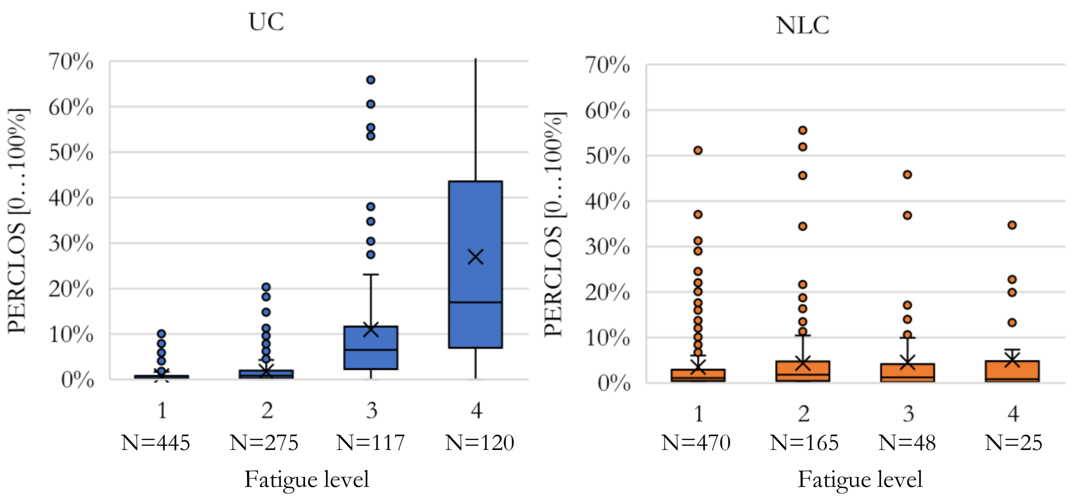


Figure 7-18. Boxplots of all available PERCLOS values depending on the fatigue level and on the activity condition. Note: Not all outliers are displayed for UC.

Table 17. Descriptive data of all available PERCLOS values depending on the fatigue level and the activity condition.

Activity condition		**Fatigue level**			
		1	**2**	**3**	**4**
UC	$N_{participants}$	18	18	10	10
	$N_{PERCLOS}$	445	275	117	120
	$M_{PERCLOS}$	0.77 %	1.80 %	11.01 %	27.00 %
	$Mdn_{PERCLOS}$	0.28 %	0.89 %	6.51 %	16.99 %
	IQR$_{PERCLOS}$	0.67 %	1.69 %	10.05 %	36.63 %
	$h_{outliers}$ **(Q3+1.5xIQR)**	10 %	10 %	10 %	1 %
	$h_{outliers7\%}$ **(>7 %)**	1 %	6 %	48 %	74 %
NLC	$N_{participants}$	20	11	6	5
	$N_{available\ PERCLOS}$	474	166	48	25
	$M_{PERCLOS}$	3.11 %	4.32 %	4.52 %	5.04 %
	$Mdn_{PERCLOS}$	1.10 %	1.79 %	1.21 %	0.81 %
	IQR$_{PERCLOS}$	2.02 %	4.29 %	3.94 %	4.61 %
	$h_{outliers}$ **(Q3+1.5xIQR)**	14 %	8 %	13 %	16 %
	$h_{outliers7\%}$ **(>7 %)**	12 %	15 %	21 %	20 %

7.1.4 Discussion

The results of data analysis revealed that a vast majority of 91 % of participants in the NLC used the available opportunity to engage in voluntary NDRAs. A detailed analysis of the behavior of the participants when engaging in NDRAs, for instance, which types of NDRAs are chosen and on which factors the selection depends, is presented in Hecht, Feldhütter, Draeger, and Bengler (2020).

This engagement in voluntary NDRAs resulted in a significantly higher proportion of participants who stayed in a completely non-fatigued state (FL1) during the entire drive and a 23 % lower proportion of participants who reached FL3 and FL4, even though this difference was not significant. Furthermore, participants in the UC reached FL3 and FL4 (not significant) earlier during the drive and remained at these high FLs significantly longer, which indicates that the fatigue state was more stable on a high level when having nothing to do and being underloaded. All these findings are also reflected in a significantly higher overall fatigue level of participants in the UC when averaging over the entire drive time of 60 minutes. The findings confirm results of previous studies, some of which had a shorter automation duration and which showed that interesting NDRAs can prevent the development of fatigue compared to when no NDRA is available (Neubauer et al., 2014; Schömig et al., 2015; Weinbeer et al., 2019) or only a fatiguing task (Jarosch et al., 2017). Therefore, it can be concluded that the manipulation of the fatigue state by means of different activity conditions was generally successful, also for a prolonged duration of CAD.

However, it also became apparent from the results of this study that voluntary NDRAs cannot completely prevent drivers from high fatigue states during a prolonged drive with CA. According to the theories described in chapter 4.6, fatigue vulnerability depends on personality traits and the transaction between the task and the personality (Ackerman et al., 2012; Phillips, 2014; Szalma, 2012), meaning that motivation and interest to engage in a task are highly individual (Ackerman et al., 2012). In this study, it was attempted to address the latter by asking the participants prior to the experiment to bring their own interesting items for extended waiting times. However, it is conceivable that for some participants in the NLC, the NDRAs brought in and offered were still not interesting or motivating enough for such a long driving duration. At least two participants in the NLC did not engage in available NDRAs for a large proportion of the driving time, which indicates that the NDRAs were not attractive enough to them or observing the CA was more attractive. It is also conceivable that other participants also felt bored or at least not sufficiently entertained by the NDRAs after a while, but nevertheless engaged in them, because they were there, even though they did not have a stimulating effect.

Varying fatigue vulnerability could also be ascertained in the UC. Despite the absence of any activity opportunity, there was still a large proportion (47 %) of participants that did not reach FL3 or FL4. Again following the theories on individual differences in fatigue vulnerability described in chapter 4.6, it is assumed that a great number of these participants who did not evolve higher fatigue states put more effort into fighting fatigue and could compensate upcoming fatigue by self-regulating or self-activating behavior. This manifested in a large number of participants who were rated at FL2 due to mannerisms that were observed (Friedrichs & Yang, 2010; Wierwille & Ellsworth, 1994). This individual vulnerability to fatigue also manifested in a highly individual point in time when FL3 or FL4 initially occurred during the drive in the UC. The first participant already reached FL3 after six minutes of CAD, the latest after 57 minutes, while the mean duration was 32 minutes. Even though the fatigue state in the UC was much more stable over a prolonged time than in the NLC, it was observable that the fatigue state in both activity conditions fluctuated between FL2, FL3 and FL4. This finding is in line with Bittner et al. (2000) and Hargutt and Tietze (2001) (cited in Karrer-Gauß, 2011) according to whom fatigue does not evolves in a linear way but episodically: After periods of increased fatigue signs, periods of absent fatigue signs follow due to increased compensatory effort, however, with a general trend towards increasing fatigue (Karrer-Gauß, 2011). This compensatory effort applies to both activity conditions, although in the UC, fatigue can only be compensated by the described mannerisms, which is assumed to be probably not as effective as NDRAs due to the predominantly successful activity manipulation. However, due to these occasional phases of increased alertness, some participants were no longer at FL3 or FL4 when the RtI sounded, even though the probability of being at FL3 or FL4 tended to be higher in the UC than in NLC.

These findings may also explain the results of the comparison of take-over performance between the activity conditions. Only AccLong differed significantly between the activity conditions, as participants in the NLC produced two times stronger decelerations than participants in the UC, indicating a lower take-over quality (Gold, 2016) for participants in a lower fatigue state. This finding is not as expected, since lower performance was rather assumed for fatigued participants (see chapter 4.5). Indeed, an increased average fatigue level was found in the UC, which was caused by three main reasons: significantly more participants in the NLC completely remained in a non-fatigued state, more participants in the UC reached a high fatigue state, and fatigue states were significantly more stable for participants in the UC. However, there were also some participants who did not correspond to the fatigue level intended for the activity condition, especially in the moment of the RtI. Therefore, it is more likely that the differences in the longitudinal acceleration are not caused by fatigue, but by a very low situation awareness caused by the intense engagement of the participants in NDRAs during the drive and right before the RtI. According to Gold et al. (2013), low situation awareness may result in higher longitudinal accelerations, because a strong braking maneuver extends the time to build up situation awareness and for making a decision on how to react. Hence, it is concluded that voluntary and naturalistic NDRAs might serve as a countermeasure for fatigue, although this is accompanied by a strong distraction. This is, however, not a fatigue-specific issue, since NDRAs are permitted in CAD anyway.

Due to the individual fatigue development beyond the initially formed activity conditions and, especially, because some participants did not correspond to the fatigue level intended for the activity condition in the moment of the RtI, RQ1 cannot be answered by solely analyzing the effect of the activity conditions. Therefore, participants were re-clustered according to their fatigue level at the time of the RtI regardless of the former affiliation to the activity condition, and the effect of the fatigue level on take-over performance was analyzed. Results indicate that participants at FL3 were between about 1 to 1.5 seconds faster in taking over control than participants at the other fatigue levels; the difference compared to FL4 was the greatest and significant. This result is not as expected, since fatigue was assumed to rather prolong take-over time. However, previous studies on take-over performance in CAD did not find a significant effect of fatigue on the take-over time either (Vogelpohl et al., 2018; Weinbeer et al., 2018). Indeed, this would match the fact that there is no deterioration of TOT between FL1 and FL2 compared to FL3 and FL4, but it cannot explain the significant difference between FL3 and FL4 and why participants at FL3 are faster in taking over control than at lower fatigue levels. One explanation might be found in the observation of Weinbeer et al. (2018) that fatigued participants frequently show startle responses when the RtI sounded, which might result in a faster take-over action. However, it remains unclear why this should not apply to participants at FL4. The analysis of take-over performance further revealed that participants at FL4 produced significantly higher lateral accelerations than participants at FL1. This indicates a lower take-over quality (Gold, 2016) for fatigued participants, which is as expected due to the decrement of elementary cognitive functions as a consequence of fatigue described in chapter 4.5. This is also in accordance with the results of Gonçalves et al. (2016), who found significantly higher lateral acceleration for fatigued participants assessed by the self-rating method SSS. Beside these two results, fatigue level has no further significant effect on the take-over performance metrics. However, the sample sizes of the individual fatigue level groups, especially FL3 and FL4, were relatively low since a variation of fatigue was aimed at by the activity conditions and not by a real-time assessment of fatigue during the drive. Thus, the results are relatively limited for a final interpretation.

The evaluation of the metric PERCLOS revealed a huge issue regarding the data availability for the NLC due to system boundaries and due to exclusion of data for quality reasons. The overall data availability is slightly higher than 50 % in the NLC and almost 90 % in UC, with slight variations during the drive. However, the PERCLOS availability of the UC never decreased below 70 %, while for the NLC, the availability of PERCLOS was below 60 % for 65 % of the driving time and even below 50 % for 20 % of the driving time. For the remaining data points in the NLC, the comparison of PERCLOS and the fatigue level rated by the experts revealed an inconsistency of the two metrics in terms of temporal development during the drive and in terms of missing variation of median and mean PERCLOS between the fatigue levels. Also notable is the high number of PERCLOS values above 7 % (false positives) at FL1 and FL2. The described discrepancy between PERCLOS and fatigue rating becomes evident when looking at the videos of the participants in the NLC: When engaging in NDRAs the gaze of the participants is dropped due to the position of the object in their hands (e.g., a smartphone or a book). Hence, the eyelid is more closed than when looking ahead at the road, resulting in a greater PERCLOS value erroneously indicating higher fatigue. Since the majority of participants almost immediately started engaging in NDRAs when the CA was turned on, the mean PERCLOS value of the participants in the NLC was continuously relatively high, already beginning in the first minute of the drive. Since the participants in the UC had no activity to engage in, the inspection of the videos revealed that most of the time they continuously looked at the road and, therefore, the PERCLOS value was not biased. At FL3 and FL4, it is the other way around: due to the great proportion of missing data in the NLC, PERCLOS does not represent the real eyelid closure behavior of the participants resulting in relatively low values. Due to this finding, the validity of PERCLOS of the NLC is highly questionable.

For the UC, however, the graphical curve comparison of the development of both metrics during the drive revealed a good match for UC, at least for the first half of the drive because both curves are rising. Furthermore, the analysis of the PERCLOS values depending on the fatigue level shows a plausible evolution: with increasing fatigue level the median and mean PERCLOS values increase. When comparing the PERCLOS mean (27 %) and median (17 %) of FL4 with PERCLOS values from literature, this reveals a certain compliance with PERCLOS values reported by Wierwille et al. (1994), Hanowski et al. (2008), and G. Wu et al. (2018). At FL3, PERCLOS mean (11 %) and median (6.5 %) do not indicate fatigue that clearly. At FL3 and FL4, there are some downward outliers with low PERCLOS values, indicating no fatigue for these participants. These outliers may be explained by the fact that not only eyelid closure behavior is incorporated in the expert rating, but also other fatigue-specific behavior (for instance mannerisms or head movements). Naturally, this cannot be displayed by the PERCLOS metric, which only includes eyelid closure behavior in the calculation. Especially for participants who are highly fatigued but make a great effort to stay awake (for instance by forcing themselves to keep their eyes open) or if fatigue does not manifest in varied eyelid closure for them, PERCLOS will naturally not be sensitive to detecting fatigue. A further explanation for the downward outliers might be found in the signal quality of the system. During the experiments and when watching the videos, it was observed that the eye-tracking system has deficits in correctly detecting the eyelid for some individuals, as the eyelid fold instead of the eyelid was detected when their eyes were closed, causing the eyes to be classified as open erroneously. These false signals were not always represented in the quality signal provided and, therefore, were not filtered out during data processing. Furthermore, PERCLOS strongly depends on the reference value, which is a fixed value individually taken in the beginning of each drive. Since the driver is not required to look at the road ahead during CAD, it is conceivable that participants in UC also looked somewhere else during the drive. For some participants, it was observed that they did this extensively which might have led to a distortion of the eyelid opening value and, hence, of PERCLOS. In his extensive research on blinking behavior, J. Schmidt (2018) found similar results, as in his studies, eyelid closure behavior varied between manual driving and CAD (see also J. Schmidt, Braunagel, et al., 2016). Therefore, an adaption of the reference value depending on the head tilt would probably also be beneficial even if drivers do not engage in NDRAs.

Limitation of the Study and Conclusion

An expert rating based on behavior observation was used as ground truth for fatigue in this study. Even though an expert rating of fatigue is considered a method of high validity, and it is scientifically recognized (see chapter 4.8.5), an absolute certainty about the true fatigue state of participants cannot be given. Especially for the participants who manifest no typical fatigue signs, the expert rating method is limited. To further raise the certainty of the true fatigue state, a second measurement method, for example, EEG or self-assessment by the KSS, have to be applied and compared, as suggested by Platho et al. (2013).

The evaluation of PERCLOS revealed a very low data availability in the NLC for PERCLOS, caused by a low signal quality and missing tracking due to the head tilt or position in the seat. The PERCLOS calculated from the remaining data showed a distortion due to the natural physiological variation of the eyelids when engaging in NDRAs, resulting in a lack of discriminatory power for fatigue in the NLC. The validity of PERCLOS in terms of indicating fatigue is highly questionable. However, for some participants without NDRAs, a different gazing behavior was also found, indicating that the PERCLOS calculation as it has been used for manual driving needs to be further developed. For instance, taking the current head position into account when determining the reference value is conceivable.

The intended manipulation regarding the activity conditions was successful: the average fatigue level and the proportion of fatigued drivers was lower in the NLC, the onset of fatigue could be deferred, and a stabilization of the fatigue state could be avoided. However, the effect of fatigue levels assumed critical for take-over performance could not properly be evaluated—neither with re-clustered groups—since the sample sizes of FL3 and FL4 before the RtI were rather low. The statistical analysis of the re-clustered groups by a one-way ANOVA instead of a 2 (activity condition) x 4 (fatigue level) ANOVA is discussable. However, it appeared to be correct, because the factor fatigue level was not controlled in the experiment, but is to some extent correlated with the factor activity condition. Either way, to obtain a clearer picture of the effect of fatigue on take-over performance in CAD, it is essential to hand over vehicle control to the driver once the participants reach the critical fatigue level. For this purpose, the fatigue state must be monitored in real time.

Personality factors were not considered in the evaluation of this study. However, especially for fatigue vulnerability, personality plays an important role (cf. chapter 4.6). However, also individual driver characteristics may have a significant effect on take-over performance (Gold, 2016). In this study, for instance, age might have had an impact on both fatigue development and take-over performance. Hecht et al. (2020) report age-related differences in preferences and behavior when engaging in NDRAs. Additionally, the evaluation of fatigue development reveals that all five participants in the NLC and six of nine (67 %) of participants in UC, who were at FL3 and FL4 before the RtI, were older than 60 years. This might be an indicator that NDRAs do not have the same countering effect on all age groups, which has already been found by Y. Wu et al. (2020). One reason might be missing motivation of older participants for the NDRAs available during the experiment. Another reason could be that elderly persons deal differently with boredom and upcoming fatigue. Results suggest that they are more vulnerable. However, Y. Wu et al. (2020) found no increase of perceived fatigue (assessed by the KSS) for participants older than 56 years during a prolonged duration of CAD, regardless of the activity condition (no NDRA, single or multiple NDRAs). On the contrary, young participants between the ages of 19 and 32 years experienced a significant increase of fatigue when they could not engage in NDRAs compared to single and multiple NDRAs. However, the sample sizes in experiment 1 are too small for a systematic examination of the factor age and a profound interpretation. Nevertheless, it cannot be ruled out that results are confounded to some extend by the factor of age. To avoid this in further experiments, age groups should be more homogeneous.

In this study, participants were not instructed in a specific way regarding the engagement with NDRAs. They were only required to behave as naturally as possible, imagining that the experimental drive was their commute. Naturally, this lead to the fact that not all participants were engaged in a NDRA during the RtI. Therefore, it has to be kept in mind that the results regarding the effect of the NDRAs on take-over performance found in this study may be limited to some extent. However, the risk accompanied by the not strictly controlled condition of NLC was approved, since one purpose of the study was to evaluate the effect of free choice of NDRAs during CAD on the fatigue development to gain insights into a realistic scenario regarding the future application of CAD. A methodological balance needs to be found between a study condition that is as natural as possible to prevent fatigue, and as controlled as possible to draw a final conclusion on the effect of NDRAs on take-over performance.

7.2 Experiment 2[3]: Effect of Automation Duration and Non-Driving-Related Activities on Fatigue Level and Take-Over Performance

7.2.1 Research Questions and Purpose of this Study

As outlined in chapter 4.6 and chapter 4.9, NDRAs might be an effective countermeasure for the development of fatigue, even though the results are not consistent, especially the effectiveness for take-over performance. In experiment 1 (cf. chapter 7.1), voluntary engagement in naturalistic NDRAs may serve as a countermeasure for fatigue for a certain proportion of drivers. However, this is accompanied by a strong distraction and missing situation awareness in the moment of the RtI, potentially leading to poor take-over quality or at least confounding the potential positive effect. Results of experiment 1 in which age varied strongly suggest that age potentially affects the interdependency with NDRAs and fatigue vulnerability. Therefore, age of participants was restricted to be below 60 years to eliminate this potentially influencing factor. As in experiment 1, two activity conditions are defined in this study: the NDRA condition (NC) and the underload condition (UC) as baseline. The UC was configured identically to the UC in experiment 1: participants were not allowed to engage in any activity to encourage the development of passive TR fatigue, assuming an increase over time. For the NC, instead of a completely free choice of NDRAs, the computer game Tetris was selected as the NDRA for the participants, in order to have more standardized experimental conditions in terms of location of the item and engagement parameters mentioned in chapter 3.3.3 and classified by Naujoks et al. (2018) or Jarosch, Gold, et al. (2019) (e.g., necessity of hand-holding, effort to disengage, etc.). This game requires continuous visual and cognitive attention and due to its popularity it is, at the same time, an interesting and motivating naturalistic activity (Befelein, Boschet, & Neukum, 2018), especially for a young to middle-aged testing group. Therefore, it is expected that Tetris will counteract the development of fatigue.

Consequently, following experiment 1, the first aim of experiment 2 is to further examine in an exploratory way

1) how an interesting and motivating NDRA using the computer game Tetris affects the driver's fatigue state during prolonged monotonous periods of CAD.

To isolate the activating effect of the NDRA on take-over performance and to compensate for the distraction and missing situation awareness by the motoric and cognitive engagement when playing Tetris, participants are requested in this experiment to stop playing Tetris right before the RtI and to monitor the driving environment. Consequently, it is assumed that participants in the NC show no greater decelerations due to missing distraction and enough situation awareness. Furthermore, like in experiment 1, better take-over performance is assumed for

[3] The study was conducted with the assistance of Kaiyu Huang as part of his Master's thesis (Huang, 2018).

participants of the NC than participants in the UC because of the absence of fatigue. How the effect of fatigue will manifest in detail in the individual take-over performance metrics is not clear, yet, since the results of experiment 1 could not give a clear direction.

Therefore, following experiment 1, the second aim of experiment 2 is to examine in an exploratory way

2) how the fatigue resultant from the manipulation from the two activity conditions affects take-over performance.

As seen in experiment 1 and as expected from literature review (cf. chapter 4.4), fatigue evolves with time-on-task and with elapsing driving time, respectively, meaning that fatigue development is a temporal process. A NDRA potentially decelerates or mitigates the fatigue evolution process evidently, whereas underload may accelerate the process as shown in experiment 1. Therefore, it is expected that the potential difference in take-over performance resulting from the manipulation by the activity conditions will only become apparent after a certain time period.

Previous studies have addressed this temporal dependency of fatigue evolution and resulting effects on take-over performance (Bourrelly et al., 2019; Feldhütter et al., 2017) by comparing short and prolonged automation durations. They found minor and major deteriorations of response times attributing to increased fatigue.

Following these studies, the third aim of experiment 2 is to examine

3) how the driving duration affects take-over performance.

For this purpose, each participant experiences two identical take-over events: the first one is issued after a very short driving duration of three minutes with CAD (CAD3) to take a baseline of the take-over performance for each participant in a non-fatigued state. The second RtI is issued after a prolonged driving duration of 35 minutes with CAD (CAD35). The driving duration is derived from experiment 1, in which participants in both activity conditions reached a high level of fatigue after about 35 minutes, on average. Moreover, in several previous studies, prolonged driving durations of 25 to 40 minutes with CAD were sufficient to develop fatigue (see chapter 4.9). It is expected that the fatigue that develops due to prolonged CAD will negatively affect take-over performance. However, as seen in experiment 1, participants who engage in NDRAs are not completely immune to fatigue: the more time elapses, the more likely fatigue will also arise in the NC. This suggests that a prolonged driving duration may also affect the take-over performance of participants in the NC, but hypothetically in a mitigated way. Therefore, it is expected that the driving duration has a stronger effect on the take-over performance of participants in the UC than of participants in the NC, suggesting an interaction effect of driving duration and activity condition on take-over performance.

Fatigue is assessed by a retrospective expert rating and the PERCLOS metric as already applied in experiment 1. To obtain further insights on the PERCLOS metric under realistic conditions, a fourth aim of experiment 2 is

4) the methodological evaluation of the PERCLOS metric for both activity conditions and an examination of the thresholds proposed in literature.

For this purpose, data availability and data quality of PERCLOS was examined depending on the activity condition and compared to the expert rating of fatigue, which functions as the ground truth. Since PERCLOS was found not to be reliable enough in the NC in experiment 1 and since Phillips (2014) recommends measuring fatigue indicators on different dimensions, an additional measurement method for assessing fatigue was applied in this study. Subjective assessment is often suggested as a powerful method assessing fatigue, and the KSS is a scale that has frequently been validated and used in the context of driving (see chapter 4.8.1). That is why the KSS was selected for experiment 2 as a further measure to assess fatigue.

7.2.2 Method

7.2.2.1 Participants

A total of 40 drivers attended the experiment who were evenly and randomly distributed over the activity conditions, so that each condition consisted of 20 participants. Nine participants (45 %) in the UC and eight participants (40 %) in NC were female. The age of the participants ranged between 20 and 59 years in the UC, with a mean age of M=28.4 years (SD=8.16), and between 21 and 30 years in the NC, with a mean age of M=25.7 years (SD=2.24). Two participants (10 %) in each condition had driving experience of less than three years; all remaining participants had driving experience of more than three years. Twenty of all participants (50 %) had participated in earlier driving simulator studies (UC: 55 %, NC: 45 %). Thirty-three of the participants (83 %) rated their knowledge in terms of automated driving on a 5-level scale (1 “very unfamiliar” to 5 “very familiar”) at a medium level (3) or lower (UC: 90 %, NC: 75 %).

7.2.2.2 Study Design

As mentioned above, there were the two activity conditions UC and NC (see Figure 7-19). Participants were tested individually and each was assigned to one of the two activity conditions prior to the experiment. Analogously to the UC in experiment 1 (cf. chapter 7.1.2), any possibilities to engage in NDRA were removed from the interior of the vehicle in the UC in this experiment, as well. Furthermore, the participants were not allowed to take any items (e.g., smartphone or watch) into the vehicle. Participants in the NC were provided with a 7-inch

Lenovo tablet computer, which was mounted above the center console at the height of the central information display, so that the tablet could be used in a comfortable sitting position. On the tablet, only the common and popular computer game Classic Tetris was installed. The game could be interrupted by pressing the *Pause* button on the tablet screen.

Participants were instructed to play Tetris for as long as possible. To further incentivize participants and motivate them to engage in the NDRA, a prize of 50 € was given to the player with the highest Tetris score. Participants in the NC were prompted to monitor the driving environment 15 seconds before the RtI by means of a gentle tone and an icon in the instrument cluster (for visual representation see Figure 7-20). The design of the request to monitor (RtM) was inspired by the "soft warning" and "hard warning" concepts of van den Beukel (2016, p. 48) and the study of Large, Burnett, Morris, Muthumani, and Matthias (2017) and has already been applied in a similar setting in the study of Lu et al. (2019).

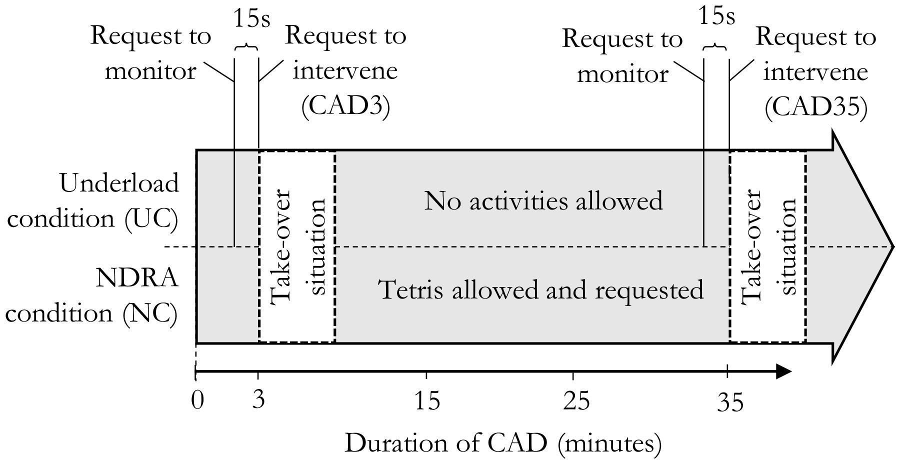


Figure 7-19. The study design of experiment 2 with two take-over situations after different automation durations (three and 35 minutes) and two activity conditions.

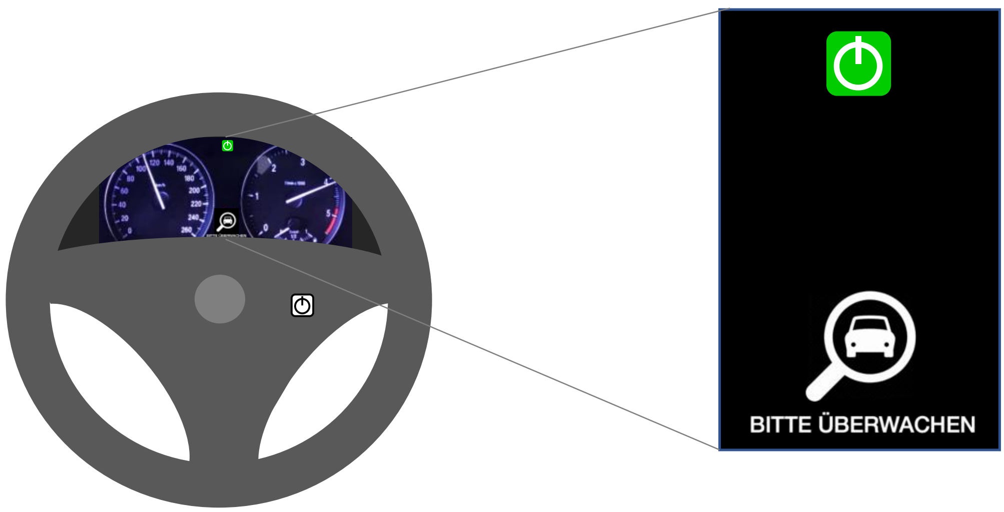


Figure 7-20. Schematic representation of the RtM. The German text says "Please monitor".

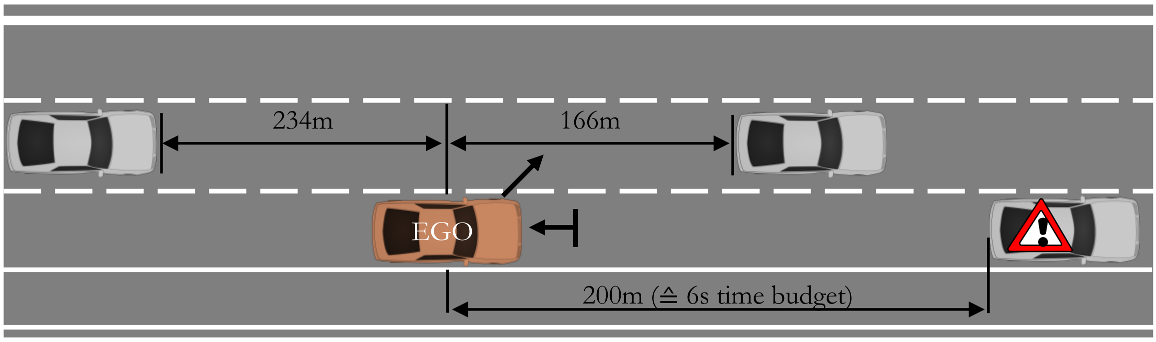


Figure 7-21. Take-over situation of experiment 2. Time budget=6 s, traffic condition=large gap.

After the automation durations of three minutes (CAD3) and 35 minutes of CAD (CAD35), a RtI was issued. The duration of 30 minutes was derived from experiment 1. The take-over situation was the same for both driving durations and was designed similarly to experiment 1: the time budget was again set to six seconds to have a medium to high urgency. The traffic condition was determined to large gap for a medium complexity (see Figure 7-21).

7.2.2.3 Fatigue Assessment

The fatigue rating was conducted analogously to experiment 1, resulting in the continuous assignment of each participant to one of the four fatigue levels FL1–FL4 (cf. chapter 7.1.2). FL3 and FL4 are assumed to be critical in terms of performance decrements and, therefore, are assumed to have a negative effect on take-over performance.

PERCLOS was calculated and filtered according to the defined quality criteria analogously to experiment 1 (cf. chapter 7.1.2.3).

As an additional measure, the German version of the nine-level KSS from Niederl (2007) (cf. chapter 4.8.1) was used to assess fatigue subjectively. The questionnaire was chosen since it is easy to apply and has a good validity (Åkerstedt & Gillberg, 1990). At a fatigue score of seven or higher, drivers have increasing difficulties to stay awake and to avoid driving errors (Platho et al., 2013; Reyner & Horne, 1998); therefore, scores of seven or higher were assumed to be critical for take-over performance. However, to not bias the participants before and during the experiment, the KSS was not provided until the second take-over event was over and the participants stopped the vehicle. Therefore, participants were asked to rate their fatigue level retrospectively for the moment right before the second RtI.

7.2.2.4 Procedure

Participants were recruited for this experiment by notifications at the university and at other public boards. Before the day of the experiment, participants were asked to complete a questionnaire on demographic questions, such as age, gender, and driving experience. Furthermore, test persons were instructed not to drink caffeinated beverages on the day of the experiment if possible. The experiment's starting time was evenly distributed over the day (morning: 8 a.m. after lunch: 1 p.m., evening: 6 p.m.). After being welcomed and introduced to the procedure of the experiment by the experimenter, the participants were instructed in how to use the driving simulator, the capabilities of the CA, and the Tetris game on the tablet (only in NC), which they could engage in while the CA was activated. Additionally, participants in the NC were instructed that the CA sometimes needs to be monitored for performance reasons, which was always indicated by a RtM. Once the RtM sounded participants were expected to interrupt playing Tetris and monitor the driving environment. In some cases, the RtM would result in a RtI. The purpose of the study was not explained until the experiment was finished to not bias the behavior of the participants. During a 15-minute training course starting and ending at a resting area, the participants familiarized themselves with the driving characteristics of the driving simulator and with the CA. To familiarize participants with the event of a take-over and to minimize the effect of training in the two take-overs of the experiment, participants experienced four take-over situations during the training course. All take-over situations were announced by the usual RtI. Participants in the NC additionally experienced the RtM prior to three of the four RtI (prior to the first, the second and the fourth one). One RtI was initiated without a RtM to avoid participants always expecting a RtI after a RtM. The training take-over situations varied between no obvious reason (no obstacle) to different obstacles in the ego-lane (tractor, tire and construction site) and between different lanes. The time budget was eight seconds. After the training course, the simulation was stopped and the eye-tracking system was calibrated. Once the participants had no more questions, the experimental track started from a resting area. Participants were requested by the experimenter to stay on the right lane, to accelerate to 120 km/h and to activate the CA after entering the three-lane freeway. After finishing the 35-minute experimental track with two take-over situations, the participants were asked to stop at the next resting area. Once the participants reached the parking position, the examiner asked them to rate their fatigue level right before the take-over event on the KSS. Then, the simulation was stopped and the participants were asked to leave the vehicle mock-up. After completing the post questionnaire about past night's sleep length and quality and how they perceived the criticality of the take-over situation on a 10-level scale ranging from 1="Not critical" to 10="Extremely critical", the participants received their compensation and were free to leave. The total duration of the experiment was around 80 minutes. After each experiment, the Tetris score was documented, and the game was reset by the experimenter.

7.2.2.5 Data Analysis

The annotation of behavioral indicators for the purpose of fatigue assessment based on video data of the face and torso captured by the camera systems was conducted with the Interact annotation software by Mangold International.

Due to quality or technical problems with the camera systems, data of some participants are partly (some minutes during the drive) or completely missing for the analysis of PERCLOS and for the expert fatigue rating, which results in smaller sample sizes for each metric. The exact sample sizes can be abstracted from the tables and illustrations.

7.2.3 Results

7.2.3.1 Fatigue Development Depending on the Activity Condition According to the Expert Rating of Fatigue

One participant in the NC had to be excluded from the expert rating and PERCLOS calculation due to abnormal blinking behavior caused by dry air and unfamiliar wearing of contact lenses. This results in a remaining sample size of N=19 participants for NC and N=20 participants for UC. Due to the study design, self-rating of fatigue by means of the KSS is only available for the moment right before the second RtI after 35 minutes of CAD, which can be compared between the activity conditions and between fatigue levels. A temporal development between different points in time of the drive is not possible.

Figure 7-22 (left) shows the development of fatigue during the experimental drive according to the expert rating conducted depending on the activity condition by displaying the mean fatigue level averaged over all participants and 95 % CI for each minute of the drive. The graphical analysis of Figure 7-22 (left) suggests that the mean fatigue level of the participants in the UC continuously rises until the end of the drive. The mean fatigue level of participants in the NC stays at a constant low level of FL1 apart from the last seven minutes when a slight increase occurred. The mean fatigue level averaged over all participants and the entire drive of 35 minutes (cf. Figure 7-22, right) was analyzed in terms of differences between the activity condition using the Welch's t-test (assumptions of normality and equal variances were violated). The assumption of interval-scaled dependent variable is valid, since the analysis was conducted with averaged ordinal-scaled data. The result reveals that participants in the UC were at a significantly higher fatigue level ($M_{\text{fatigue level,UC}}$=2.04, SD=0.70) than participants in the NC ($M_{\text{fatigue level,NC}}$=1.01, SD=0.05) when averaging over the entire drive, with a large effect size ($t(19.193)=6.577$, $p<0.001$, $d=2.053$).

Figure 7-24 (only for NC) and Figure 7-25 (only for UC) give a more detailed insight into the development of fatigue and display the number of participants at each fatigue level for each minute of the drive. In the UC, eleven participants (55 %) reached FL3 or FL4 at least once during the experimental drive, while this was not the case for any of the participants in the NC (0 %). The Chi-squared test indicates a significant association between activity condition and the probability of reaching FL3 or FL4 with a large effect size ($\chi^2(1, N=39)=14.555$, $p<0.001$, $\varphi=0.611$).

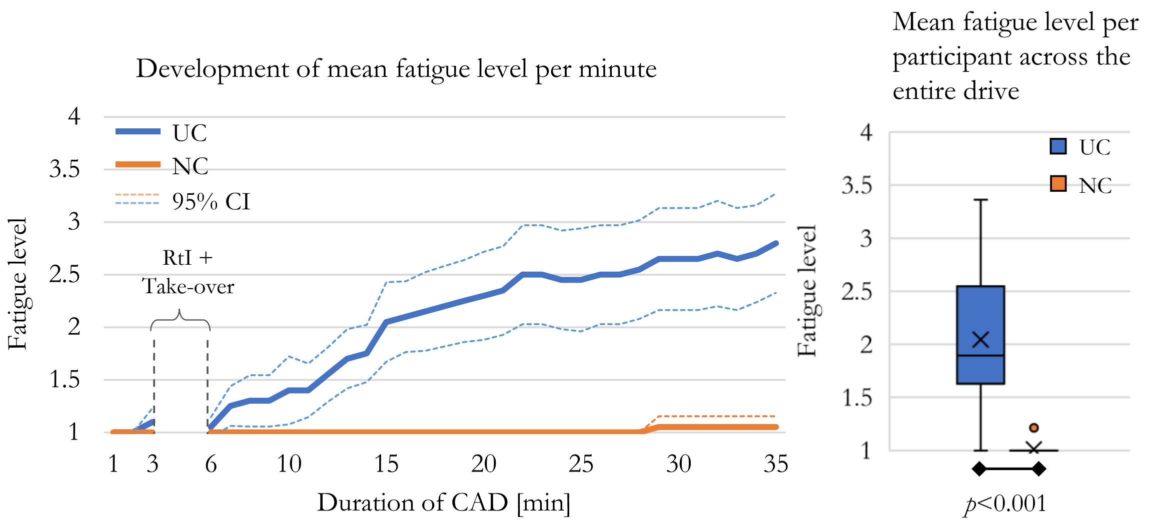


Figure 7-22. Left: Mean fatigue level and corresponding upper/lower 95 % CI of each minute of the drive depending on the activity condition. Right: Boxplot of mean fatigue level per participant averaged over the entire drive depending on the activity condition.

Figure 7-25 shows that the point in time when reaching FL3 or FL4 was highly individual in the UC, but with a general trend towards a higher probability to reach one of the high fatigue levels with elapsing time. On average, participants in the UC reached FL3 or FL4 after 19 minutes of CAD (*SD*=7, Min=8, Max=35) and remained at these high levels, on average, for 13 minutes (*SD*=9). Some individuals remained at FL3 or FL4 until the end of the drive once they reached it, others fluctuated between FL2, FL3 and FL4. As a result, not all participants had their fatigue maximum at the end of the drive right before the second RtI.

Only one of 19 participants (5 %) in the NC reached FL2 (in minute 29 of the drive), while the remaining participants (95 %) stayed at FL1 during the entire drive (cf. Figure 7-24). In the UC, only two of the 20 participants (10 %) remained at FL1 during the entire drive. The Chi-squared test indicates a significant association between activity condition and the probability of remaining at FL1 with a large effect size ($\chi^2(1, N=39)=28.003$, $p<0.001$, $\varphi=0.847$). Eight-teen of 20 participants (90 %) in the UC were at FL1 and two of 20 (10 %) participants were at FL2 prior to the first RtI (after three minutes) (cf. Figure 7-25). In the NC, all 19 participants (100 %) were at FL1 prior to the first RtI (cf. Figure 7-24). Prior to the second RtI (after a total driving time of 35 minutes), 18 of 19 participants (95 %) in the NC were at FL1 and one participant (5 %) at FL2. In the UC, two of 20 participants (10 %) were at FL1, eight (40 %) at FL2, two (10 %) at FL3 and eight (40 %) at FL4.

When analyzing participants' ratings on the KSS, the Mann-Whitney U-test shows that participants in the UC ($M_{KSS,UC}=7.1$, $SD=1.832$) rated themselves at a significantly higher fatigue level than participants in the NC ($M_{KSS,NC}=5.1$, $SD=1.889$), with a large effect size ($U=319.5$, $p=0.001$, $r_B=0.598$) (cf. Figure 7-23, left). The Mann-Whitney U-test was used, since data are ordinal-scaled. A statistical analysis of the effect of the fatigue level on the KSS rating does not make sense, since the sample sizes of some of the groups are very small (above all FL3 with N=2) and very different (cf. Figure 7-23, N_{FL1}/N_{FL3}=1/10). A descriptive examination of Figure 7-23 indicates a slightly rising KSS rating with increasing fatigue level. Especially the difference between FL1 ($M_{KSS,L1}$=5.00, SD=1.67) and FL4 ($M_{KSS,L4}$=7.75, SD=1.85) appears to be rather large ($M_{KSS,L2}$=7.22, SD=1.40; $M_{KSS,L3}$=7.00, SD=1.00).

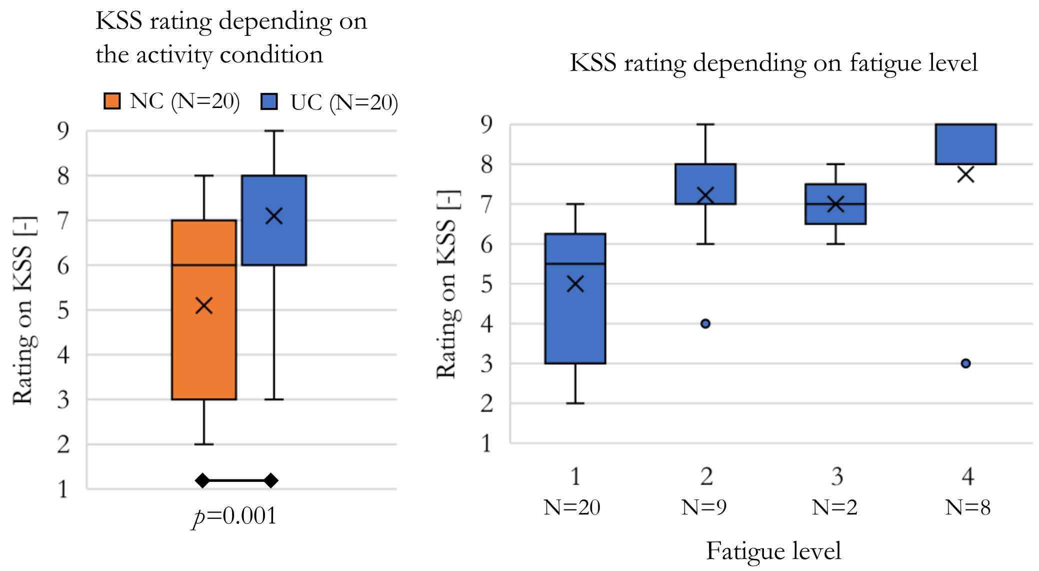


Figure 7-23. Boxplots of participants' KSS rating. Left: depending on the activity condition; right: depending on the fatigue level.

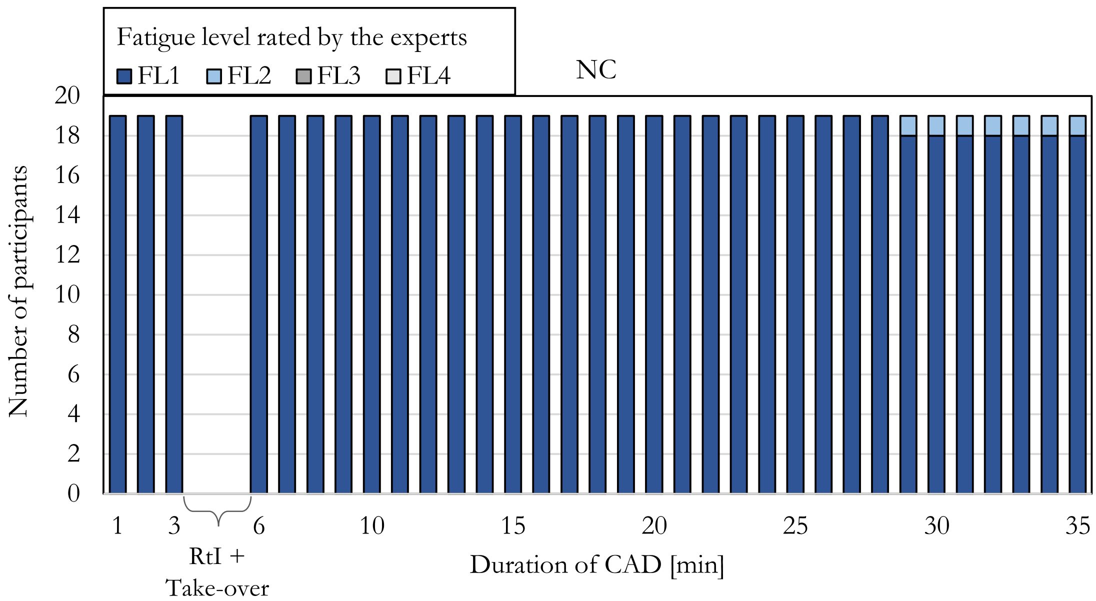


Figure 7-24. Number of participants being in one of the four fatigue levels per minute of the drive for the NC.

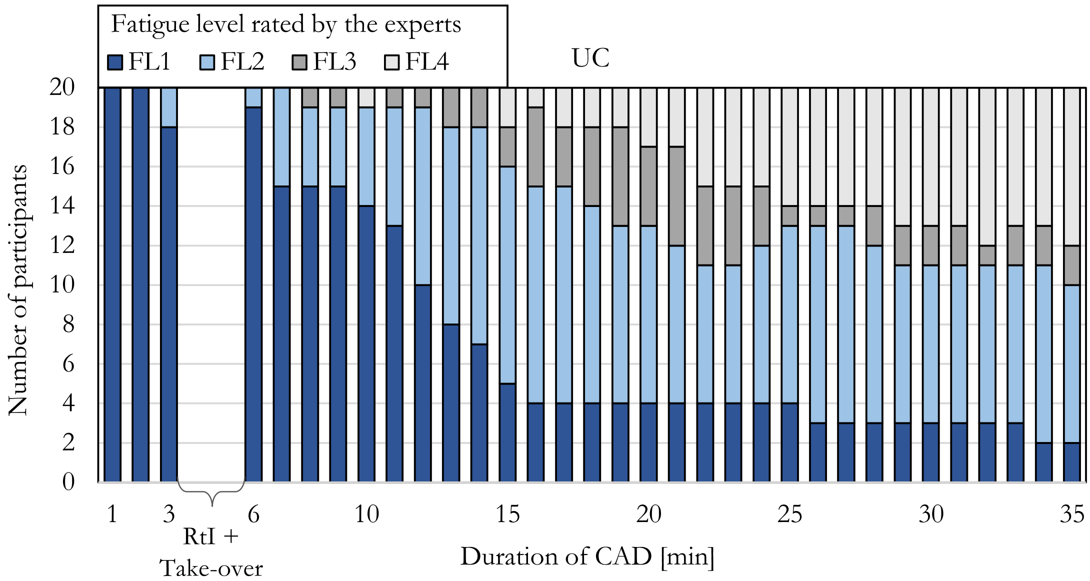


Figure 7-25. Number of participants being in one of the four fatigue levels per minute of the drive for the UC.

7.2.3.2 Effect of the Activity Condition on Take-over Performance

The study was designed based on the hypothesis that the manipulation by means of different activity conditions depends on a temporal factor, meaning that a potential difference between the conditions would become apparent only after a certain time has elapsed. The take-over event after three minutes of CAD was implemented to be able to additionally examine the effect of driving duration and interaction effects; however, a manipulation by means of the activity condition was not expected for this take-over. For this reason, only the performance of the second take-over was compared (tagged by CAD35) between the activity conditions to analyze their effect. All participants of the NC paused playing Tetris as soon as the second RtM sounded and were, thus, not engaged in an activity when the second RtI was issued.

Two-tailed Student's t-tests for independent samples were conducted to compare means of TOT_CAD35, AccLong_CAD35, AccLat_CAD35 and TTC_CAD35 in the two activity conditions (UC and NC). In case of a violation of the normality assumption, the non-parametric Mann-Whitney U-test was performed. In case of a violation of the variance homogeneity, the Welch's t-test was performed. When both assumptions were violated, the Welch's t-test was chosen over the Mann-Whitney U-test as proposed by Rasch et al. (2011) and Ruxton (2006). Effect size was given by the rank biserial correlation r_B for the Mann-Whitney U-test, for the other tests by Cohen's *d*. Fisher's exact tests were calculated instead of Chi-squared tests to investigate the existence of an association between the activity conditions and InRe_CAD35, FinRe_CAD35 and MC_CAD35 (bi- or multinomial data), since all variables had expected frequencies smaller than five. For crash_CAD35, no statistical test was calculated because there was no variance in values (see Figure 7-28, left). Effect size was either calculated by the phi coefficient (φ) (2x2 contingency table) or by Cramér's *V* (variables other than dichotomous).

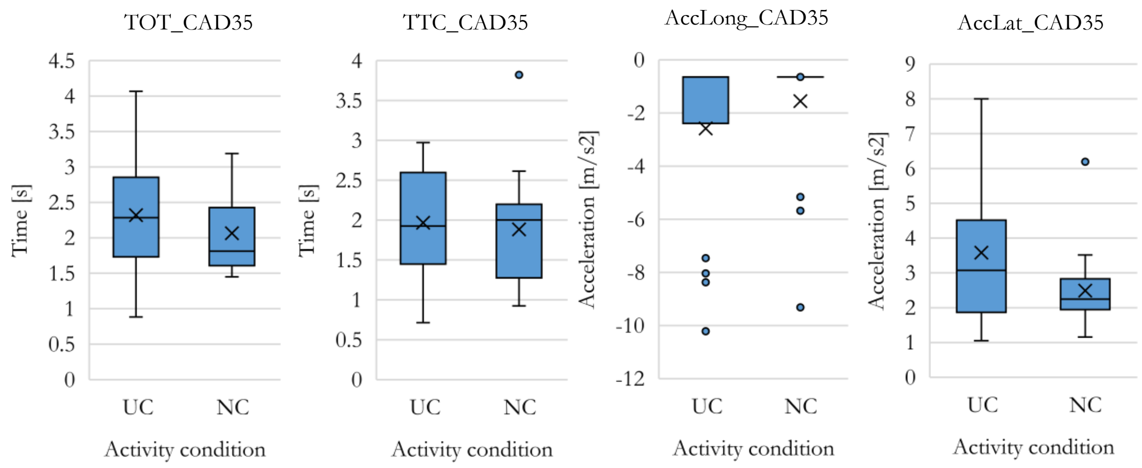


Figure 7-26. Boxplots of all metrical dependent variables of the take-over performance after an automation duration of 35 minutes depending on the activity condition.

Figure 7-26 shows the boxplots of the take-over performance variables TOT_CAD35, TTC_CAD35, AccLong_CAD35 and AccLat_CAD35 depending on the activity condition, and Table 18 the corresponding descriptive data. Figure 7-27 and Figure 7-28 display the bi- and multinomial take-over performance variables InRe_CAD35, FinRe_CAD35, crash (CAD35) and MC_CAD35 depending on the activity condition.

Results of the statistical analysis of all metrical take-over performance data (see Table 19 and Figure 7-26) did not show any significant differences between the activity conditions. There was a tendency of AccLat_CAD35 being higher for participants in the UC than participants in the NC ($\Delta M_{\mathrm{AccLat\_CAD35,NC\text{-}UC}}$=1.09 m/s^2) with a medium effect size (t(28.088)=2.039, p=0.051, d=0.645). Furthermore, there was no significant association between activity condition and any of the bi- or multinomial take-over parameters (InRe_CAD35, FinRe_CAD35 and MC_CAD35) (cf. Table 20).

Table 18. Descriptive data of all metrical dependent variables of take-over performance after 35 minutes of CAD depending on the activity condition.

	Condition	*N*	Mean	*SD*	Min	Max
TOT_CAD35 [s]	UC	20	2.32	0.82	0.88	4.07
	NC	20	2.07	0.57	1.45	3.19
TTC_CAD35 [s]	UC	20	1.97	0.70	0.71	2.97
	NC	20	1.88	0.69	0.92	3.82
AccLong_CAD35 [m/s²]	UC	20	−2.58	3.48	−10.22	−0.64
	NC	20	−1.55	2.34	−9.32	−0.64
AccLat_CAD35 [m/s²]	UC	20	3.58	2.14	1.05	8.00
	NC	20	2.49	1.08	1.15	6.19

Table 19. Results of mean comparisons of the metrical dependent variables after 35 minutes of CAD between the activity conditions UC and NC.

	Test	Statistic $t\,(U)$	df	p	Effect size $d\,(r_B)$
TOT	Mann-Whitney	238.500	-	0.304	0.193
TTC	Student	0.375	38	0.709	0.119
AccLong	Welch	−1.098	33.320	0.280	−0.347
AccLat	Welch	2.039	28.088	**0.051**	0.645

Note: For the Mann-Whitney U-test, test statistic is given by U-value and effect size by the rank biserial correlation r_B.

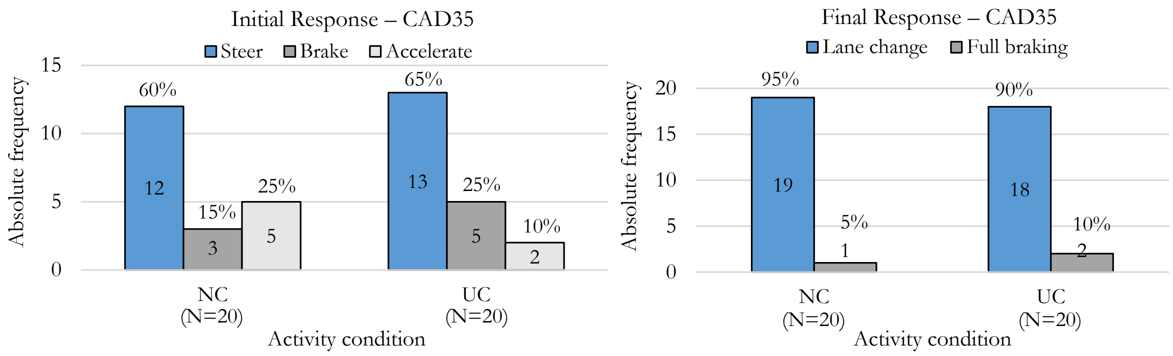


Figure 7-27. Frequency of initial response types (steer, brake, accelerate) and final response types (lane change, full braking) after 35 minutes of CAD depending on the activity condition.

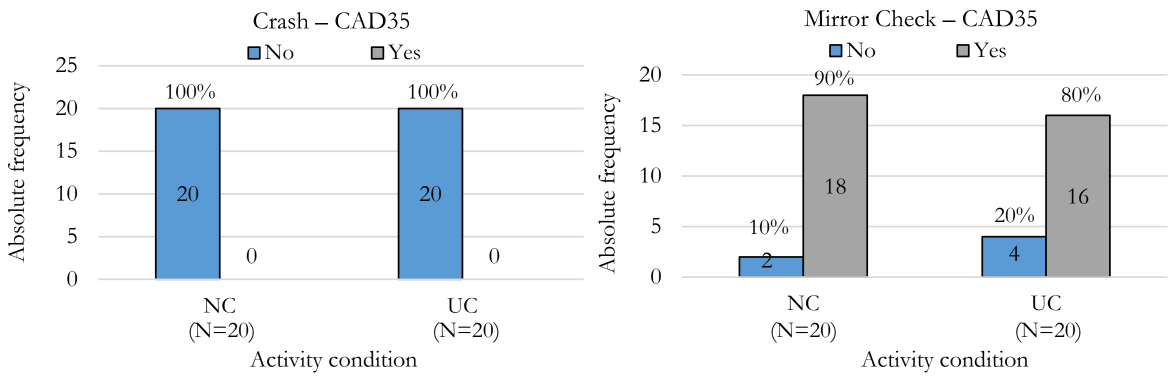


Figure 7-28. Frequency of crash and mirror check after 35 minutes of CAD depending on the activity condition.

Table 20. Results of Chi-squared tests to compare the frequencies of the bi- or multinomial dependent variables of take-over performance (InRe, FinRe, crash and MC) after 35 minutes of CAD between the activity conditions UC and NC.

	Test	Statistic χ^2	*N*	*df*	*p*	Effect size φ/V
InRe_CAD35	Fisher's exact	1.826	40	2	0.401	0.214
FinRe_CAD35	Fisher's exact	0.360	40	1	1.000	−0.095
Crash_CAD35[a]	-	-	-	-	-	-
MC_CAD35	Fisher's exact	0.784	40	1	0.661	0.140

Note: In case that two cells had an expected frequency smaller than five, Fisher's exact test was conducted. For 2x2 contingency tables, effect size is given by phi coefficient φ. For other contingency tables by Cramér's *V*.

[a] No statistical test was calculated, since there was no variance in values.

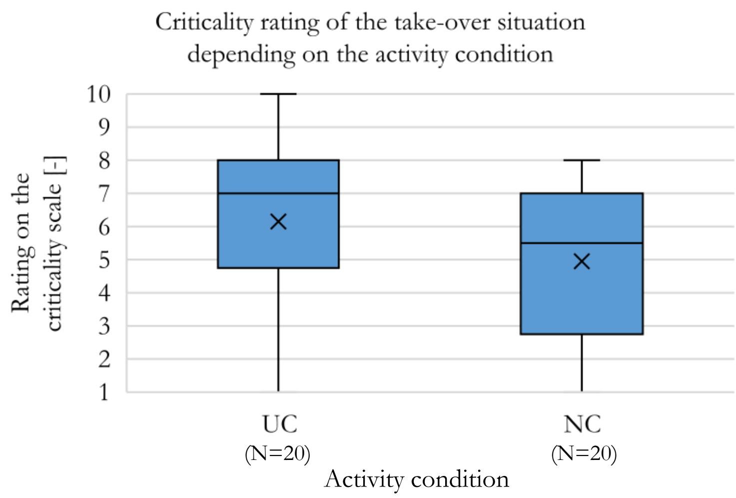


Figure 7-29. Boxplot of the criticality rating of the second take-over situation after 35 minutes of CAD depending on the activity condition.

Figure 7-29 displays the boxplot of the perceived criticality of the second take-over situation depending on the activity condition ($M_{\text{criticality rating,UC}}$=6.15, *SD*=2.52; $M_{\text{criticality rating,NC}}$=4.95, *SD*=2.44). Statistical differences between the activity conditions were tested using a Mann-Whitney U-test due to the ordinal-scaled dependent variable. Effect size is given by the rank biserial correlation r_B. Results reveal that there is no significant difference in the perceived criticality between the activity conditions (*U*=254.500, *p*=0.140, r_B=0.273).

7.2.3.3 Effect of Driving Duration and Activity Condition on Take-over Performance

All participants of the NC paused playing Tetris both times as soon as the RtM sounded and were, thus, not engaged in an activity when the first and the second RtI were issued.

Figure 7-30 shows the boxplots of the take-over performance variables TOT, TTC, AccLong and AccLat depending on the activity condition (UC, NC) and the automation duration (CAD3, CAD35), and Table 21 the corresponding descriptive data. A mixed ANOVA with the between-subjects two-level factor activity condition (UC, NC) and the two-level within-subjects factor automation duration (CAD3, CAD35) was conducted for each of the metrics TOT, TTC, AccLong and AccLat. The assumption of normal distribution was violated since Shapiro-Wilk test was significant for one group of TOT, all groups of AccLong, and for three groups of AccLat. However, this assumption violation was not considered, since the mixed ANOVA was proved to be relatively robust against non-normal distribution (Bühner & Ziegler, 2017, p. 368; Glass, Peckham, & Sanders, 2016; Harwell, Rubinstein, Hayes, & Olds, 2016; Salkind, 2010). Homogeneity of variances was assessed by Levene's test, which was significant for AccLat_CAD35 and for AccLong_CAD3. Hartley's F_{max} test, however, indicates a variance ratio smaller than ten for both metrics ($F_{max,AccLong\_CAD3}=2.2$; $F_{max,AccLat\_CAD35}=3.9$). According to Bühner and Ziegler (2017, p. 369), this means that the actual α-error probability is not increased and the mixed ANOVA can be interpreted without α-correction. Effect size is given by partial eta-squared η_p^2.

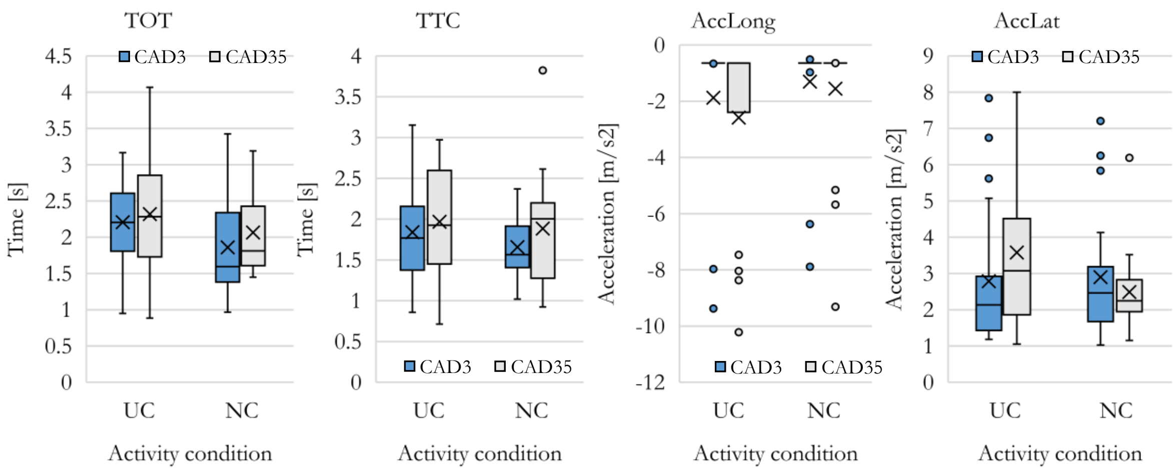


Figure 7-30. Boxplots of all metrical dependent variables of take-over performance depending on the activity condition and automation duration.

Boxplots suggest greater TOT and greater TTC for CAD35 when comparing to CAD3, and greater TOT and greater TTC for the UC when comparing to NC (see Figure 7-30 and Table 21). However, results of the mixed ANOVA indicate no significant main effects of activity condition and automation duration on TOT and TTC, even though there is a tendency of TOT to be greater for UC and a tendency of TTC to be greater for CAD35 (cf. Table 22). Additionally, no significant interaction effect between the two factors on TOT and TTC was found (see Table 22). Furthermore, the boxplots of AccLat and AccLong suggest higher accelerations for the subgroup UC CAD35. However, the result of the mixed ANOVA does not indicate a significant interaction effect between activity condition and automation duration on both metrics (see Table 22). Additionally, the main effects of both factors are not significant (cf. Table 22), even though the boxplot of AccLong indicates slightly greater accelerations for UC and for CAD3.

Table 21. Descriptive data of all metrical dependent variables of take-over performance depending on the activity condition and automation duration.

		Automation duration (within factor)					
		CAD3			CAD35		
	Activity condition (between factor)	***M* (*SD*)**	**Min**	**Max**	***M* (*SD*)**	**Min**	**Max**
TOT [s]	UC	2.21 (0.55)	0.95	3.17	2.32 (0.80)	0.88	4.07
	NC	1.86 (0.65)	0.97	3.43	2.07 (0.56)	1.45	3.19
TTC [s]	UC	1.84 (0.61)	0.86	3.15	1.97 (0.69)	0.71	2.97
	NC	1.66 (0.40)	1.02	2.37	1.88 (0.67)	0.92	3.82
AccLong [m/s²]	UC	−1.87 (2.95)	−9.38	−0.53	−2.58 (3.39)	−10.22	−0.64
	NC	−1.30 (1.96)	−7.89	−0.51	−1.55 (2.28)	−9.32	−0.64
AccLat [m/s²]	UC	2.79 (1.91)	1.18	7.84	3.58 (2.08)	1.05	8.00
	NC	2.90 (1.69)	1.03	7.20	2.49 (1.05)	1.15	6.19

Table 22. Results of the mixed ANOVA (activity condition x automation duration) for TOT, TTC, AccLong and AccLat.

	Effect	$df_{numerator}$	$df_{denumerator}$	F	p	η_p^2
TOT	Activity condition	1	38	3.150	0.084	0.077
	Automation duration	1	38	1.635	0.209	0.041
	Activity condition * automation duration	1	38	0.138	0.712	0.004
TTC	Activity condition	1	38	0.631	0.432	0.016
	Automation duration	1	38	2.975	0.093	0.073
	Activity condition * automation duration	1	38	0.235	0.631	0.006
AccLong	Activity condition	1	38	1.038	0.315	0.027
	Automation duration	1	38	1.512	0.226	0.038
	Activity condition * automation duration	1	38	0.341	0.563	0.009
AccLat	Activity condition	1	38	1.284	0.264	0.033
	Automation duration	1	38	0.287	0.595	0.007
	Activity condition * automation duration	1	38	2.837	0.100	0.069

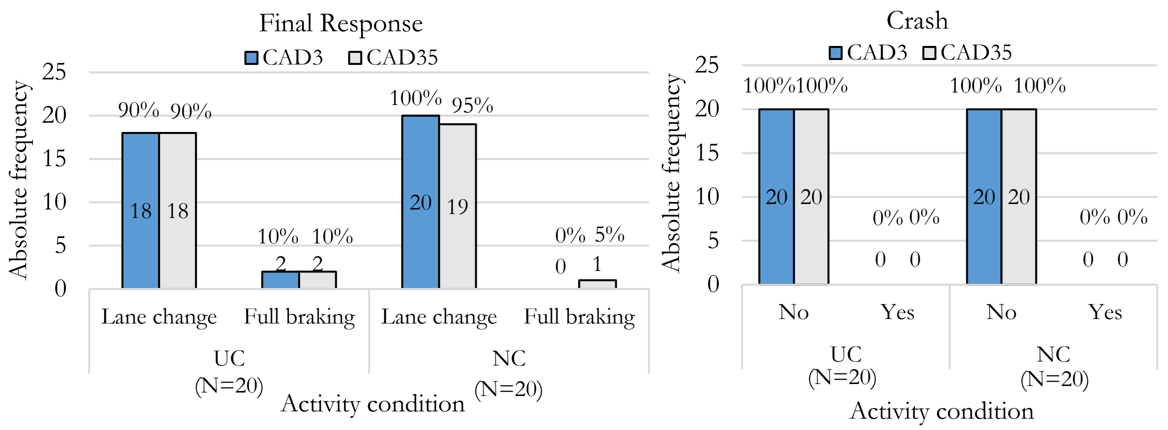


Figure 7-31. Frequency of initial response types (steer, brake, accelerate) and mirror check depending on the activity condition and automation duration.

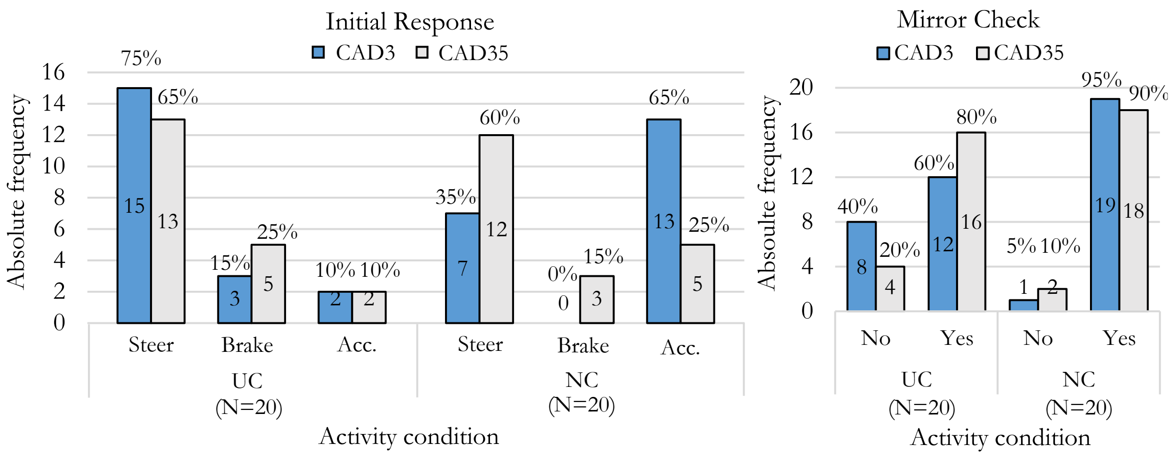


Figure 7-32. Frequency of final response types (lane change, full braking) and crash depending on the activity condition and automation duration.

Figure 7-31 and Figure 7-32 display the bi- and multinomial take-over performance variables InRe, FinRe, crash and MC depending on activity condition and automation duration. None of the participants produced a collision in one of the take-over events. For more than 90 % of the participants, FinRe was a lane change in both activity conditions and after both automation durations (cf. Figure 7-32, left). For InRe, it is remarkable that a large proportion of 65 % of participants in the NC accelerated after three minutes of CAD, whereas after 35 minutes a large proportion of 60 % initially steered (cf. Figure 7-32, left). This is not visible for the UC, which suggests that there might be an interaction effect between the factors of automation duration and activity condition. Furthermore, it is noticeable that 90 % or more participants of the NC conducted a mirror check for both automation durations (CAD3 and CAD35), whereas for the UC the picture is not that clear: In CAD3, 60 % of the participants conducted a mirror check, whereas in CAD35, the proportion was 80 % (cf. Figure 7-32, right).

7.2.3.4 Evaluation of PERCLOS and Comparison to Fatigue Rating

Thirty-three minutes of the 35-minute experimental drive (three minutes before the first RtI, 30 minutes before the second RtI) were considered for PERCLOS calculation. Two minutes, counted from the first RtI onwards, were left out of the PERCLOS analysis because the take-over situation and process as well as a short manual drive and the reactivation of the CA took place during this phase. This resulted in a maximum of 660 PERCLOS data points in the UC (20 participants x 33 minutes of CAD) and 627 PERCLOS data points in the NC (19 participants x 33 minutes of CAD) theoretically available from this experiment. However, for quality reasons the data processing described in chapter 7.1.2 results in a reduced PERCLOS availability for both activity conditions. In the UC, 607 of 660 (92 %) PERCLOS values were available and, in the NC, 223 of 627 (36 %). The unavailability of PERCLOS values led to a strongly varying sample size when calculating the mean PERCLOS per minute across all participants of the respective activity condition: the sample size in the UC was between 17 and 20 (of a maximum of 20 PERCLOS values per minute) and between four and 17 (of a maximum of 19 PERCLOS values per minute) in the NC (see Figure 7-33). Figure 7-34 displays the frequency distribution of PERCLOS availability per minute: for the UC, the PERCLOS availability was 80 % or higher in 100 % of the minutes (equal to 33 minutes), while for the NC, in only 3 % (equal to one minute) of the evaluated 33 minutes the PERCLOS availability was 80 % or higher. Furthermore, for the NC, the PERCLOS availability was 50 % or less in the majority of the minutes (79 % or 26 minutes) of the drive, and in 9 % of the minutes (equal to three minutes) the availability was less than 30 %.

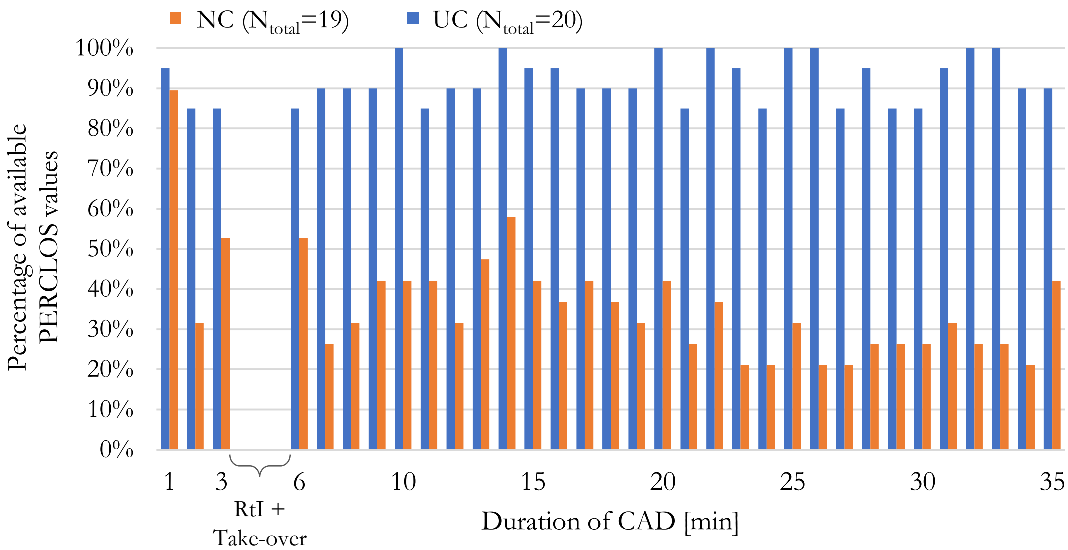


Figure 7-33. Percentage of available PERCLOS values for each minute of the drive relative to the maximum available PERCLOS values depending on the activity condition.

Figure 7-35 (left) displays the mean PERCLOS and 95 % confidence interval using the available PERCLOS values for each minute of the drive depending on the activity condition. It is noticeable that the mean PERCLOS of the NC is at a low level during the entire drive, while the mean PERCLOS of the UC steadily rises, indicating increasing fatigue with elapsing time. This development of PERCLOS during the drive corresponds to the development of fatigue level rated by the experts in both activity conditions: the trend lines of the mean fatigue level and the mean PERCLOS of the UC (cf. Figure 7-36) have a positive gradient, indicating increasing fatigue for both metrics with elapsing time. When making the same comparison for the NC (cf. Figure 7-37), the progression of the trend line of PERCLOS and the fatigue level coincide on an almost constant level during the entire drive (with a slightly positive gradient for PERCLOS). This result is also reflected in the statistical analysis of PERCLOS when averaging over the entire drive and comparing the activity conditions. Since the Shapiro-Wilk test and Levene's test are significant, Welch's t-test is chosen over the Mann-Whitney U-test as proposed by Rasch et al. (2011) and Ruxton (2006). As was already the result for the mean fatigue level rated by the experts (cf. Figure 7-22 and subchapter 7.2.3.1), mean PERCLOS was significantly higher in the UC than in the NC with a medium effect size (see Figure 7-35 right; $M_{\mathrm{PERCLOS,UC}}$=5.96 %, SD=10.4; $M_{\mathrm{PERCLOS,NC}}$=0.92 %, SD=1.2; t(19.515)=2.161, p=0.043, d=0.684).

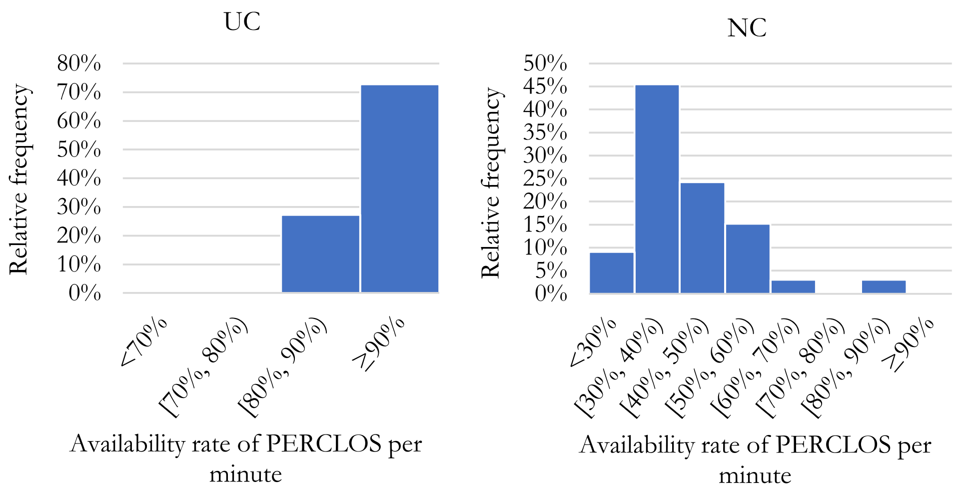


Figure 7-34. Frequency distribution of PERCLOS availability rate per minute depending on the activity condition.

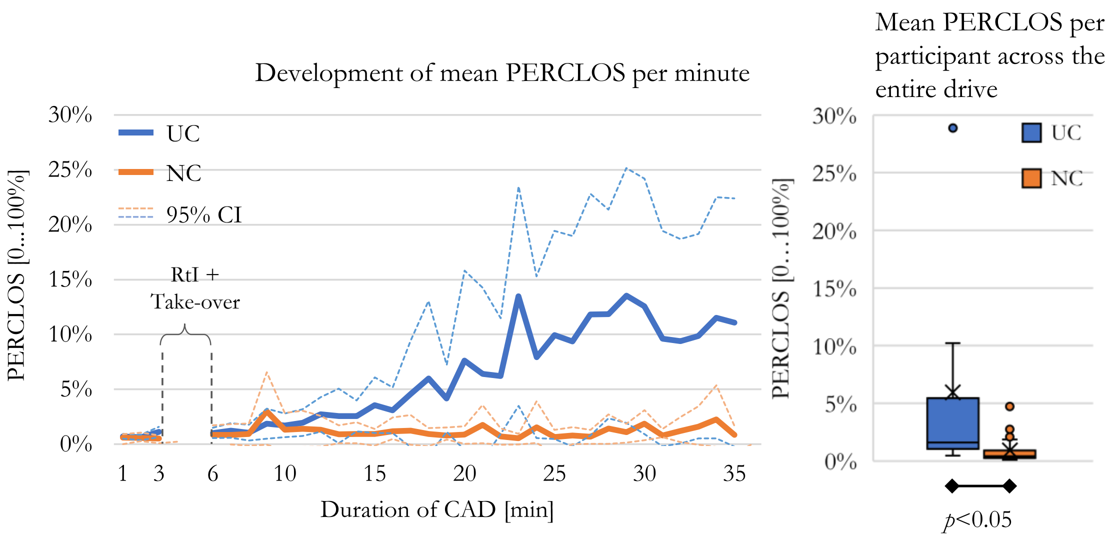


Figure 7-35. Left: Development of mean PERCLOS and 95 % CI for each minute of the drive depending on the activity condition. Right: Boxplot of PERCLOS averaged over the entire drive per participant depending on the activity condition. Note: Not all outliers in the UC are displayed.

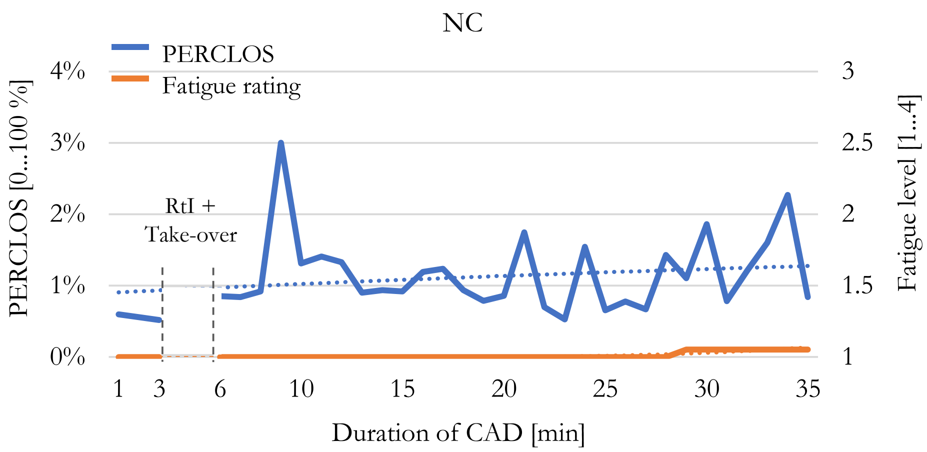


Figure 7-36. Mean PERCLOS and mean fatigue level averaged over all participants of the NC per minute of the drive with a polynomial trend line (each of degree two).

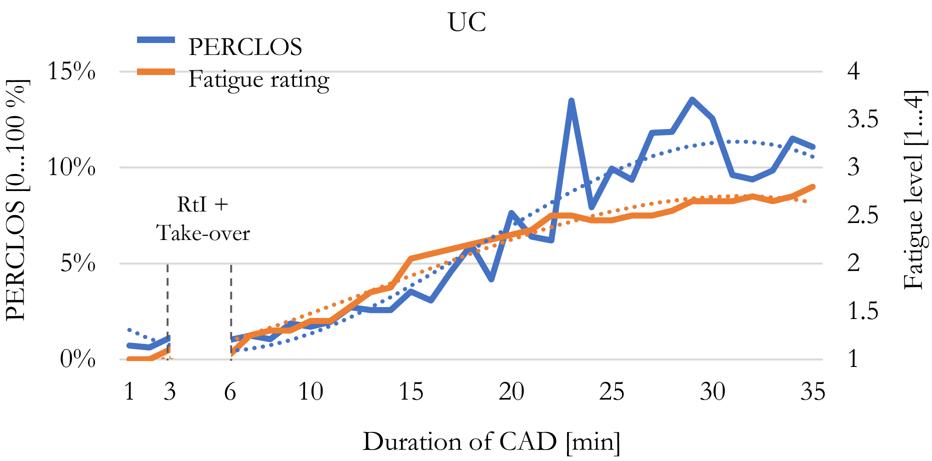


Figure 7-37. Mean PERCLOS and mean fatigue level averaged over all participants of the UC per minute of the drive with a polynomial trend line (each of degree three).

More insights into the evaluation of validity of PERCLOS are provided by Figure 7-38, which shows the boxplot of all available PERCLOS values depending on the fatigue level and the activity condition. Table 23 provides selected descriptive data corresponding to the boxplots. The selected parameters are identical to the ones defined in experiment 1 (see chapter 7.1.3.4).

In the UC, a clear increase of M_{PERCLOS} and Mdn_{PERCLOS} is noted for increasing fatigue levels: M_{PERCLOS} of FL2 is 2.5 times greater than M_{PERCLOS} of FL1 ($Mdn_{\text{PERCLOS,FL3}} = 1.9\text{x}Mdn_{\text{PERCLOS,FL2}}$), M_{PERCLOS} of FL3 is three times greater than M_{PERCLOS} of FL2 ($Mdn_{\text{PERCLOS,FL3}} = 3.4\text{x}Mdn_{\text{PERCLOS,FL2}}$), and M_{PERCLOS} of FL4 is four times greater than M_{PERCLOS} of FL3 ($Mdn_{\text{PERCLOS,FL4}} = 3\text{x}Mdn_{\text{PERCLOS,FL3}}$). For the NC, this evaluation is not possible, since only PERCLOS data of participants in FL1 is available (apart from two data points in FL2). h_{outliers} is relatively high at FL1 and FL2 in the UC and at FL1 in the NC and it is at a similar level between 19 % and 25 %. IQR_{PERCLOS} for FL1 is 1.5 times smaller in the UC than in the NC. Furthermore, $h_{\text{outliers7\%}}$ (as defined in subchapter 7.1.3.4) at FL1 and FL2 (only for UC) is similarly low in both activity conditions ($h_{\text{outliers7\%,UC,FL1}} = 0\,\%$, $h_{\text{outliers7\%,UC,FL2}} = 6\,\%$; $h_{\text{outliers7\%,NC,FL1}} = 2\,\%$).

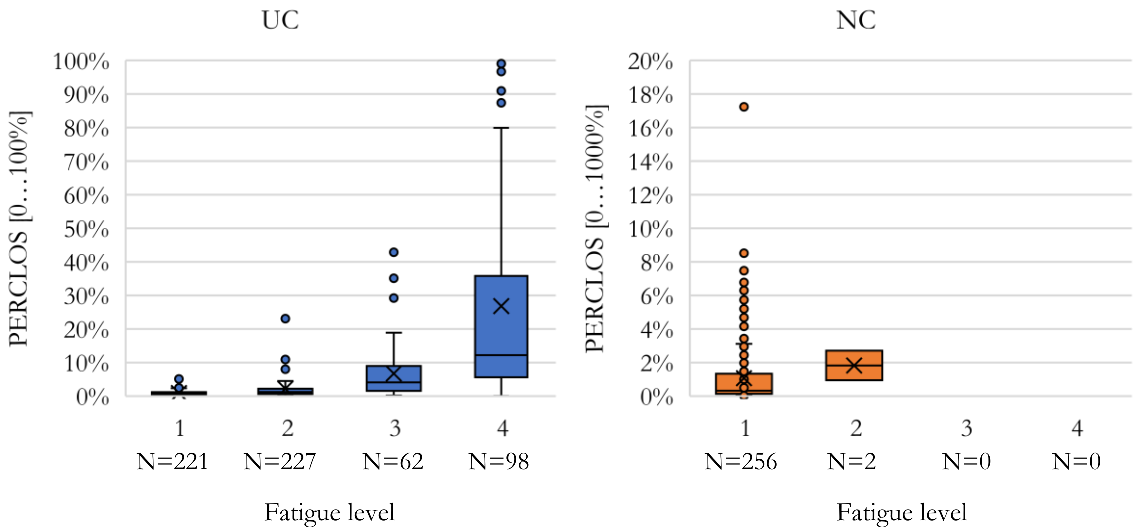


Figure 7-38. Boxplots of all available PERCLOS values depending on the fatigue level and on the activity condition.

Table 23. Descriptive data of all available PERCLOS values (independent of the participant) depending on the fatigue level and the activity condition.

Activity condition		Fatigue level			
		1	2	3	4
UC	$N_{\text{participants}}$	20	18	11	9
	N_{PERCLOS}	221	227	62	98
	M_{PERCLOS}	0.90%	2.22%	6.74%	26.84%
	Mdn_{PERCLOS}	0.65%	1.22%	4.10%	12.22%
	$\text{IQR}_{\text{PERCLOS}}$	0.79%	1.58%	7.40%	30.19%
	h_{outliers} (Q3+1.5xIQR)	5%	11%	15%	24%
	$h_{\text{outliers7\%}}$ (>7%)	0%	6%	34%	72%
NC	$N_{\text{participants}}$	20	1	0	0
	$N_{\text{available PERCLOS}}$	256	2	0	0
	M_{PERCLOS}	1.07%	1.83%	-	-
	Mdn_{PERCLOS}	0.32%	1.83%	-	-
	$\text{IQR}_{\text{PERCLOS}}$	1.19%	0.88%	-	-
	h_{outliers} (Q3+1.5xIQR)	8%	0%	-	-
	$h_{\text{outliers7\%}}$ (>7%)	2%	0%		

7.2.4 Discussion

In this study, all participants of the NC complied with the instructions and were engaged in the Tetris computer game during the automated drive. As expected from experiment 1 (see chapter 7.1.4) and many previous studies (cf. chapter 4.9 or e.g., Jarosch et al., 2017; Neubauer et al., 2014; Schömig et al., 2015; Weinbeer et al., 2019), this engagement in the NDRA caused the manipulation intended by the activity condition to be successful: the proportion of participants staying at FL1 during the entire drive was significantly lower in the NC, the proportion of participants reaching FL3 and FL4 was significantly higher in the UC, and the overall fatigue level was significantly lower for participants in the NC than in the UC when averaging over the entire drive. The successful manipulation by means of the activity condition is also reflected in the perceived fatigue of the participants assessed by the KSS, which was significantly higher for participants in the UC. The manipulation in this experiment was even more successful than in experiment 1 because none of the participants in the NC reached FL3 or FL4; in fact, they all—except one participant—remained at FL1 (non-fatigued) during the entire drive, which is also represented in the large effect size for all abovementioned statistical analyses. In the first place, this finding might convey that playing Tetris is a more effective countermeasure for fatigue than a free choice of NDRAs, which would disprove the hypothesis of experiment 1. However, according to the theories described in chapter 4.6 and discussed in chapter 7.1.4, fatigue vulnerability strongly depends on personality traits and their interaction with the task. Since it was not the same sample of participants who were tested in the two studies, and only less is known about the personality traits of the test groups relevant for the interaction with Tetris or other NDRAs, a generalized conclusion regarding the effectiveness of different NDRAs is not possible. The only conscious variation in terms of the participants' characteristics was the upper limitation of age, since one result of experiment 1 was that older participants showed a different way of handling the activities offered (Hecht et al., 2020), and that there is a potential correlation of fatigue vulnerability and age in the NC with elapsing time. It is conceivable that age may be one of the influencing factors on the effectiveness of NDRAs as a countermeasure for fatigue as it was suggested by Y. Wu et al. (2020). However, this cannot be conclusively clarified with the two present studies.

In this study, a general varying fatigue vulnerability in the UC was again found: about half of the participants (45 %) did not reach FL3 and FL4 within 35 minutes of CAD, which is almost identical to the proportion in experiment 1, despite a clearly shorter driving duration in experiment 2. This also resulted in a lower average driving duration until reaching FL3 and FL4 in experiment 2 than in experiment 1 (19 vs. 32 minutes). This result may indicate that participants of experiment 2 were potentially more susceptible to fatigue, emphasizing that the ability to resist fatigue depends on personal factors, also in the absence of NDRAs (see chapter 4.6). However, other factors may also have caused this result. For instance, varying daytimes, the instruction not to drink caffeinated beverages or the individual form on the day (independent of the personality traits) may have had an impact on the fatigue vulnerability, as well. One needs to consider, though, that these factors also apply to the NC, where the findings in terms of fatigue vulnerability were diametrically opposite. Observations regarding the fatigue development in the UC were very similar to those in experiment 1 (cf. chapter 7.1.4): many participants who stayed completely at FL2 or fluctuated between FL2, FL3 and FL4 exhibited mannerisms. It is assumed that mannerisms are part of a self-activating mechanism to compensate upcoming fatigue. The point in time of reaching FL3 and FL4 was again highly individual, ranging from eight to 35 minutes.

Previous studies found a significant effect of the automation duration on take-over performance. A prolonged automation duration compared to a short one, prevalently, caused a deterioration of rather automatic reactions such as gaze reaction time (Feldhütter et al., 2017) or hands-on-steering-wheel time and feet-on-pedal time (Bourrelly et al., 2019), but also an increase of the take-over time (Bourrelly et al., 2019). Reaction times of automatic reactions were not assessed in this study; hence, these results cannot be compared. However, none of the assessed take-over performance variables showed any significant difference between the automation durations either. Therefore, the findings of Bourrelly et al. (2019) could not be confirmed. One reason might be that the automation duration of 60 minutes in Bourrelly et al. (2019) was clearly longer than in the present study, and therefore, the fatigue of participants higher. Even though Bourrelly et al. (2019) assessed fatigue by a self-appraisal as well, no comparison of the proportion of fatigued participants between the experiments in the moment of the RtI is possible, since the scales used are too different.

The only peculiarity in terms of interaction effect between the activity condition and driving duration was revealed by the initial response: participants in the NC accelerated more frequently after a short automation duration than the same group after the prolonged automation duration. The proportion in the UC was on the same (low) level in both take-overs. This finding is difficult to interpret. Usually, accelerating in such a situation is not considered a proper reaction, meaning that participants reacted worse after a short duration than after a prolonged one. However, when looking at the remaining take-over parameters of the participants who accelerated, it becomes apparent that they all performed a lane change maneuver with relatively short TOTs and very low AccLong and AccLat. All these factors indicate a controlled and well-planned maneuver. Therefore, it is conceivable that these participants were completely back in the loop immediately after the RtI, and accelerating was a conscious decision in order not to lose speed when fulfilling the lane change maneuver. This potentially good situation awareness may be attributed to the RtM that participants of the NC received prior to the RtI. It may have two reasons why participants did not show the same reaction in the second take-over: First, the 15 seconds of the RtM were not sufficient anymore to build up sufficient situation awareness after a longer automation duration. Second, participants learned that accelerating is not necessarily required in this scenario. Since all participants again showed a completely controlled maneuver in the second take-over, the latter is the more likely explanation. For the remaining take-over performance variables (TOT, TTC, AccLong and AccLat), the interaction effect between the activity condition and driving duration supposed beforehand was not found. Taking the results of experiment 1 into account, this finding is surprising: while some participants of the NC in experiment 1 reached FL3 or FL4 despite a NDRA, this was not the case for the participants with a NDRA in experiment 2. Due to a perfect fatigue manipulation in the NC in experiment 2 (complete absence of FL3 and FL4) and a considerable proportion of participants at FL3 and FL4 in the UC, a deterioration of the take-over performance of participants in the UC could have been expected when comparing the two activity conditions. However, this was not the case. This might be attributed to the fact that only half of the participants in the UC was at FL3 or FL4 when the RtI sounded, as was already the case in experiment 1. The proportion of participants in the UC who were in a fatigued state at the moment of the RtI was potentially too low to cause a significant difference. This also indicates, however, that—if there is a statistically relevant effect of fatigue on take-over performance—this effect is not a large one, since half of the participants in the UC was in the intended fatigue state anyway. This would also explain that there was no significant interaction effect between automation duration and activity condition. A separate additional analysis of take-over performance of CAD35 based on the fatigue level as it was conducted in experiment 1 did not make sense in this study, since the sample size of participants at FL3 and FL4 was too small.

Contrary to experiment 1, there was no difference between activity conditions in the AccLong in this study. Apparently, the approach with the RtM 15 seconds prior to the RtI in the NC was effective for participants to build up sufficient situation awareness for the take-over. As a result, participants did not perform a strong braking maneuver. This result confirms the effectiveness of the concept of van den Beukel (2016) and Large et al. (2017) and emphasizes the relevance of both a sufficient time budget to take over control and a suitably designed HMI-concept for the transition phase.

The evaluation of PERCLOS revealed very similar findings as in experiment 1 regarding the data availability issue with a strong difference between the activity conditions: while in the UC, the overall data availability of about 90 % was approximately the same in both experiments, the data availability of less than 40 % in the NC was even worse in this study than in experiment 1. The consideration of data availability over the course of the experimental drive indicates the same result. In the UC, the data availability per minute never dropped below 80 %, whereas in the NC, the majority of the driving time data availability was less than 50 %. This was probably caused by the fixed position of the tablet in the middle console, on which participants played Tetris continuously. The analysis of recorded videos revealed that most participants turned their upper bodies towards the middle console and strongly inclined their heads towards the tablet. This body posture was obviously even more adverse for the eye-tracking system than an engagement in arbitrary NDRAs, since the data availability was even worse in this study than in experiment 1. This result is not as expected because one of the three infrared cameras was intentionally mounted in the middle console to cover this type of posture.

The comparison of PERCLOS with the fatigue rating showed a strong conformity in this study for both activity conditions: the significant difference between activity conditions found for the overall mean fatigue level (over all participants and all minutes) was also ascertained for the overall mean PERCLOS. Furthermore, the mean PERCLOS per one-minute increments each over all participants evolved uniformly when compared to the mean fatigue level over the driving time of 35 minutes. However, at least in the NC, the reliability of this result is very limited, since the sample size for calculating the mean PERCLOS per minute was reduced by more than 50 % in most cases. The additional analysis of PERCLOS depending on the fatigue level shows similar plausible findings for the UC as in experiment 1. The PERCLOS median and mean steadily rises with increasing fatigue level. Again, PERCLOS mean and median at FL4 are comparable to literature reviews (cf. chapter 7.1.4), while PERCLOS mean and median at FL3 do not indicate fatigue as distinctly. The downward outliers found in experiment 1 for FL3 and FL4 manifest in this study, as well. Their potential origin has already been traced back on the different natural boundaries of the PERCLOS metric and the natural boundaries of the eye-tracking system, which has been described in chapter 7.1.4. For the NC, no conclusion can be drawn for the validity of PERCLOS for FL2, FL3 or FL4, since no or only a few data points are available for analysis in this study. For FL1, it is noted, however, that PERCLOS mean and median are on a low level. Additionally, the number of false positives (PERCLOS indicating increased fatigue at FL1 and FL2) is not even half the size as in experiment 1. This apparent increase of PERCLOS validity in the NC compared to experiment 1 might be caused by the following circumstance observed when analyzing the videos: when participants were completely involved in the NDRA causing them to close their eyes more because they were looking down at the tablet, there was no eyelid detection at all. As soon as the participants looked at the road, for instance, to briefly check the environment, their eyelid was detected, but in that case, their eyes were opened normally. This would explain the greater consistency with FL1 but also the rather poor data availability.

Limitation of the Study and Conclusion

One of the main findings of this study is that NDRAs proved to be effective as a countermeasure for fatigue. When comparing results of experiment 1 and 2, Tetris was even more effective than a free choice of NDRAs. However, playing Tetris is probably not generally the better countermeasure for fatigue compared to a free choice of NDRAs, but it was for these specific (young to middle-aged) individuals in this study. Also, without a NDRA slight differences in fatigue vulnerability became apparent between the two experiments. Age was hypothesized to be one possible reason for these differences; however, this factor was not explicitly examined. Future work should make an effort to identify the personal factors relevant for fatigue vulnerability, taking the presence and absence of NDRAs into account.

Even though the manipulation by activity condition worked very well in terms of a low fatigue level in the NC, it was not successful to the same extent in the UC, since not all participants in the UC reached FL3 and FL4. Therefore, the sample sizes for FL3 and FL4 were too low in this study to analyze take-over performance depending on the fatigue level. However, similar to experiment 1, this makes a clear conclusion impossible regarding the effect of fatigue on take-over performance. As already concluded from experiment 1, to avoid this issue in future studies and to be able to examine the effect of fatigue on take-over performance more effectively, an experimental design is necessary in which the RtI and take-over event are initiated as soon as the critical fatigue level is reached and not after a fixed duration of CAD. To realize this approach in a subsequent study, an appropriate method needs to be designed.

The value of KSS for fatigue assessment was only evaluable descriptively due to the small sample size for FL3 and FL4, which showed a plausible correlation, though. A significant difference was found between the activity conditions, indicating that playing Tetris had a positive effect on the perceived fatigue. Nevertheless, the validity of KSS applied in that way—retrospective assessment for the moment right before the second RtI—has to be treated with caution, since the participants' memory may be biased by the take-over event experienced and the time elapsed in the meantime.

A remote eye tracking system showed its weaknesses when drivers engage in a NDRA, which caused a low data availability for PERCLOS in the NC. Contrary to experiment 1, the validity of PERCLOS in the condition with a NDRA appeared to be much better in experiment 2. However, the reliability of the result has to be treated with caution due to the very low data availability. In the UC, results of experiment 1 could be confirmed: PERCLOS showed a good face validity and a good conformity with literature reviews. Nevertheless, PERCLOS evaluation in experiment 1 and experiment 2 showed that fatigue assessment should not only be based on only one metric in order to increase validity and reliability.

7.3 Experiment 3: Effect of Fatigue on Take-Over Performance

This study[4] and its results have been pre-published in Feldhütter et al. (2018). Some parts of the written text were adopted from this paper. Figures, tables, and data analyses were adapted for a consistent representation in this thesis.

7.3.1 Research Question and Purpose of this Study

One of the main research questions of this thesis is how fatigue affects take-over performance in CAD (RQ1). Experiment 1 and 2 demonstrated that fatigue vulnerability and the temporal fatigue development are highly individual. Therefore, RQ1 cannot be addressed completely by a study design with a fixed duration of CAD. It was concluded that a study design is required instead in which a RtI is only triggered once a participant has reached the fatigue level to be evaluated. In this experiment 3, such a fatigue-state-dependent study design was applied to examine the take-over performance of fatigued participants. For this purpose, two fatigue conditions were defined: the fatigued condition (FC) and the non-fatigued condition (NFC) to generate a baseline. Participants from both fatigue conditions experienced an identical take-over situation. In this study, no NDRA was available to focus solely on the evaluation of the effect of fatigue on take-over performance. Past research with similar study settings has yielded different findings concerning the effect of fatigue on take-over performance (for details see chapter 4.9). Due to these controversial and not distinct findings, the effect of fatigue on take-over performance was examined in an exploratory way.

Fatigue was mainly assessed by an expert rating that has already been used in experiment 1 and 2, but was adopted in such a way that it could be applied in real-time. Weinbeer et al. (2018) (see also Wehlack, 2019) and Vogelpohl et al. (2018) have already shown that this approach is useful. To have an even better basis for evaluating fatigue in real-time, the experts were supported by additional objective physiological measures that were also collected and processed in real-time. As in the two previous experiments, the PERCLOS metric was evaluated in terms of data availability and quality to obtain further insights into this metric under realistic conditions. For this purpose, the real-time expert rating of fatigue functioned as the ground truth. Comparisons between the experiments as well as to literature were made. As recommended by Phillips (2014) (see chapter 4.2 and chapter 4.8), in addition to the objective measures, fatigue was assessed subjectively by the KSS, which had already been used in experiment 2. The KSS was applied after the take-over situation to not bias the participants. Therefore, it can be considered as a retrospective control measure for the validity of the remaining measures.

[4] This study was conducted with the assistance of Dominik Kroll as part of his Master's this (Kroll, 2018).

7.3.2 Method

7.3.2.1 Participants

A total of 57 participants took part in the experiment, of which 30 participants were in the fatigued condition and 27 in the non-fatigued one. Technical problems caused two participants in the NFC and one participant in the FC to be excluded from data analysis. Seven of the remaining 29 participants (24 %) tested in the FC did not reach the pre-defined fatigue level within the 90 minutes of CAD and were excluded from the analysis of take-over performance. However, this group (termed CAD90 in the following) was kept for evaluating PERCLOS. All participants of the CAD90 were male and had a mean age of M=23 years (SD=1.51), ranging from 21 to 26 years. Gender was deliberately not varied in this group, since it was not used for take-over performance examination. The remaining sample for the take-over performance evaluation consisted of 47 participants (22 in the FC and 25 in the NFC). Six of the 22 participants (27 %) in the FC and five of the 25 participants (20 %) in the NFC were female. Their ages ranged between 20 and 27 years in the FC (M=23.50 years, SD=1.67) and between 20 and 47 years in the NFC (M=25.04 years, SD=5.17). The criterion for participation in the experiment was the possession of a valid driver's license for at least four years to prevent inexperienced drivers' behavior from biasing the results. Eight participants in each condition (FC: 36 %, NFC: 32 %) had possessed their driving license for four to five years, the remaining participants for six years or longer. Sixteen of 47 participants (34 %) had participated in a driving simulator experiment before (FC: 50 %, NFC: 20 %). Thirty-one of the 47 participants (66 %) rated their knowledge in terms of CAD (1 "very unfamiliar" to 5 "very familiar") at a medium level (3) or lower (FC: 55 %, NFC: 76 %) on a 5-level scale.

7.3.2.2 Study Design

Each participant was assigned to one of the two fatigue conditions prior to the experiment. For participants in the FC, the fatigue-state-dependent study design was realized by assessing fatigue in real-time and triggering the take-over situation once a pre-defined fatigue level was reached. For feasibility reasons, a RtI was triggered after 90 minutes of CAD at the latest, even if the participant had not reached the pre-defined fatigue level (cf. Figure 7-39). This cut-off criterion was assumed to be reasonable, since Weinbeer et al. (2018) found no further increase of fatigue after 75 minutes of CAD, and also the results of experiment 1 revealed that—if participants reach higher fatigue levels—this arises after 15 to 60 minutes of CAD. Apart from the usual standard monotonous course (as already applied in the previous experiments 1 and 2, see chapter 6.4) two further, smaller measures were taken to promote the development of passive TR fatigue: participants were instructed in advance not to drink caffeinated beverages on the day of the experiment, and the start of the experimental time slot was oriented towards the usual circadian low (before 8 a.m., after lunch between 1 p.m. and 2 p.m. or after 6 p.m).

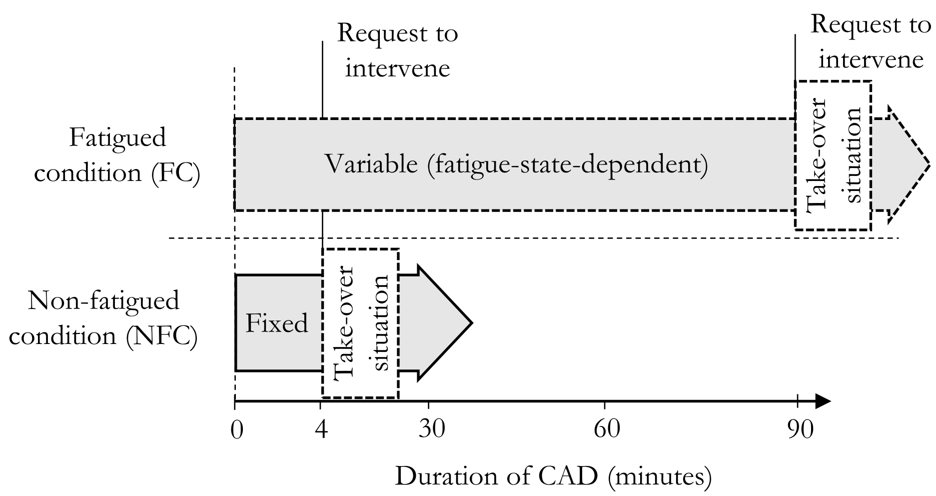


Figure 7-39. The fatigue-state-dependent study design of experiment 3 with two fatigue conditions.

For participants in the NFC, a non-fatigue state was mainly ascertained by a very short fixed automation duration of four minutes. Additionally, some more measures were taken to keep participants in a non-fatigued state. The simulated track was designed in a varied way with curves, an interesting landscape, and maneuvers such as speed changes and overtaking. Furthermore, participants had to do one minute of light rope skipping right before starting the experiment, since Weinbeer et al. (2019) showed that light physical exercise may prevent the development of fatigue. Moreover, participants in the NFC started the experiment between 10 a.m. and 12 p.m. or between 3 p.m. and 5 p.m. Participants in neither condition were offered any NDRA during the test drive to avoid the influencing factor of distraction or stimulation prior to the RtI. Furthermore, experiment 1 and 2 have shown that PERCLOS is of a rather low quality in the presence of a NDRA, which was to be avoided in this study.

As soon as each condition's (NFC and FC) triggering condition for a take-over was fulfilled, all participants experienced an identical take-over situation using the same scenario with a time budget of six seconds as in experiment 1 and experiment 2. The only variation consisted of an adapted traffic condition to a medium gap instead of a large gap (see Figure 7-40). A between-subjects design was used to completely exclude any biasing effects (e.g., training or anticipation of a take-over).

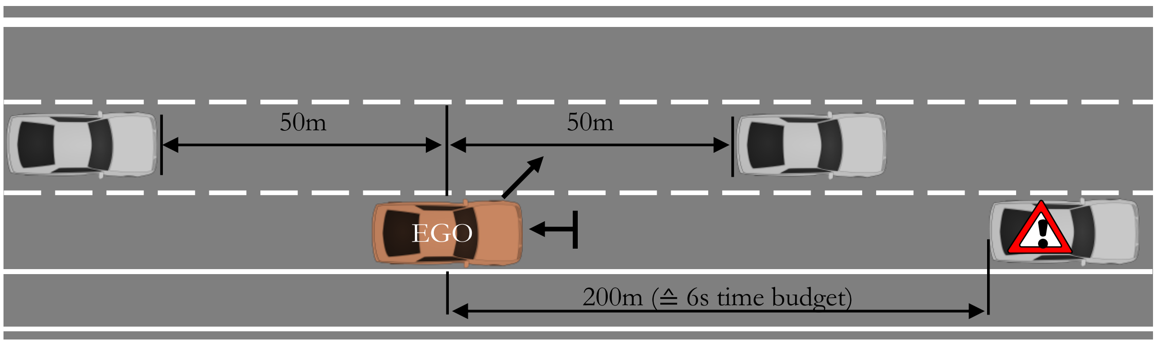


Figure 7-40. The take-over situation of experiment 3. Time budget=6 s, traffic condition=medium gap.

7.3.2.3 Fatigue Assessment

In the FC, two experts independently rated participants' fatigue in real-time during the test drive according to a rating protocol (see Appendix) watching real-time webcam material of the participants' faces and upper torsos. The experts were trained intensely in advance with video material and other training material similar to the protocol proposed by Wiegand et al. (2009) (see also chapter 4.8.5). The rating protocol is similar to the annotation guideline used in experiment 1 and experiment 2 and is based on the four-level fatigue scales developed by Wierwille and Ellsworth (1994) and Karrer-Gauß (2011), supplemented with further fatigue indicators listed in chapter 4.8.5. Based on the indicators protocolled, each participant was continuously assigned to one of the four fatigue levels FL1–FL4, which are analogous to experiment 1 and experiment 2 (cf. chapter 7.1.2.3 and chapter 7.2.2.3). Corresponding to experiment 1 and experiment 2, FL3 and FL4 are assumed to be critical in terms of performance decrements. As applied in the study of Weinbeer et al. (2018), the two experts synchronized their decision on rating the participant at FL3 or FL4 before triggering the RtI in order to raise the certainty about a correct rating. Due to this consultation, the calculation of an inter-rater reliability score is not possible. A good conformity of the rating behavior of the two raters was ascertained, though, by the previous collective intense training and the clear rating protocol according to which participants were classified.

Furthermore, the experts were supported by a software tool[5] proposed in Feldhütter et al. (2018), which was implemented in the driving simulator. The system is based on the approach of J. Schmidt, Braunagel, et al. (2016) (see also J. Schmidt, 2018) and evaluates multiple objective eye tracking metrics in real-time that are difficult for experts to recognize by observing the participants: PERCLOS, number of microsleep events (NMS) and head movement behavior (PHM) related to fatigue. These three metrics were chosen, since they can be assessed by the eye tracking system used and are also promising for assessing fatigue (see chapter 4.8.2.1 and 4.8.2.5). According to an algorithm fusing the three metrics (see C1–C5 in Figure 7-41), participants are binarily classified as fatigued or non-fatigued. PERCLOS thresholds were derived from the two previous experiments and literature review. Thresholds of NMS and PHM were derived from literature (cf. chapter 4.8.2.1 and 4.8.2.5). Apart from supporting the experts in their rating, the purpose of using this software tool was to test whether the validity of PERCLOS without a NDRA found in experiment 1 and experiment 2 can be proven under real-time conditions in a driving experiment. A detailed description of the tool and the examination in terms of sensitivity and specificity is not part of this thesis, but can be found in Feierle (2017), Kroll (2018) and Feldhütter et al. (2018). In this thesis, a retrospective evaluation of the PERLCOS metric was carried out analogously to experiment 1 and experiment 2. For

[5] The system was developed and implemented in close collaboration with Luis Kalb and Alexander Feierle as part of their Master's theses (Kalb, 2017; Feierle, 2017). With their assistance, two studies were conducted to collect data as basis for the development of the system. A detailed analysis of these data is documented in Kalb (2017) and Feierle (2017), and results are not repeated in this thesis.

quality reasons, data processing was done in an identical way as in experiment 1 and experiment 2 (cf. chapter 7.1.2.3).

Once both observers rated a participant at FL3 or FL4 and the software tool simultaneously classified her/him as fatigued, a take-over situation was triggered. In cases of an implausible output of the software tool (mostly due to poor data quality), the classification of the human raters alone also sufficed to trigger a take-over situation.

For the NFC, fatigue was assessed analogously to experiment 1 and experiment 2 with a fatigue rating conducted retrospectively by two experts (cf. chapter 7.1.2.3).

Additionally, in both fatigue conditions, fatigue was subjectively assessed by the German version of the KSS (Niederl, 2007), analogous to experiment 2 (cf. chapter 7.2.2.3). At a KSS scores of seven or higher, drivers have increasingly difficulties to stay awake and to avoid driving errors (Platho et al., 2013; Reyner & Horne, 1998); therefore, scores of seven or higher were assumed to be critical for take-over performance. However, to not bias the participants before and during the experiment, the KSS was not provided until after the take-over event was over and the participants had stopped the vehicle. Subsequently, the participants were asked to rate their fatigue level retrospectively for the moment right before the second RtI.

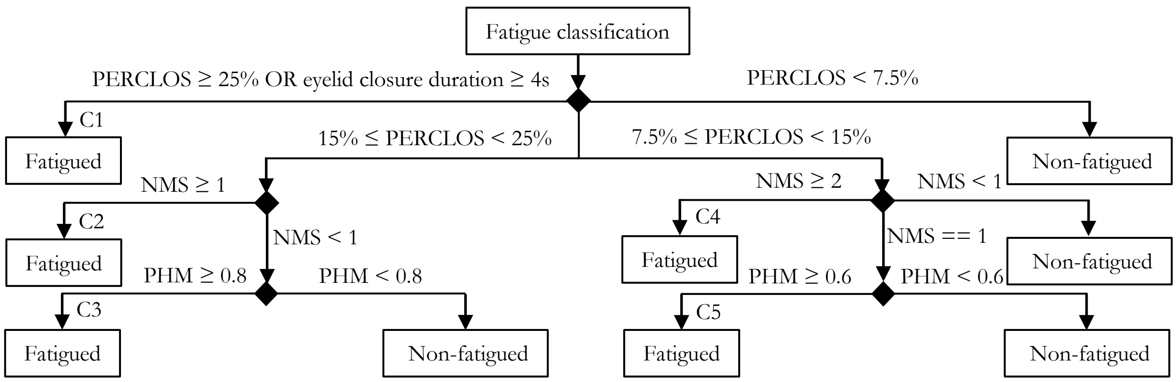


Figure 7-41. Fusing algorithm of the software tool to classify participants as fatigued or non-fatigued, modified from Feldhütter et al. (2018).

7.3.2.4 Procedure

Before the day of the experiment, participants were asked to fill in an online questionnaire containing socio-demographic questions, with which they also received a specific introduction depending on the fatigue condition. After welcoming the participants, the examiner gave a condition-specific introduction to the experiment and the participants could familiarize themselves with the driving simulator. Before entering the driving simulator, all participants were asked to leave all personal items (e.g., smartphone, watch) outside to prevent participants from vitalizing or distracting themselves. During a ten-minute training drive on a simulated three-lane German Autobahn, participants became familiar with the manual driving quality of the simulator and were asked to conduct different driving maneuvers, such as lane changing, accelerating and full braking. Furthermore, participants trained handling the CA (activation/deactivation) and experienced a RtI. Participants of the NFC subsequently completed a light physical exercise. The eye tracking system was calibrated. Afterwards, the participants experienced the condition-specific experimental track. Right after the participants had resolved the take-over situation, the simulation was stopped, and a post-interview was conducted. Apart from the KSS, the interview included questions about the participants' sleep behavior the night before the experiment (hours and quality of sleep) and about the perceived criticality of the take-over situation just experienced on a 10-level scale ranging from 1="Not critical" to 10="Extremely critical". Afterwards, participants received their compensation and were free to leave. For the NFC, the overall duration of the experiment was about 30 minutes, for the FC, between 45 minutes and 120 minutes.

7.3.2.5 Data Analysis

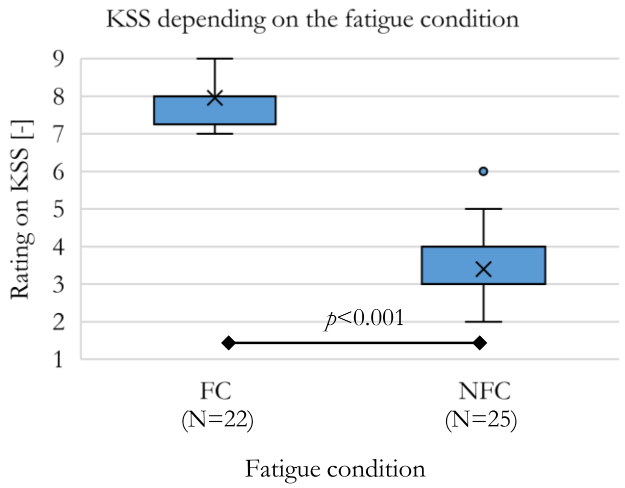
Due to quality or technical problems with the camera systems, data of some participants are partly (some minutes during the drive) or completely missing for the analysis of PERCLOS and for the expert fatigue rating, which results in smaller sample sizes for each metric. The exact sample sizes can be abstracted from the tables and illustrations.

7.3.3 Results

7.3.3.1 Fatigue Development

Due to the study design, self-rating of fatigue by means of the KSS is only available for the moment right before the RtI, which can be compared between the fatigue conditions. A temporal development of the KSS between different points in time during the drive is not possible.

For seven of the 29 participants (24 %) tested in the FC, 90 minutes of CAD were not enough to reach FL3 or FL4. For the remaining 22 participants, Figure 7-42 (right) displays the boxplot of the driving durations until participants in the FC reached FL3 or FL4. On average, it took 42 minutes until participants in the FC reached the critical fatigue level (*SD*=18, Min=18 min, Max=80 min). For 20 of the 22 participants (91 %), the RtI was triggered while they were at FL3, whereas two of the 22 participants were at FL4, since they jumped directly from FL2 at FL4. All of the seven participants who did not reach FL3 or FL4 (CAD90) were at FL2 in the end of the drive. In the NFC, 22 of 25 participants (88 %) were at FL1 before the RtI and three at FL2.



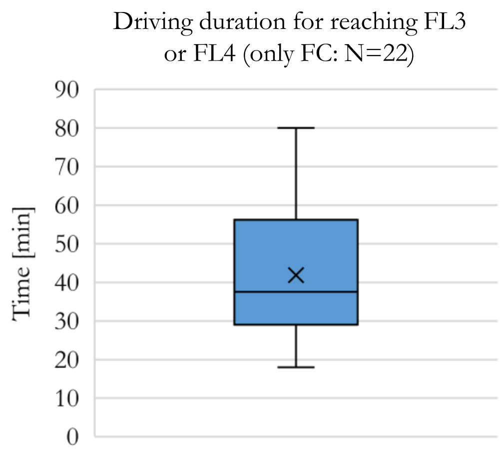


Figure 7-42. Left: Boxplot of the KSS rating depending on the fatigue condition. Right: Boxplot of the driving duration for reaching FL3 or FL4, only for the FC.

Figure 7-42 (left) shows the boxplot of participants' mean KSS rating depending on the fatigue condition. No participant in the FC rated herself/himself on a lower level than seven, and no participant in NFC on a higher level than six. For a statistical analysis of the effect of fatigue condition, the Mann-Whitney U-test was used, since data are ordinal-scaled. The result shows that participants in the FC ($M_{KSS,FC}$=7.96, SD=0.72) rated themselves at a significantly higher fatigue level than participants in the NFC ($M_{KSS,NFC}$=3.40, SD=1.00) with a large effect size (U=550.000, p<0.001, r_B=1.000).

7.3.3.2 Effect of Fatigue on Take-over Performance

Two-tailed Student's t-tests for independent samples were conducted to compare means of TOT, TTC, AccLong and AccLat in the two fatigue conditions (FC and NFC). In case of a violation of the normality assumption, the non-parametric Mann-Whitney U-test was performed. In case of a violation of the variance homogeneity, the Welch's t-test was performed. When both assumptions were violated, the Welch's t-test was chosen over the Mann-Whitney U-test as proposed by Rasch et al. (2011) and Ruxton (2006). Effect size is given by the rank biserial correlation r_B for the Mann-Whitney U-test, and for the other tests by Cohen's d. Chi-squared tests were calculated to investigate the existence of an association between the fatigue conditions and InRe, FinRe, crash and MC (bi- or multinomial data). In case of an expected frequency smaller than five, Fisher's exact tests were performed instead. Effect size was either calculated by the phi coefficient (φ) (2x2 contingency table) or by Cramér's V (variables other than dichotomous).

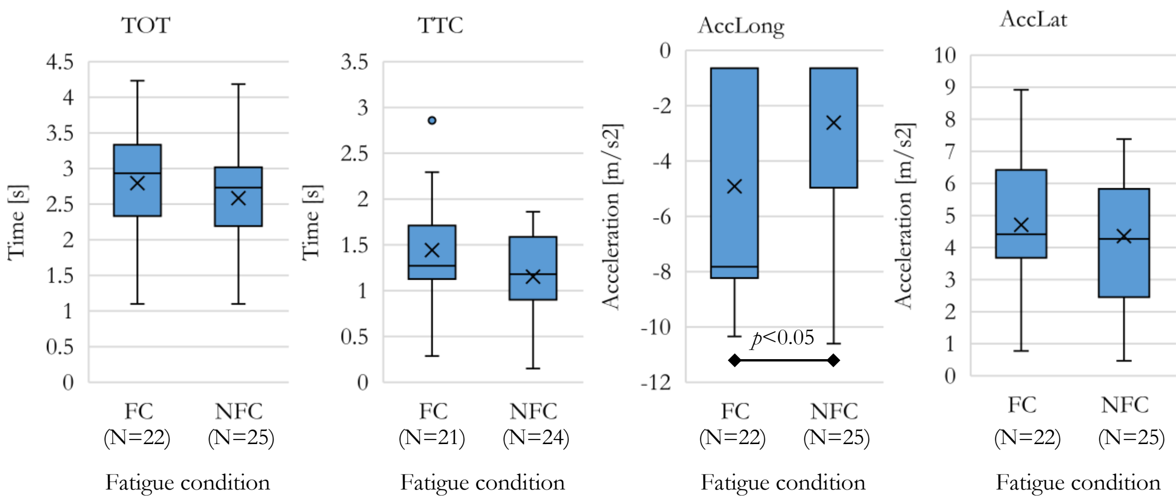


Figure 7-43. Boxplots of all metrical dependent variables of take-over performance depending on the fatigue condition.

Table 24. Descriptive data of all metrical dependent variables of take-over performance depending on the fatigue condition.

	Fatigue condition	***N***	**Mean**	***SD***	**Min**	**Max**
TOT [s]	FC	22	2.78	0.83	1.10	4.23
	NFC	25	2.59	0.79	1.10	4.18
TTC [s]	FC	21	1.44	0.67	0.29	2.90
	NFC	24	1.15	0.45	0.15	1.86
AccLong [m/s²]	FC	22	−5.11	4.19	−10.34	−0.64
	NFC	25	−2.61	3.41	−10.60	−0.64
AccLat [m/s²]	FC	22	4.69	2.35	0.78	8.92
	NFC	25	4.35	2.04	0.47	7.38

Note: In case of a collision with the broken-down vehicle, TTC is 0, hence, it was not considered for further analyses of TTC and sample size was reduced respectively.

Table 25. Results of mean comparison of the metrical dependent variables between the fatigue conditions.

	Test	**Statistic *t* (*U*)**	***df***	***p***	**Effect size *d* (r_B)**
TOT	Student	0.801	45	0.427	0.234
TTC	Student	1.688	43	0.099	0.505
AccLong	Welch	−2.226	40.585	**0.032**	−0.655
AccLat	Student	0.519	45	0.607	0.152

Note: For the Mann-Whitney-U test, test statistic is given by U-value and effect size by the rank biserial correlation r_B.

Figure 7-43 displays the boxplots for TOT, TTC, AccLong and AccLat, and Table 24 lists the corresponding descriptive data. The graphical inspection of the boxplots indicates greater TOT, greater TTC and greater accelerations (lateral and longitudinal) for participants in the FC. Results of the t-tests reveal that there is only a significant effect of the fatigue condition on the AccLong with a medium to large effect size (cf. Table 25). This means that participants in the FC decelerated stronger than participants in NFC ($\Delta M_{\mathrm{AccLong,FC\text{-}NFC}}$=2.5 m/s^2). For the TTC, results show a tendency to be greater for participants in the FC than for participants in the NFC with a medium effect size ($\Delta M_{\mathrm{TTC,FC\text{-}NFC}}$=0.29 s). Figure 7-44 and Figure 7-45 display the frequencies of InRe, FinRe, crash and MC. Results of the Chi-square tests show that participants in the FC carried out a full braking maneuver significantly more frequently than participants in the NFC with a small to medium effect size (see Table 26). Furthermore, there is a slight tendency for an association between InRe and the fatigue condition (fatigued participants more frequently showed a braking reaction as an initial response than non-fatigued participants) (cf. Table 26 and Figure 7-44).

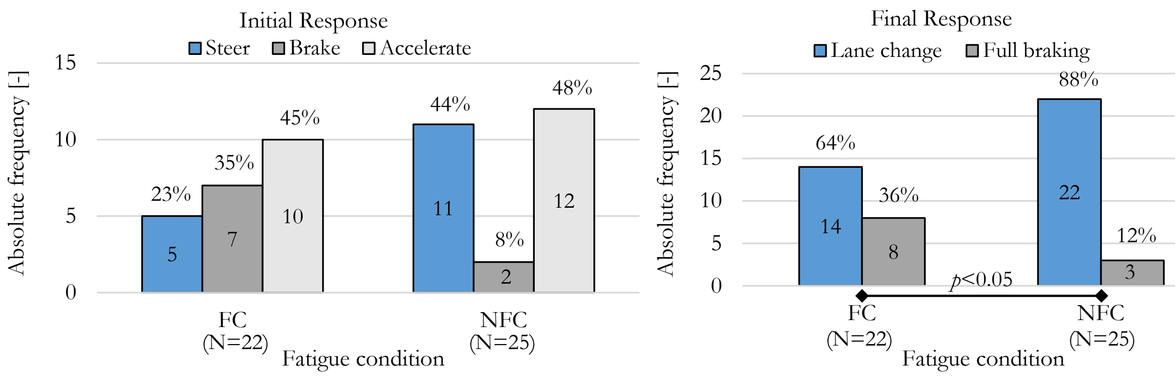


Figure 7-44. Frequency of initial response types (steer, brake, accelerate) and final response types (lane change, full braking) depending on the fatigue condition.

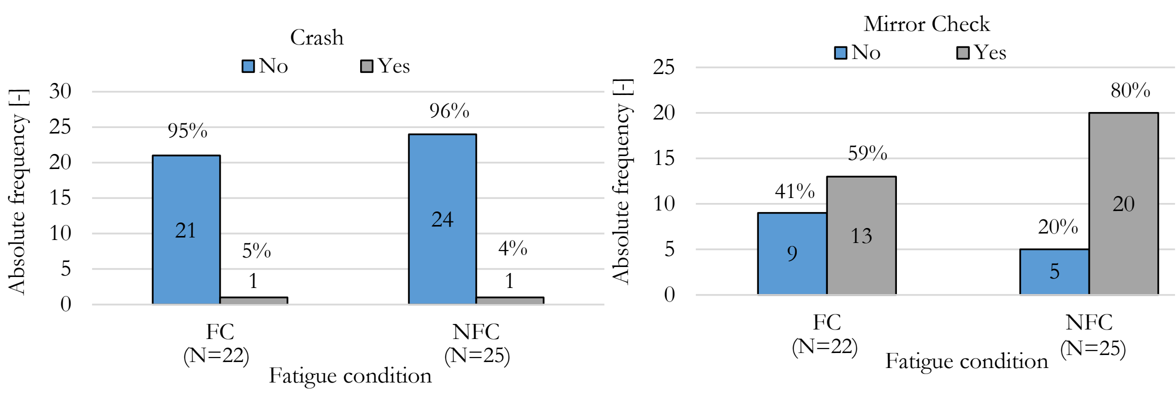


Figure 7-45. Frequency of crash and mirror check depending on the fatigue condition.

Table 26. Results of the Chi-squared tests to compare the frequencies of the bi- or multinomial dependent variables of take-over performance (InRe, FinRe, crash and MC) between the fatigue conditions FC and NFC.

	Test	Statistic χ^2	N	df	p	Effect size φ/V
InRe	Fisher's exact	4.860	47	2	0.095	0.327
FinRe	Pearson's Chi-square	3.875	47	1	**0.049**	0.287
Crash	Pearson's Chi-square	0.009	47	1	0.926	0.013
MC	Pearson's Chi-square	2.446	47	1	0.118	0.228

Note: In case that two cells have an expected frequency smaller than five, Fisher's exact test was conducted. For 2x2 contingency tables, effect size is given by phi coefficient φ. For other contingency tables by Cramér's V.

Figure 7-46 displays the boxplot of the perceived criticality of the take-over situation depending on the fatigue condition. The Mann-Whitney U-test was used instead of the Student's t-test to statistically analyze the effect of the fatigue condition, since data are ordinal-scaled. Effect size is given by the rank biserial correlation r_B. The result reveals that participants in the FC rated the same situation significantly more critical ($M_{\text{criticality rating,FC}}$=6.86, SD=2.73) than participants in the NFC conditions ($M_{\text{criticality rating,NFC}}$=5.16, SD=2.21) with a medium to large effect size (U=389.000, p=0.015, r_B=0.415).

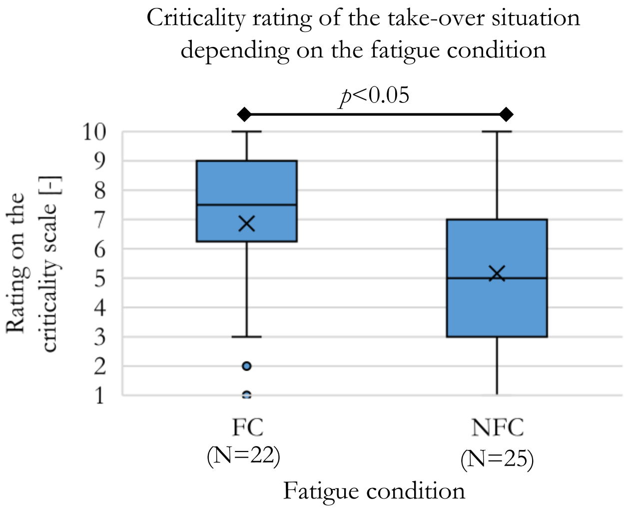


Figure 7-46. Subjective criticality rating of the take-over situation depending on the fatigue condition.

7.3.3.3 Evaluation of PERCLOS and Comparison to Fatigue Rating

Since no NDRA was available in both fatigue conditions, both conditions had the same preconditions in terms of data availability. While experts rated fatigue in real-time in the CAD90 and the FC, the rating was done retrospectively in the NFC. Even though no effect of this methodological variation is expected because the experts for both procedures are the same persons, an analysis of PERCLOS depending on the condition is carried, since nothing was found on it in literature. Due to the fatigue-state-dependent study design of the FC, an evaluation of mean PERCLOS per minute—as was done in experiment 1 and experiment 2—does not make sense, since the sample size of participants for each minute varies strongly. Furthermore, the sample size of PERCLOS values at FL3 or FL4 was rather low, also due to the study design, since the RtI was triggered and the experimental drive was stopped once participants were at a stable FL3 or FL4. Another reason for low sample sizes for FL3 and FL4 was the body posture of participants at these levels. They often slouched in the driving seat in such a way that they were out of the cameras' view.

For the NFC, 100 PERCLOS values (25 participants x four minutes of CAD) were theoretically available, for the FC, 921 PERCLOS values (22 participants x an average of 41.86 minutes of CAD), and for the CAD90, 630 PERCLOS values (seven participants x 90 minutes of CAD). Therefore, in total, there were 1651 theoretically available data points for PERCLOS calculation. The data processing that was chosen for quality reasons described in chapter 7.1.2.3 caused 130 of 1651 possible PERCLOS values to be removed, which resulted in an overall data availability for PERCLOS of 92 % (FC: 93 %; NFC: 91 %; CAD90: 91 %).

More insights on the evaluation of validity of PERCLOS is provided by Figure 7-47, which shows the boxplots of all available PERCLOS values depending on the fatigue level rated by the experts and the fatigue condition. Table 27 provides selected descriptive data corresponding to the boxplots. The selected parameters are identical to the ones defined in experiment 1 (see chapter 7.1.3.4). FL4 of the FC and FL2 of NFC are displayed (see Table 27 and Figure 7-47), but are not considered in the following evaluation due to the small sample sizes and, consequently, their limited validity.

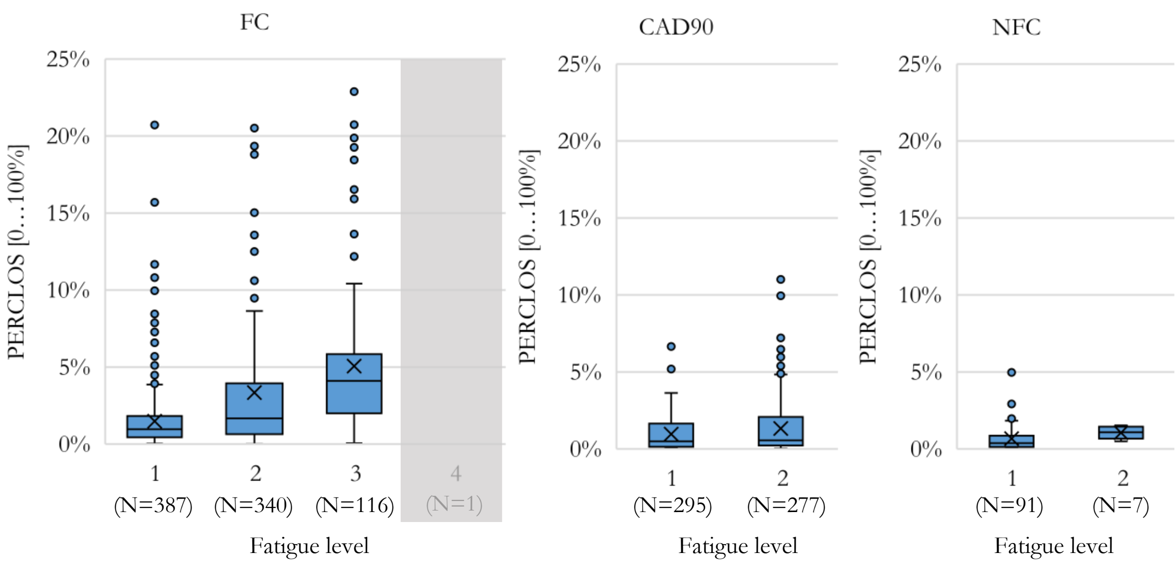


Figure 7-47. Boxplots of all available PERCLOS values depending on the fatigue condition (FC, CAD90, NFC) and the fatigue level. Note for FC: a) for a better readability, not all single outliers are displayed b) due to not being representative and for better readability, the single PERCLOS value of FL4 is not displayed.

In the FC, an increase of M_{PERCLOS} and Mdn_{PERCLOS} from FL1 to FL3 is notable: M_{PERCLOS} of FL2 is 2.2 times greater than M_{PERCLOS} of FL1 ($Mdn_{\text{PERCLOS,FL2}}$=1.7x$Mdn_{\text{PERCLOS,FL1}}$) and M_{PERCLOS} of FL3 is 1.5 times greater than M_{PERCLOS} of FL2 ($Mdn_{\text{PERCLOS,FL2}}$=2.5x$Mdn_{\text{PERCLOS,FL1}}$). For CAD90 and NFC, only a comparison between FL1 and FL2 is possible, even if the comparison is of questionable validity for the NFC. While there is almost no difference between FL1 and FL2 in CAD90 ($M_{\text{PERCLOS,FL2}}$=1.05x$M_{\text{PERCLOS,FL1}}$, $Mdn_{\text{PERCLOS,FL2}}$=1.1x$Mdn_{\text{PERCLOS,FL1}}$), there is a slight increase between FL1 and FL2 in the NFC ($M_{\text{PERCLOS,FL2}}$=1.5x$M_{\text{PERCLOS,FL1}}$, $Mdn_{\text{PERCLOS,FL2}}$=1.7x$Mdn_{\text{PERCLOS,FL1}}$). h_{outliers} is at a similar level around 5 % at FL1 and FL2 in all three conditions at (NFC is not included in this consideration due to the small sample size), except FL1 of the CAD90, in which h_{outliers} is 1 %. IQR_{PERCLOS} of FL1 is the smallest in the NFC (between about 2 and 4.5 times smaller than in the other conditions), and the highest in the FC. Furthermore, there are none or only very few extreme outliers ($h_{\text{outliers7\%}}$) for FL1 and FL2 in the CAD90 and NFC. In the FC, $h_{\text{outliers7\%}}$ is slightly higher.

Table 27. Descriptive data of all available PERCLOS values depending on the fatigue level and the fatigue condition.

Fatigue condition		**Fatigue level**			
		1	**2**	**3**	**4**
CAD90	$N_{\text{participants}}$	7	7	0	0
	N_{PERCLOS}	295	277	-	-
	M_{PERCLOS}	0.96 %	1.33 %	-	-
	Mdn_{PERCLOS}	0.50 %	0.56 %	-	-
	IQR_{PERCLOS}	1.51 %	1.86 %	-	-
	h_{outliers} (Q3+1.5xIQR)	1 %	5 %	-	-
	$h_{\text{outliers7\%}}$ (>7 %)	0 %	2 %^^	-	-
FC	$N_{\text{participants}}$	22	22	20	1
	N_{PERCLOS}	387	340	116	1
	M_{PERCLOS}	1.49 %	3.34 %	5.06 %	90.13 %*
	Mdn_{PERCLOS}	0.96 %	1.67 %	4.10 %	90.13 %*
	IQR_{PERCLOS}	1.38 %	3.31 %	3.84 %	-
	h_{outliers} (Q3+1.5xIQR)	5 %	5 %	11 %	-
	$h_{\text{outliers7\%}}$ (>7 %)	3 %	9 %	20 %	-
NFC	$N_{\text{participants}}$	25	3	0	0
	$N_{\text{available PERCLOS}}$	91	7	-	-
	M_{PERCLOS}	0.68 %	1.05 %*	-	-
	Mdn_{PERCLOS}	0.41 %	1.08 %*	-	-
	IQR_{PERCLOS}	0.85 %	0.78 %*	-	-
	h_{outliers} (Q3+1.5xIQR)	4 %	0 %*	-	-
	$h_{\text{outliers7\%}}$ (>7 %)	0 %	0 %*	-	-

Note: Numbers marked with * are not conclusive due to the small sample size.

7.3.4 Discussion

In the FC of this study, the RtI was triggered once participants reach FL3 or FL4, which was assessed in real-time by two experts. On average, it took participants about 40 minutes of CAD to reach FL3 or FL4, while the vast majority was in FL3 when the RtI was initiated. The wide range of resulting driving durations with CAD—varying from 18 and 80 minutes—substantiates the findings of experiments 1 and 2 and previous studies reported in literature that fatigue development is highly individual. The potential mechanisms and theories have already been discussed extensively in the context of experiment 1 and 2 (see chapter 7.1.4 and 7.2.4) and are not repeated here. However, it is emphasized once again that the specific factors that predict fatigue vulnerability are mostly unknown. Results of the KSS rating, which were significantly different between participants in the FC and NFC, indicate that the manipulation of the two conditions and the applied method succeeded. All participants in the FC placed themselves between level seven and nine on the KSS, which is considered to be critical for driving (Platho et al., 2013; Reyner & Horne, 1998), while all participants in the NFC placed themselves below a score of seven.

Slower reaction times due to slowed information processing and increased decision-making time are probably the most prominent hypothesized effects of fatigue when driving manually (see chapter 4.5), and these have also been found in previous studies on automated driving (see chapter 4.9). However, this could not be confirmed in this study, as TOT were not significantly lower for participants in the FC. Instead, fatigued participants exhibited full-braking maneuvers more frequently than non-fatigued participants to resolve the take-over situation, consequently pursued from the initial response (more frequently a brake reaction for fatigued participants by tendency) and with significantly higher decelerations than non-fatigued participants. In previous research (Gold et al., 2013), strong decelerations have been associated with missing situation awareness, because a strong braking maneuver extends the time to build up situation awareness and for making a decision on how to react. In this context, this argumentation also makes sense. The RtI was initiated when the participants in the FC were on the cusp of falling asleep. One main effect of fatigue is a "progessive withdrawal of attention from road and traffic demands" (Brown, 1994, p. 311). Both the visual attention is limited due to prolonged eye closure durations, as is cognitive attention due to slowed information processing or mind wandering ("driving without awareness" according to Brown (1994, p. 312), also known as *look but failed to see* problem), which causes reduced situation awareness. As a result, Brown (1994) and other authors (e.g., Dinges, 1995) report an increased crash probability, which was, however, not found in this study. Collisions were probably prevented in this study because the participants in the FC reflexively showed quick TOTs (comparable to the ones of the non-fatigued participants), which was probably caused by the extremely salient and, consequently, strongly stimulating RtI. An awakening effect of a warning signal in manual driving (Baulk et al., 2001) or of a RtI in CAD (Schömig et al., 2015) has already been suggested in earlier studies. Startled by this RtI, participants carried out a strong braking maneuver to minimize risk and to buy time, to build up situation awareness and to cope with the emergency situation. Another potential explanation for the absence of a slower reaction time in the fatigued drivers is that sleep deprivation is found to extend response times in manual driving rather than task-induced fatigue (Phillips, 2014), and the latter was tested in this study. In addition, in more recent studies Phillips (2014) found indicators that, particularly for complex tasks, which taking over control from CAD can be considered to be, fatigue may not necessarily cause a complete response inhibition, but difficulties to develop an optimal strategy to respond to a task. This indicates that the ability to recall knowledge-based behavior required to cope with such complex and unfamiliar tasks (Rasmussen, 1986) may be affected by fatigue. The results of the criticality rating additionally support the assumption that fatigued participants had more difficulties to cope with the take-over situation than non-fatigued participants, as fatigued participants perceived the identical take-over situation to be significantly more critical than non-fatigued participants. By contrast, almost 90 % of the non-fatigued participants decided to change lanes to resolve the situation without braking, while the large majority (80 %) also checked the mirrors before executing the maneuver. This indicates that non-fatigued participants did not have the difficulties with lacking situation awareness and the recall of knowledge-based behavior to cope with the complex situation, but were able to react consciously and deliberately.

It has also been found in other studies that there is no impairment of the TOT when comparing fatigued and non-fatigued drivers (Gonçalves et al., 2016; J. Schmidt et al., 2017; Weinbeer et al., 2018), which emphasizes the above-stated hypothesis. Other take-over performance parameters are not comparable to the studies of J. Schmidt et al. (2017) and Weinbeer et al. (2018) due to the differing take-over situations tested and parameters assessed. Gonçalves et al. (2016) found higher AccLat for fatigued drivers instead of the higher AccLong. Since Gonçalves et al. (2016) instructed their participants that monitoring the road and driving environment was mandatory at all times, missing situation awareness may not have been a problem in this study, which would explain the divergent findings. However, missing situation awareness may also occur when the driver's eyes are on the road ("driving without awareness"; Brown, 1994, p. 312). By contrast, Vogelpohl et al. (2018) found that fatigued participants driving with CAD more frequently chose to brake and stay in the lane as a final response to a RtI. However, the baseline was a group of fatigued manual drivers in this study, who experienced the same scenario with a warning signal and who often overtook the braking lead vehicle without braking. Nevertheless, Vogelpohl et al. (2018) argue it was missing situation awareness that caused fatigued participants to conduct a strong braking maneuver to gain more time for making a decision. It remains unclear whether the more frequent braking maneuvers were only an automation effect or maybe an interaction of automation and fatigue.

When comparing the results on take-over performance of this study with experiment 1 and experiment 2, the high proportion of almost 50 % who accelerated as an initial reaction to the RtI regardless of the fatigue condition is remarkable. In the previous studies, the proportion was between 5 % and 25 % (with one exception in the first take-over situation of the NC in experiment 2). The main variation compared to experiment 1 and experiment 2 was a different traffic condition by decreasing the gap between the vehicles on the adjacent lane. Since the large proportion of participants who accelerated was found in both fatigue conditions, it can be concluded that this reaction was to some extent caused by this traffic condition variation. Even though the gap on the adjacent lanes was still clearly large enough to easily change lanes, participants may have had the feeling that they had to accelerate and increase the distance to the vehicle behind them when they detected by looking into the mirror. This can also be caused by the limitation of a driving simulator, namely that distances are more difficult to estimate than in reality.

The evaluation of PERCLOS revealed a high data availability of more than 90 % in all three fatigue conditions (FC, NFC and CAD90), which is comparable to the result of experiment 1 and experiment 2 when participants were not engaged in a NDRA. The comparison with the fatigue rating showed that the mean, median and the dispersion (assessed by the IQR) of PERCLOS values at FL1 were comparably low in all three conditions. However, when comparing the increase of mean and median PERCLOS in the FC from FL2 to FL3, the increase was not as distinct in this study as it was in experiment 1 and experiment 2 for the UC (without a NDRA). The dispersion and the proportion of false positives of PERCLOS at FL1 and FL2 for the FC is comparable to the previous experiments. By contrast, the very low proportion of false positives at FL1 and FL2 for CAD90 and NFC is remarkable. The results of FL4 are not comparable, since only one data point is available in this study.

Results suggest for FL3 in the FC that PERCLOS is not as valid as in the conditions without a NDRA in the two previous experiments. One reason for this difference might be that the expert rating, which was the ground truth for fatigue and the basis for the evaluation of PERCLOS, was assessed in real-time for the FC instead of post-hoc, as was the case for the NFC as well as for the previous two experiments. Even though the training, the guideline and the indicators for the rating as well as the experts were identical in the studies, the real-time rating has some methodical limitations. Especially, the slight variations/fluctuation in the driver's fatigue state, especially when she/he fluctuates between fatigue levels, are not considered as well as in the post-hoc rating where raters annotate each manner according to a clear guideline and where raters have the possibility to fast-forward or rewind in case of an uncertainty. Furthermore, a shift to the next fatigue level was not documented before the experts agreed that the fatigue level was now stable. Therefore, it might have been the case that the participant had already had many microsleep events in one minute or had even had her/his eyes completely closed for a while before the experts changed the fatigue level to FL3 or FL4. This causes the expert rating to lag behind PERCLOS. At FL2, there were a lot of data points with a correct classification, which could compensate the outliers at least for the mean calculation. This was not the case for FL3, in which the sample size was comparatively low, since the experiment was stopped once participants reached this fatigue level. Even though this is a methodological weakness in this experiment, this finding provides a good indicator for the usage of PERCLOS under real driving conditions in a series implementation: a certain tolerance threshold needs to be considered, since single higher PERCLOS values might also occur due to other reasons than fatigue, and the PERCLOS history or temporal progression should be taken into account, for instance, by increasing the evaluation time frame to more than one minute.

Limitation of the Study and Conclusion

The results of the PERCLOS evaluation suggest that the real-time expert rating is probably not as accurate as the post-hoc rating in terms of being the minute-wise ground truth for PERCLOS evaluation, since slight fluctuations between fatigue levels are not displayed and fatigue levels have to become stable before they are set in the rating documentation. However, this is mostly accounted for by methodical reasons to realize a fatigue-state-dependent study design. Since rating participants at FL3 or FL4 is the triggering event for the take-over, raters have to be as certain as possible about the fatigue state. Therefore, the jump to a higher fatigue level was set only when both experts agreed on it and when the new fatigue level was apparently stabilized. Results of the KSS additionally showed that the fatigue manipulation was successful in the two conditions. Therefore, the validity of the real-time expert rating method for a fatigue-state-dependent study design is not questionable, but it is rather the minute-wise ground truth for PERCLOS evaluation.

The individual automation durations that were needed in this study by the participants to reach the pre-defined fatigue level confirm that fatigue development is highly individual. Furthermore, it emphasizes that the use of a fatigue-state-dependent study design is necessary to comprehensively address research questions which require a specific fatigue state or the controlled variation of fatigue states (in this thesis RQ1).

Results of examining the take-over performance showed that fatigue did not have the previously assumed impact on take-over performance, namely prolonging TOT, indicating that the RtI had a strong stimulating effect. However, the take-over quality of fatigued participants was lower due to supposed lacking situation awareness. Situation awareness is a fundamental prerequisite for a safe operation of vehicles and for a safe take-over from CAD (cf. chapter 3.3.2). In this study, missing situation awareness did not lead to an increased crash probability; however, different take-over situations need to be tested to further investigate these findings. It is conceivable that fatigued drivers' less confident behavior might lead to a more obvious deterioration of the driving performance in even more complex situations.

7.4 Experiment 4: Effect of Fatigue on Take-Over Performance in Different Urgent Situations

This study[6] and its results have been pre-published in Feldhütter et al. (2019). Some parts of the written text were adopted from this paper. Figures, tables, and data analyses were adapted for a consistent representation in this thesis.

7.4.1 Research Questions and Purpose of the Study

In experiment 3, no effect of fatigue on the TOT was found. This finding was surprising in the first place because previous studies, mainly in the context of manual driving, but also some on automated driving found that fatigue prolongs reaction times to (emergency) events (see chapter 4.5 and chapter 4.9). It was concluded that the salient, and consequently, stimulating RtI probably had an awakening effect, leading to a startle reaction to cope with the situation. Additionally, the results of experiment 3 showed that missing situation awareness is probably one main problem of fatigued drivers, leading to a strong braking reaction. Even though the reaction of fatigued drivers was less confident and reflexive, this behavior did not lead to more accidents. The conclusion of experiment 3 was that different take-over situations should be tested to confirm the findings of this study.

J. Schmidt et al. (2017) already compared various take-over situations in different fatigue states assessed by the KSS. Scenarios comprise strong braking of the leading vehicle, accident on the ego-lane (similar to the scenario used in this thesis), deviation of the ego-vehicle from the road in a curve, broken-down vehicle on the shoulder and box on the ego-lane with low time budget. The situations were tested in a within-subjects design (in the indicated order). The take-over scenarios were rather different from each other in terms of complexity, predictability, criticality and urgency (for the definitions of the taxonomy used see chapter 3.3.1). KSS evaluation revealed significantly rising ratings with each take-over situation. J. Schmidt et al. (2017) compared hands-on-steering-wheel time between the five scenarios and between the KSS rating in each situation. It was only the scenario that had a significant effect on the reaction time, but not the KSS rating. The first result is in line with previous research (see chapter 3.3.3), the latter with the results of experiment 3. However, no conclusion can be drawn from the study of J. Schmidt et al. (2017) in terms of the take-over quality (e.g., response types or accelerations) not least because of the high variations in characteristics of the take-over situations.

Therefore, only one situational factor influencing the take-over performance is varied in this study to make it comparable to experiment 3. At the same time, the fatigue-state-dependent study design with the two fatigue conditions FC and NFC is applied to replicate the findings of experiment 3 and to contribute to a greater transparency in terms of the effect of fatigue on

[6] This study was conducted with the assistance of Alexandra Ruhl as part of her Master's thesis (Ruhl, 2018).

take-over performance (RQ1). The situational factor to be varied was chosen to be a higher urgency of the take-over situation, which is why the time budget to react to the RtI was reduced from six seconds to five seconds. A time budget of five seconds had already been used for a critical testing scenario, for instance, by Gold et al. (2013). From the results of experiment 3, missing situation awareness was assumed to be the main problem for fatigued drivers. Situation awareness can only be built up with sufficient time. Therefore, it is supposed that a smaller time budget will reinforce the effect of fatigue, worsening take-over quality found in experiment 3, since there would be less time to correct hectic and erroneous behavior. Hectic and erroneous behavior manifested in startled reactions and strong braking maneuvers in experiment 3. This behavior is hypothesized to occur more frequently or in an intensified way with a smaller time budget; however, different behavior patterns not yet found in experiment 3 that could contribute to a poor take-over quality may also arise. Eventually, a poorer take-over quality might increase the occurrence of accidents, as well.

In addition, to be able to evaluate potential interaction effects of time budget and fatigue on take-over performance, the consolidated data from both experiments together are analyzed. To make a statistical comparison to experiment 3 valid, the identical method in terms of study design and fatigue assessment is applied, and only the time budget is varied between six seconds (TB6) and five seconds (TB5).

7.4.2 Method

7.4.2.1 Participants

Fifty-two participants took part in this experiment, 29 participants of which were in the FC and 23 in the NFC. Technical problems caused three participants in the NFC and one participant in the FC to be excluded from data analysis. Eight of the remaining 28 participants (29 %) tested in the FC did not reach the defined fatigue level within the 90 minutes of CAD and were excluded from the analysis of take-over performance. However, this group (termed CAD90 in the following) was kept for evaluating PERCLOS. Two of the eight participants (25 %) in the CAD90 were female and had a mean age of *M*=25 years (*SD*=1.87), ranging from 21 to 26 years. The remaining sample for the take-over performance evaluation consisted of 40 participants (20 in the condition). Eight of the 20 participants (40 %) in the FC and seven of the 20 participants (35 %) in the NFC were female. The age ranged between 22 and 31 years in the FC (*M*=25.15 years, *SD*=2.03) and between 23 and 59 years in the NFC (*M*=30.25 years, *SD*=9.28). Criterion for participation in the experiment was the possession of a valid driver's license for at least four years to prevent inexperienced drivers' behavior from biasing the results. Two out of 20 participants (10 %) in the FC and one out of 20 participants (5 %) in NFC had held their driving license for five years, the remaining participants for six years or longer. Fifteen of 40 participants (38 %) had participated in a driving simulator experiment before (FC: 40 %, NFC: 35 %). Twenty-nine of the 40 participants (58 %) rated their knowledge in terms of CAD on a 5-level scale (1 “very unfamiliar” to 5 “very familiar”) at a medium level (3) or lower (FC: 60 %, NFC: 55 %).

7.4.2.2 Study Design

The study design was identical to the one from experiment 3 (cf. chapter 7.3.2.2) and is displayed in Figure 7-39.

Each participant was assigned to one of the two fatigue conditions (FC and NFC) prior to the experiment, using a between-subjects design to completely exclude any biasing effects (e.g., training or anticipation of a take-over). For participants in the FC, the fatigue-state-dependent study design was realized by assessing fatigue in real-time and triggering the take-over situation once a pre-defined fatigue level had been reached. Experiment 3 revealed that a proportion of almost 25 % participants did not reach FL3 or FL4 within 90 minutes of CAD. However, it was decided to adhere to this cut-off criterion, because pre-studies showed that some participants became tense, nervous, angry or even aggressive when the automation duration was further extended. The purpose of the pre-studies was the technical and methodical training of the examiners and the fatigue raters. Data of these pre-studies were not evaluated. As already applied in experiment 3, the same measures were taken to gently promote the development of passive TR fatigue: the standard monotonous experimental course (as in experiment 1 and experiment 2, see chapter 6.4), the previous introduction not to drink caffeinated beverages on the day of the experiment and the start of the experimental time slot within a potential circadian low (before 8 a.m., after lunch between 1 p.m. and 2 p.m. or after 6 p.m).

Since the study design and method of experiment 3 for participants in the NFC succeeded in inducing a non-fatigue state, the same approach was used for this condition in this study: a fixed automation duration of four minutes, a diverse and interesting simulated test track, one minute of light rope skipping right before starting the experiment, and starting the experiment beyond the usual circadian low phases (between 10 a.m. and 12 p.m. or between 3 p.m. and 5 p.m.).

As soon as each condition’s (NFC and FC) triggering condition for a take-over was fulfilled, all participants experienced an identical take-over situation using the same scenario with the identical traffic condition as in experiment 3. The only variation consisted of an adapted time budget of five seconds instead of six seconds (see Figure 7-48).

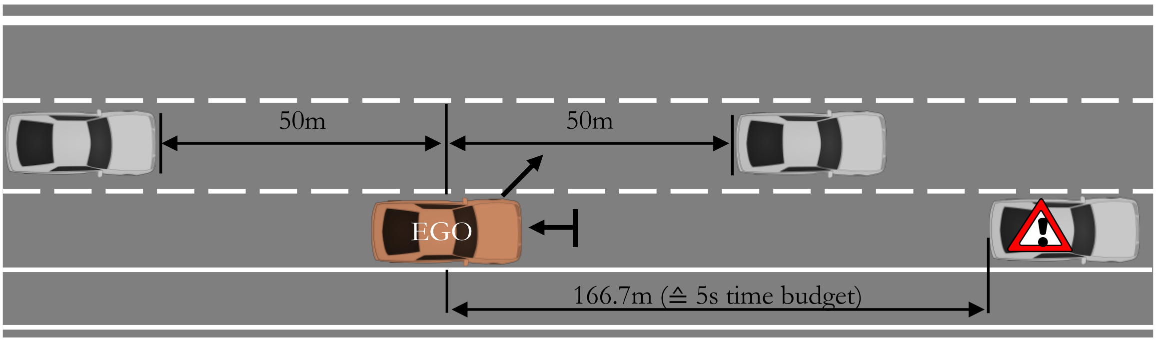


Figure 7-48. Take-over situation of experiment 4. Time budget=5 s, traffic condition=medium gap.

7.4.2.3 Fatigue Assessment

Since the method for assessing fatigue proved to be successful in experiment 3, and data should be as comparable as possible to experiment 3, the identical method was applied in this study. Therefore, the description can be found in chapter 7.3.2.3 and is not repeated here.

7.4.2.4 Procedure

The procedure of this study was identical to the one used in experiment 3 (cf. chapter 7.3.2.4), which is why it is not repeated here.

7.4.2.5 Data Analysis

Due to quality or technical problems with the camera systems, data of some participants are partly (some minutes during the drive) or completely missing for the analysis of PERCLOS and for the expert fatigue rating, which results in smaller sample sizes for each metric. The exact sample sizes can be abstracted from the tables and illustrations.

7.4.3 Results

7.4.3.1 Fatigue Development

For eight of the 28 participants (29 %) tested in the FC, 90 minutes of CAD were not enough to reach FL3 or FL4. For the remaining 20 participants, Figure 7-49 (right) displays the boxplot of the driving duration of all participants in the FC until they reached FL3 or FL4. On average, it took 38 minutes until participants in the FC reached the critical fatigue level (*SD*=20, Min=11 min, Max=73 min). For 18 of the 20 participants (90 %), the RtI was triggered while they were at FL3, whereas two of the 20 participants were at FL4, since they jumped directly from FL2 to FL4. All of the eight participants who did not reach FL3 or FL4 (this group is termed CAD90 analogously to experiment 3) were at FL2 at the end of the drive. In the NFC, 17 of 20 participants (85 %) were at FL1 before the RtI, and the remaining three at FL2.

Figure 7-49 (left) shows the boxplot of participants' KSS rating depending on the fatigue condition. Nineteen of the 20 participants (95 %) in the FC rated themselves on a KSS-level of seven or higher (one on level four), while no participant in NFC rated herself/himself on a level higher than six. For a statistical analysis of the effect of fatigue condition on the KSS rating, the Mann-Whitney U-test was used, since data are ordinal-scaled. The result shows that participants in the FC ($M_{KSS,FC}$=7.70, SD=1.22) rated themselves on a significantly higher fatigue level than participants in the NFC ($M_{KSS,NFC}$=4.1, SD=1.29) with a large effect size (U=390.500, p<0.001, r_B=0.953).

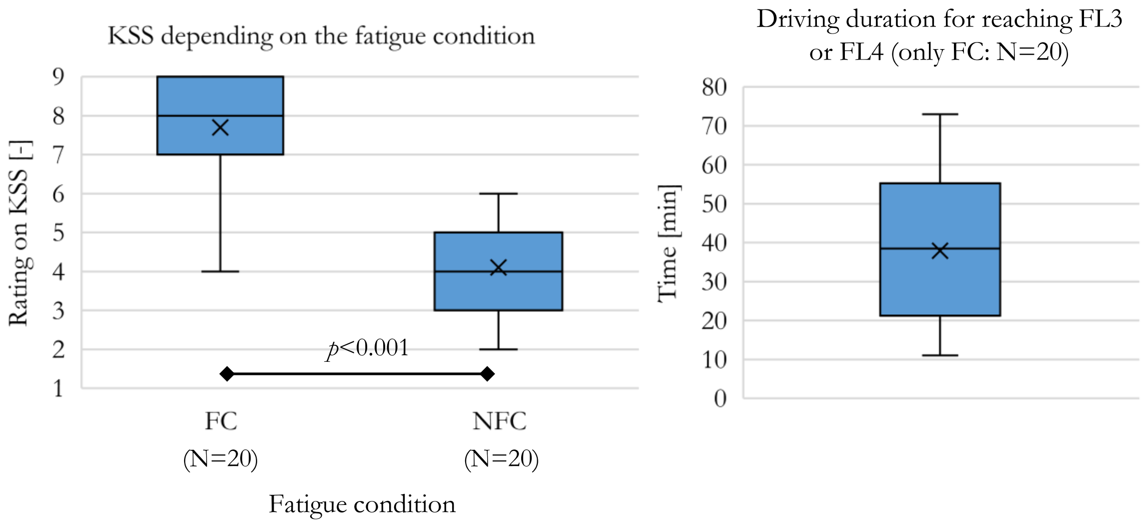


Figure 7-49. Left: Boxplot of the KSS rating depending on the fatigue condition. Right: Boxplot of the driving duration for reaching FL3 or FL4 (only for the FC).

7.4.3.2 Effect of Fatigue on Take-over Performance

Two-tailed Student's t-tests for independent samples were conducted to compare means of TOT, TTC, AccLong and AccLat in the two fatigue conditions (FC and NFC). In case of a violation of the normality assumption, the non-parametric Mann-Whitney U-test was performed. In case of a violation of the variance homogeneity, the Welch's t-test was performed. When both assumptions were violated, the Welch's t-test was chosen over the Mann-Whitney U-test as proposed by Rasch et al. (2011) and Ruxton (2006). Effect size is given by the rank biserial correlation r_B for the Mann-Whitney U-test, and for the other tests by Cohen's d. Chi-squared tests were calculated to investigate the existence of an association between the fatigue conditions and InRe, FinRe, crash and MC (bi- or multinomial data). In case of an expected frequency smaller than five, Fisher's exact test was performed instead. Effect size was either calculated by the phi coefficient (φ) (2x2 contingency table) or by Cramér's V (variables other than dichotomous).

Figure 7-50 displays the boxplots for TOT, TTC, AccLong and AccLat, and Table 28 lists the corresponding descriptive data. The graphical inspection of the boxplots indicates almost identical TOT, TTC and AccLat in both conditions, whereas the variance of TTCs in the FC is clearly wider than in the NFC. Mean AccLong appears to be greater for participants in the FC with clearly smaller variance ($Q1_{AccLong,FC}=-5.6\,m/s^2$, $Mdn_{AccLong,FC}=-8.04\,m/s^2$, $Q3_{AccLong,FC}=-8.31\,m/s^2$, $Q1_{AccLong,NFC}=-0.64\,m/s^2$, $Mdn_{AccLong,NFC}=-7.31\,m/s^2$, $Q3_{AccLong,NFC}=-8.17\,m/s^2$, cf. Figure 7-50). Results of the t-tests reveal that there is no significant effect of fatigue on any of the metrical take-over performance parameters (cf. Table 29).

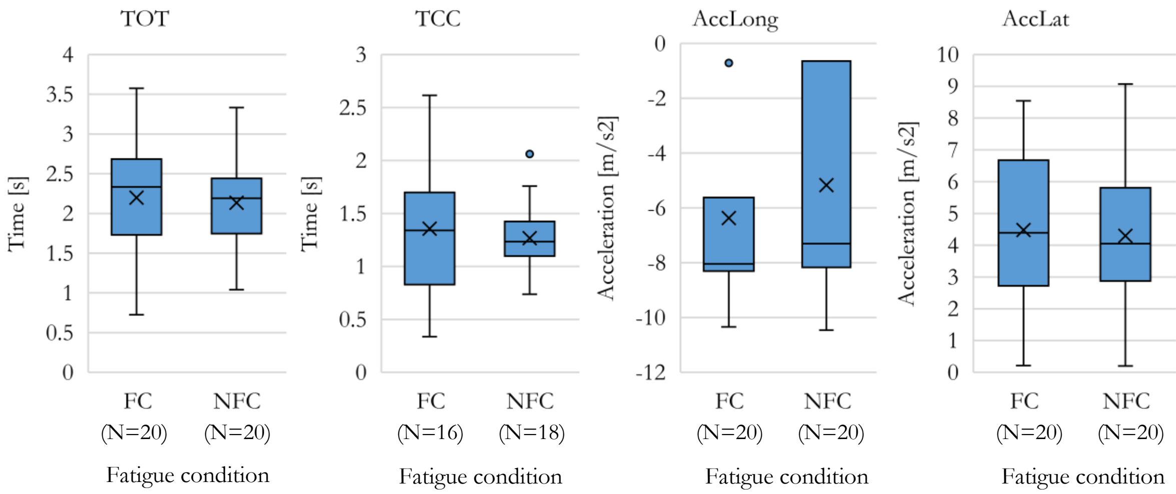


Figure 7-50. Boxplots of all metrical dependent variables of take-over performance depending on the fatigue condition.

Table 28. Descriptive data of all metrical dependent variables of take-over performance depending on the fatigue condition.

	Fatigue condition	*N*	Mean	*SD*	Min	Max
TOT [s]	FC	20	2.20	0.71	0.73	3.58
	NFC	20	2.14	0.57	1.04	3.33
TTC [s]	FC	16	1.36	0.65	0.34	2.61
	NFC	18	1.27	0.33	0.74	2.06
AccLong [m/s²]	FC	20	−6.37	4.19	−10.34	−0.64
	NFC	20	−5.17	3.41	−10.46	−0.64
AccLat [m/s²]	FC	20	4.48	2.63	0.21	8.54
	NFC	20	4.30	2.22	0.20	9.07

Note: In case of a collision with the broken-down vehicle, TTC is 0, hence, it was not considered for further analyses of TTC and sample size was reduced respectively.

Table 29. Results of mean comparison of the metrical dependent variables between the fatigue conditions FC and NFC.

	Test	Statistic t (U)	df	p	Effect size d (r_B)
TOT	Student	0.312	38	0.757	0.099
TTC	Welch	0.503	21.889	0.620	0.176
AccLong	Mann-Whitney	170.000	-	0.422	−0.150
AccLat	Student	0.236	38	0.815	0.075

Note: For the Mann-Whitney-U test, test statistic is given by U-value and effect size by the rank biserial correlation r_B.

Figure 7-51 and Figure 7-52 display the frequencies of InRe, FinRe, crash and MC. Results of the Chi-square tests show that there is a significant association between fatigue condition and InRe (cf. Table 30). Post-hoc tests of the single types of InRe with Bonferroni correction revealed no significant result. However, when inspecting Figure 7-51 (left) and the comparison between expected frequencies h_{expected} and observed frequencies h_{observed} (see Appendix), it becomes apparent that participants in the FC steered considerably less frequently as an initial response ($h_{\text{steering,observed,FC}}$=10 %) than expected ($h_{\text{steering,expected}}$=28 %). By contrast, participants in the NFC steered more frequently ($h_{\text{steering,observed,NFC}}$=45 %) than expected. Furthermore, the high proportion of $h_{\text{accelerating,observed,FC}}$=55 % in the FC who initially accelerated after the RtI is noticeable, whereas in the NFC, the proportion was $h_{\text{accelerating,observed,NFC}}$=30 %. For crash, participants crashed twice as frequently as participants in the NFC did, but no significant association between crash and fatigue condition was found.

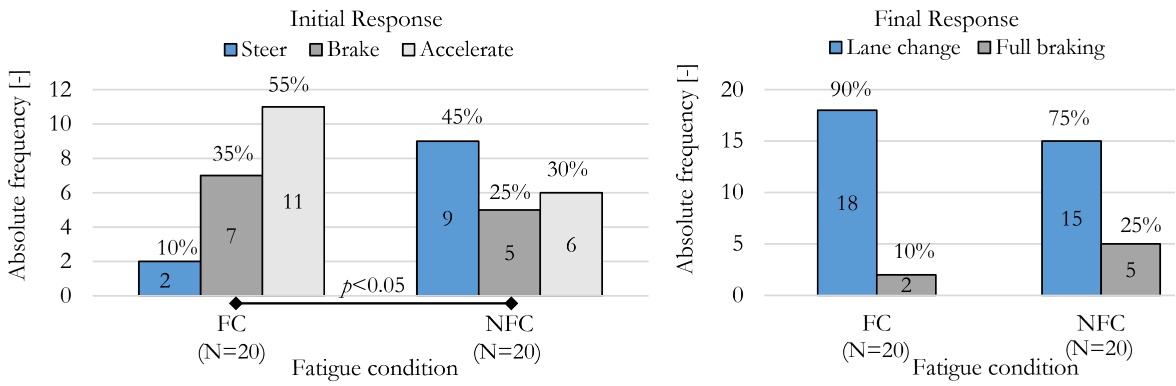


Figure 7-51. Frequency of initial response types (steer, brake, accelerate) and final response types (lane change, full braking) depending on the fatigue condition.

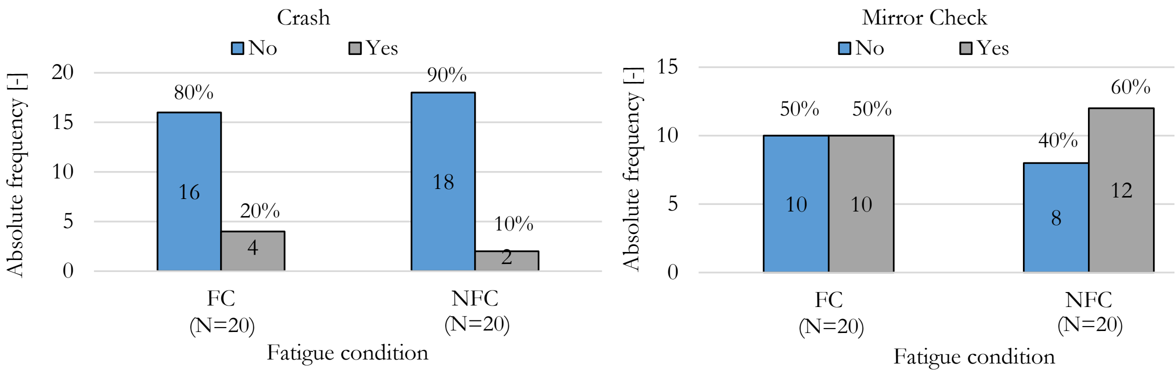


Figure 7-52. Frequency of crash and mirror check depending on the fatigue condition.

Table 30. Results of the Chi-squared tests to compare the frequencies of the bi- or multinomial dependent variables of take-over performance (InRe, FinRe, crash and MC) between the fatigue conditions FC and NFC.

	Test	Statistic χ^2	N	df	p	Effect size φ / V
InRe	Pearson's Chi-square	6.258	40	2	**0.049**	0.396
FinRe	Fisher's exact	1.558	40	1	0.407	0.197
Crash	Fisher's exact	0.784	40	1	0.661	0.013
MC	Pearson's Chi-square	0.404	40	1	0.751	0.101

Note: In case that two cells have an expected frequency smaller than five, Fisher's exact test was conducted. For 2x2 contingency tables, effect size is given by phi coefficient φ. For other contingency tables by Cramér's V.

Figure 7-53 displays the boxplot of the perceived criticality of the take-over situation depending on the fatigue condition. The Mann-Whitney-U test was used to statistically analyze the effect of fatigue condition since data are ordinal-scaled. Effect size is given by the rank biserial correlation r_B. There is no significant effect of the fatigue condition on the perceived criticality of the take-over situation ($M_{\text{criticality rating,FC}}$=7.85, SD=1.98; $M_{\text{criticality rating,NFC}}$=7.25, SD=2.47; U=225.500, p=0.489, r_B=0.128).

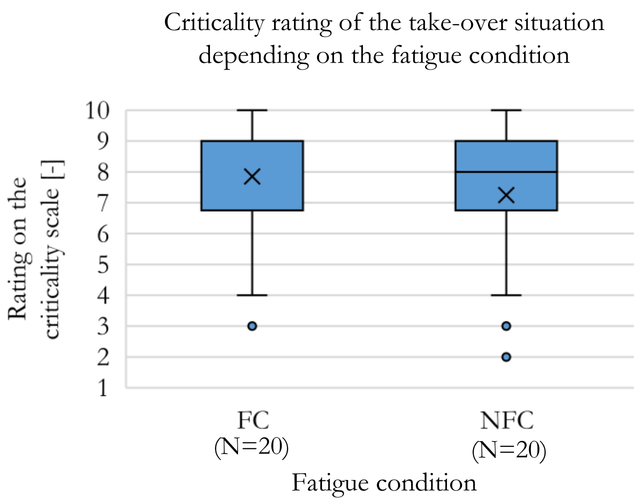


Figure 7-53. Subjective criticality rating of the take-over situation depending on the fatigue condition.

7.4.3.3 Effect of Fatigue on Take-over Performance Comparing Two Take-over Situations

To analyze the main effect and interaction effect of time budget (TB6, TB5) and fatigue condition (FC, NFC) across the data of experiment 3 and experiment 4, a 2x2 ANOVA was conducted for all metrical take-over performance metrics (TOT, TTC, AccLong, AccLat). In case of a significant result of the Levene's test, Hartley's F_{max} test was calculated. Due to the different sample sizes, omega-squared ω^2 was used instead of partial eta-squared η_p^2 to indicate effect size.

Figure 7-54 displays boxplots of TOT, TTC, AccLong and AccLat depending on the fatigue condition and the time budget, and Table 31 the corresponding descriptive data. The graphical inspection indicates longer TOTs for TB6, slightly greater TTC for TB6 and greater AccLong for both TB6 and the FC. Results of the ANOVA confirm a significantly longer TOT for TB6 ($\Delta M_{\text{TOT,TB5-TB6}}=0.52\,\text{s}$) with a medium to large effect size (cf. Table 32). For AccLong, the Shapiro-Wilk test and the Levene's test are significant. Hartley's F_{max} test, however, indicates a variance ratio smaller than ten (<2:1) at a maximum sample size ratio of 1.3. According to Bühner and Ziegler (2017, p. 369), this means that the actual α-error probability is not increased and the ANOVA can be interpreted without α-correction. To account for the non-normally distributed data, the non-parametric Kruskal-Wallis H-test was calculated. Since the Kruskal-Wallis H-test yielded the same result as the ANOVA in terms of being below the significance level of 0.05 and ANOVA was proved to be relatively robust against non-normal distribution (Bühner & Ziegler, 2017, p. 368; Glass et al., 2016; Harwell et al., 2016; Salkind, 2010), the result of the ANOVA is considered valid. Results of the ANOVA revealed a significant main effect of fatigue condition and time budget on the AccLong, both with a small to medium effect size (cf. Table 32). Participants in the FC decelerated significantly more strongly than participants in the NFC ($\Delta M_{\text{AccLong,FC-NFC}}=1.85\,\text{m/s}^2$), and participants with TB5 decelerated significantly more strongly than participants with TB6 ($\Delta M_{\text{AccLong,TB5-TB6}}=1.91\,\text{m/s}^2$). For the TTC, Shapiro-Wilk test for one group and Levene's test are significant. Hartley's F_{max} test, however, indicates a variance ratio smaller than ten (4:1) at a maximum sample size ratio of 1.5, which means that the result of the Levene's test is negligible (see above). In terms of the non-normal distribution of one subgroup of the TTC, the result of ANOVA was considered valid for the same reasons as for the AccLong. The result of the ANOVA revealed that there is no main effect of fatigue condition and time budget on TTC. Furthermore, no main effect of fatigue condition and time budget on AccLat was found. Moreover, there is no significant interaction effect of the two factors on TOT, TTC, AccLong or AccLat (see Table 32).

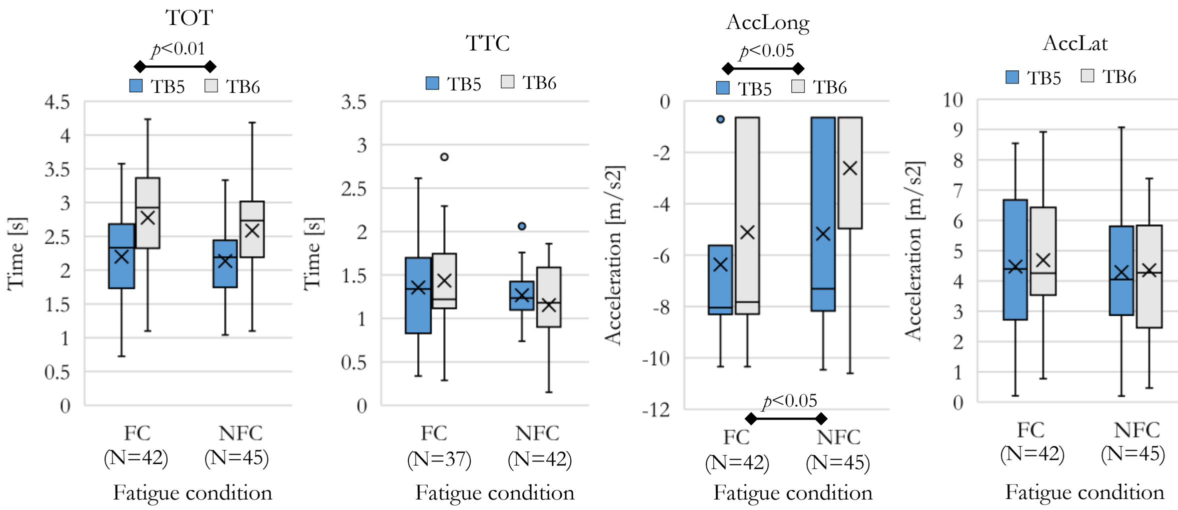


Figure 7-54. Boxplots of all metrical dependent variables of the take-over performance depending on the fatigue condition and the time budget.

Table 31. Descriptive values of all metrical dependent variables of take-over performance depending on the fatigue condition and the time budget.

		Time budget (between factor)					
		TB5 (*N*=40)[a]			**TB6 (*N*=47)[a]**		
	Fatigue condition (between factor)	***M* (*SD*)**	**Min**	**Max**	***M* (*SD*)**	**Min**	**Max**
TOT [s]	FC (*N*=42)	2.20 (0.71)	0.73	3.58	2.78 (0.83)	1.10	4.23
	FNC (*N*=45)	2.14 (0.57)	1.04	3.33	2.59 (0.79)	1.10	4.18
TTC [s]	FC (*N*=37)	1.84 (0.61)	0.34	2.61	1.97 (0.69)	0.29	2.90
	FNC (*N*=42)	1.66 (0.40)	0.74	2.06	1.88 (0.67)	0.15	1.86
AccLong [m/s²]	FC (*N*=42)	−6.37 (3.43)	−10.38	−0.64	−5.11 (4.18)	−10.34	−0.64
	FNC (*N*=45)	−5.17 (3.97)	−10.46	−0.64	−2.61 (3.41)	−10.60	−0.64
AccLat [m/s²]	FC (*N*=42)	4.48 (2.63)	0.21	8.54	4.69 (2.35)	0.78	8.92
	FNC (*N*=45)	4.30 (2.22)	0.20	9.07	4.35 (2.04)	0.47	7.38

[a] The sample size varies for TTC (TB5: *N*=34; TB6: *N*=45), since collisions (TTC=0 s) are not considered in the calculation.

Table 32. Results of the two-way ANOVA (fatigue condition x time budget) for mean TOT, TTC, AccLong and AccLat.

	Effect	$df_{numerator}$	$df_{denumerator}$	F	p	ω^2
TOT	Fatigue condition	1	83	0.634	0.428	0.000
	Time budget	1	83	10.420	**0.002**	0.099
	Fatigue condition * time budget	1	83	0.158	0.692	0.000
TTC	Fatigue condition	1	75	2.325	0.132	0.017
	Time budget	1	75	0.021	0.885	0.000
	Fatigue condition * time budget	1	75	0.617	0.435	0.000
AccLong	Fatigue condition	1	83	5.236	**0.025**	0.044
	Time budget	1	83	5.577	**0.021**	0.048
	Fatigue condition * time budget	1	83	0.646	0.424	0.008
AccLat	Fatigue condition	1	83	0.268	0.606	0.000
	Time budget	1	83	0.073	0.787	0.000
	Fatigue condition * time budget	1	83	0.023	0.879	0.000

Figure 7-55 and Figure 7-56 display the InRe, FinRe, crash and MC. It is noticeable that 50 % of participants in the FC (TB5: 55 %, TB6: 45 %) accelerated as an initial response to the RtI, while in the NFC the proportion is at 39 % (TB5: 30 %, TB6: 48 %). Furthermore, there is a high proportion of 80 % in the NFC who checked the mirror before performing a maneuver, while in all other groups the proportion is 60 % or lower. Moreover, participants of TB6 (70 %) checked the mirror more frequently before performing a maneuver than participants of TB5 (55 %). FinRe does not show any obvious differences. Crash probability in the FC (13 %) is almost twice as high as in the NFC (7 %), and 3.5 times higher for TB5 (15 %) than for TB6 (4 %).

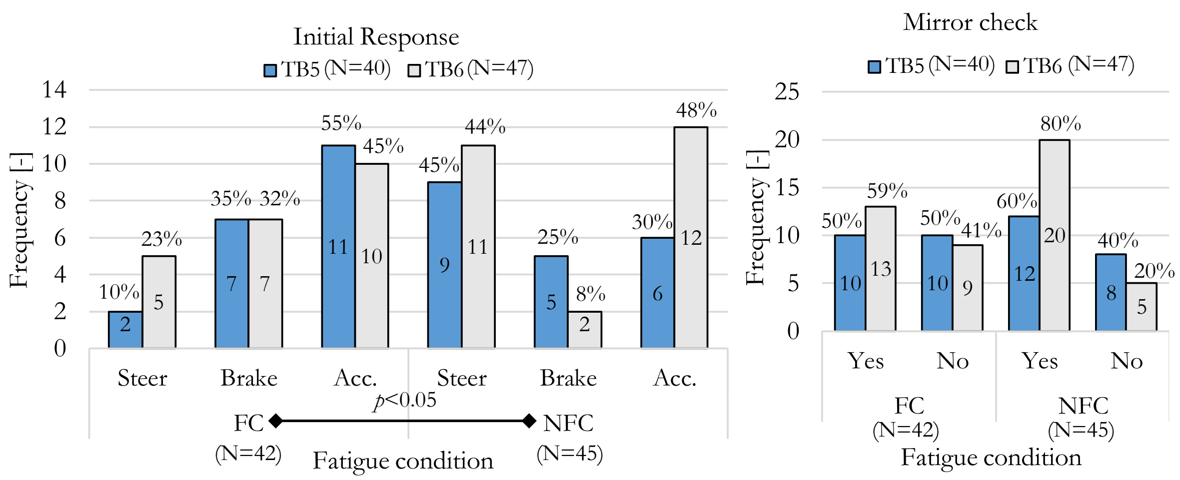


Figure 7-55. Frequency of initial response types (steer, brake, accelerate) and mirror check depending on the fatigue condition and time budget.

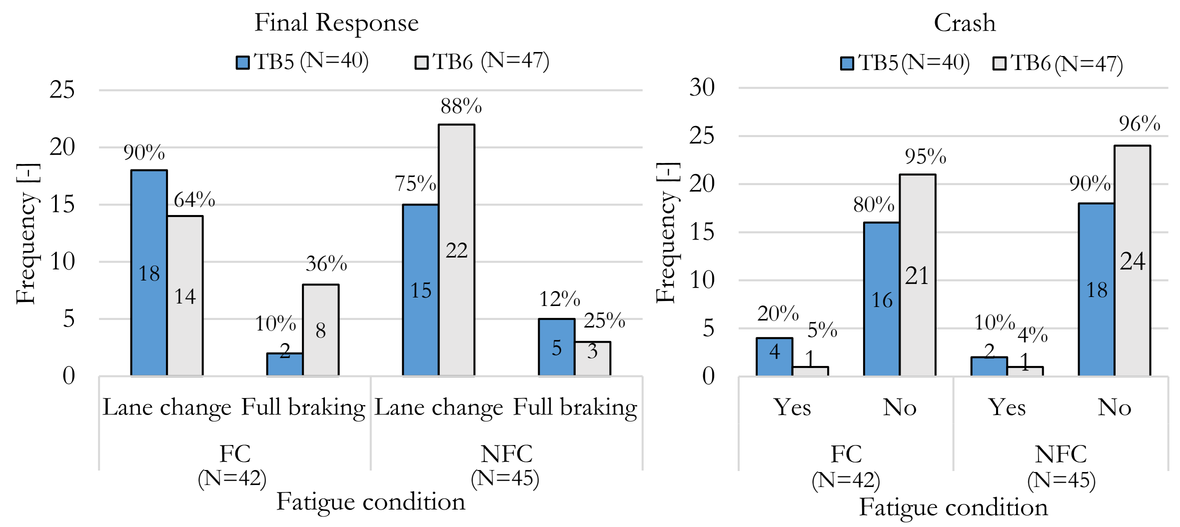


Figure 7-56. Frequency of final response types (lane change, full braking) and crash depending on the fatigue condition and time budget.

To statistically analyze the main effect of time budget and fatigue condition on the bi- and multinomial take-over performance metrics (InRe, FinRe, crash and MC) across the data of experiment 3 and experiment 4, a MLR was conducted. The analysis of an interaction effect of time budget and fatigue condition was considered to not make sense for these metrics, because the sample size of the subgroups was too small. For InRe, results of the MLR show that the full model is a significant improvement in fit over a null model ($\chi^2(4)=10.385$, $p=0.034$). Looking at the parameter estimates (cf. Table 33), fatigue condition is a significant predictor for InRe. B-coefficient and odds ratio Exp(B) indicate that participants in the FC (coded with 3) were less likely to steer than to brake ($p=0.007$) and to accelerate ($p=0.027$) as initial response compared to participants in the NFC (coded with 4). Time budget is not a significant predictor for InRe. For FinRe, crash and MC, results of the MLR show that the full models are not a significant improvement in fit over the null models.

Table 33. Parameter estimates of the MLR with fatigue condition (FC, NFC) and time budget (TB5, TB6) as independent variables and InRe as a dependent variable.

Parameter estimates[a]							
Initial response		***B***	**Std. error**	**Wald**	***df***	***p***	**Exp(B)**
Steer	Intercept	−4.896	2.169	5.094	1	0.024	
	Condition	1.204	0.544	4.894	1	**0.027**	3.333
	Time budget	0.177	0.523	0.050	1	0.823	1.124
Brake	Intercept	2.029	2.105	0.928	1	0.335	
	Condition	−0.539	0.567	0.905	1	0.342	0.583
	Time budget	−0.546	0.550	0.985	1	0.321	0.579

[a] The reference category is accelerate.

Parameter estimates[b]							
Initial response		***B***	**Std. error**	**Wald**	***df***	***p***	**Exp(B)**
Steer	Intercept	−6.924	2.500	7.670	1	0.006	
	Condition	1.743	0.642	7.361	1	**0.007**	5.715
	Time budget	0.663	0.619	1.148	1	0.284	1.940
Acc.	Intercept	−2.029	2.105	0.928	1	0.335	
	Condition	0.539	0.567	0.905	1	0.342	1.715
	Time budget	0.546	0.550	0.985	1	0.321	1.726

[b] The reference category is brake.

Figure 7-57 displays the boxplot of the perceived criticality of the take-over situation depending on the fatigue condition and time budget. On average, participants in the FC rated the criticality of the take-over situation at $M_{\text{criticality rating,FC}}$=7.36 ($SD$=2.43) and participants in NFC at $M_{\text{criticality rating,NFC}}$=6.09 ($SD$=2.53). Participants of TB6 rated the criticality of the take-over situation on average at $M_{\text{criticality rating,TB6}}$=5.96 ($SD$=2.59) and participants with TB5 at $M_{\text{criticality rating,TB5}}$=7.55 ($SD$=2.23). To analyze the main effect of time budget and fatigue condition, the non-parametric Kruskal-Wallis H-test was conducted, since the data are ordinal-scaled. The result reveals that there is a significant main effect of fatigue condition ($\chi^2(1)=5.712$, $p=0.017$) and time budget ($\chi^2(1)=8.813$, $p=0.003$) on the perceived criticality.

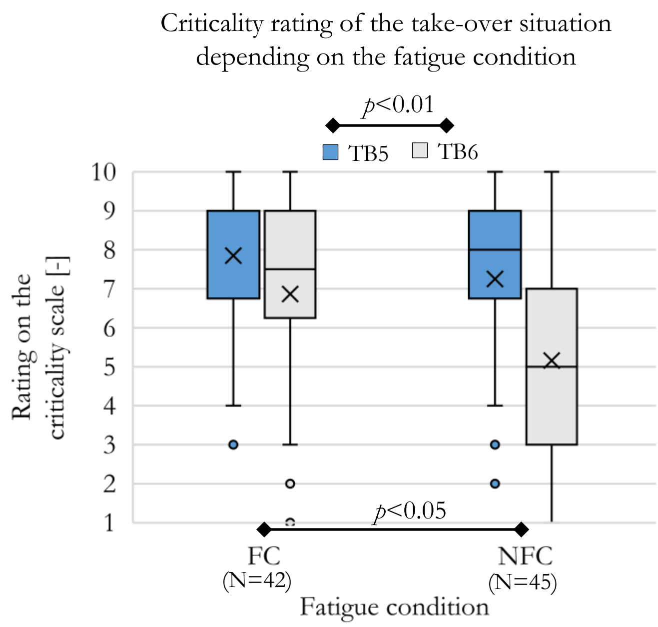


Figure 7-57. Boxplot of the criticality rating depending on the fatigue condition and the time budget.

7.4.3.4 Evaluation of PERCLOS and Comparison to Fatigue Rating

Since no NDRA was available for any of the two fatigue conditions, both conditions had the same preconditions in terms of data availability. While in the CAD90 and FC, experts rated the fatigue level in real-time, in the NFC the rating was done retrospectively. Due to the fatigue-state-dependent study design of the FC, an evaluation of mean PERCLOS per minute—as it was done in experiment 1 and experiment 2—does not make sense, since the sample size of participants for each minute strongly varies. Furthermore, the sample size of PERCLOS values at FL3 or FL4 was rather low due to the study design, since the RtI was triggered and the experimental drive was stopped once participants were at a stable FL3 or FL4. Since results of experiment 3 indicated that PERCLOS availability and validity differed between the fatigue conditions for methodological reasons, a condition-wise analysis was carried out.

For the NFC, 80 PERCLOS values (20 participants x four minutes of CAD) were theoretically available, for the FC, 758 PERCLOS values (20 participants x an average of 37.5 minutes of CAD), and for the CAD90, 708 PERCLOS values (eight participants x an average of 88.5 minutes of CAD). For CAD90, some trials were stopped up to three minutes earlier when the experts were certain that the participants would not reach FL3 and FL4 within 90 minutes due to her/his behavior. Therefore, in total, there were 1546 theoretically available data points for PERCLOS calculation. The data processing for quality reasons described in chapter 7.1.2.3 caused 87 of 1546 possible PERCLOS values to be unavailable, which resulted in an overall data availability of PERCLOS of 94 % (FC: 93 %; NFC: 95 %; CAD90: 95 %).

More insights into the evaluation of validity of PERCLOS is provided by Figure 7-58 and Figure 7-59, which shows the boxplots of all available PERCLOS values depending on the fatigue level and the fatigue condition. Table 34 provides selected descriptive data corresponding to the boxplots. The selected parameters are identical to the ones defined in experiment 1 (see chapter 7.1.3.4). FL4 of the FC, FL2 of NFC and FL3 of the CAD90 are displayed in Table 34 and Figure 7-59, but are not considered in the following evaluation due to the small sample sizes and, consequently, the limited validity.

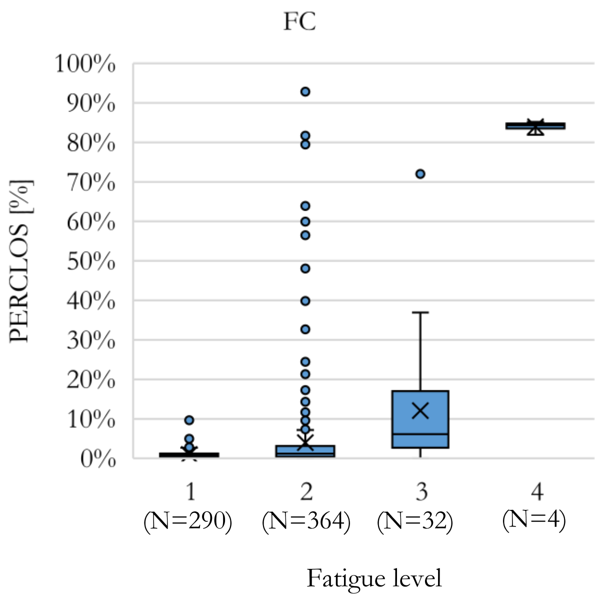


Figure 7-58. Boxplots of all available PERCLOS values of the FC depending on the fatigue level.

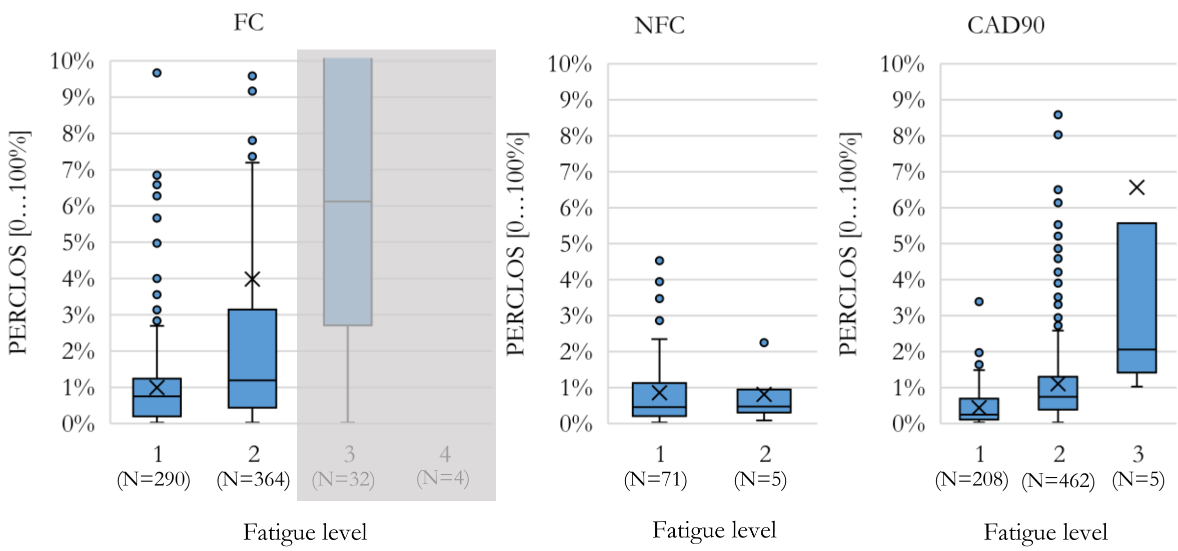


Figure 7-59. Boxplots of all available PERCLOS values depending on the fatigue level and the fatigue condition. For a better readability and comparability of lower FLs, the display range is reduced and for FC (FL3 and FL4) and CAD90 (FL3) some values are cut off.

In the FC, an increase of $M_{PERCLOS}$ and $Mdn_{PERCLOS}$ per fatigue level is notable: $M_{PERCLOS}$ of FL2 is 4 times greater than $M_{PERCLOS}$ of FL1 ($Mdn_{PERCLOS,FL2}$=1.6x$Mdn_{PERCLOS,FL1}$), and $M_{PERCLOS}$ of FL3 is 3 times greater than $M_{PERCLOS}$ of FL2 ($Mdn_{PERCLOS,FL2}$=5.2x$Mdn_{PERCLOS,FL1}$). In the CAD90, $M_{PERCLOS}$ of FL2 is 2.5 times greater than $M_{PERCLOS}$ of FL1 ($Mdn_{PERCLOS,FL2}$=3x$Mdn_{PERCLOS,FL1}$). In the NFC, $M_{PERCLOS}$ and $Mdn_{PERCLOS}$ of FL1 are almost identical to $M_{PERCLOS}$ and $Mdn_{PERCLOS}$ of FL2. When comparing all three conditions and all fatigue levels, $h_{outliers}$ is the lowest at FL1 of CAD90 with 1 % and the highest at FL2 of CAD90 with 10 %. The $IQR_{PERCLOS}$ is the smallest at FL1 of CAD90, even though the $IQR_{PERCLOS}$ at FL1 of all conditions is relatively small. The $IQR_{PERCLOS}$ is the greatest at FL2 of the FC. Furthermore, $h_{outliers7\%}$ is 0 % at FL1 and FL2 of CAD90 and at FL1 of the NFC. In the FC, $h_{outliers7\%}$ is higher with between 1 % (FL1) and 10 % (FL2).

Table 34. Descriptive data of all available PERCLOS values depending on the fatigue level and the fatigue condition.

Fatigue condition		**Fatigue level**			
		1	**2**	**3**	**4**
CAD90	$N_{participants}$	8	8	1	0
	$N_{PERCLOS}$	208	462	5	-
	$M_{PERCLOS}$	0.44 %	1.11 %	6.56 %*	-
	$Mdn_{PERCLOS}$	0.25 %	0.74 %	2.06 %*	-
	$IQR_{PERCLOS}$	0.58 %	0.91 %	4.15 %*	-
	$h_{outliers}$ (Q3+1.5xIQR)	1 %	10 %	20%*	-
	$h_{outliers7\%}$ (>7 %)	0 %	0.4 %	20 %*	-
FC	$N_{participants}$	20	20	20	2
	$N_{PERCLOS}$	290	364	32	4
	$M_{PERCLOS}$	0.99 %	3.98 %	12.10 %	83.94 %*
	$Mdn_{PERCLOS}$	0.75 %	1.19 %	6.13 %	84.33 %*
	$IQR_{PERCLOS}$	1.04 %	2.71 %	14.31 %	1.29 %*
	$h_{outliers}$ (Q3+1.5xIQR)	6 %	9 %	3 %	0 %*
	$h_{outliers7\%}$ (>7 %)	1 %	10 %	50 %	100 %*
NFC	$N_{participants}$	20	3	0	0
	$N_{available\ PERCLOS}$	71	5	-	-
	$M_{PERCLOS}$	0.86 %	0.81 %*	-	-
	$Mdn_{PERCLOS}$	0.46 %	0.47 %*	-	-
	$IQR_{PERCLOS}$	0.92 %	0.64 %*	-	-
	$h_{outliers}$ (Q3+1.5xIQR)	7 %	20 %*	-	-
	$h_{outliers7\%}$ (>7 %)	0 %	0 %*		

Note: Numbers marked with * are not conclusive due to the small sample size.

7.4.4 Discussion

This study replicated the method used in experiment 3 with the same take-over scenario but with a more urgent time budget. It was hypothesized that the results of experiment 3 would be replicated in terms of missing situation awareness and that a smaller time budget would create an even clearer picture of the effect of fatigue on take-over performance, for instance, more crashes in the FC.

As in experiment 3, it could be substantiated by the results of the KSS that the intended fatigue state manipulation in the FC and NFC was successful. All participants in the FC placed themselves between levels seven and nine on the KSS, which is considered to be critical for driving (Platho et al., 2013; Reyner & Horne, 1998), while all participants in the NFC placed themselves below a score of seven. Furthermore, the mean KSS score of the FC was significantly higher than the one of the NFC. In the FC, the mean CAD duration of about 40 minutes was enough to reach FL3 or FL4, while a vast majority was at FL3 when the RtI was triggered. The variance in driving durations was extremely high, ranging from eleven to 73 minutes. All results found regarding the fatigue development are in strong accordance with experiment 3 and other research findings (e.g., Wehlack, 2019).

Regarding take-over performance, the results of this study do not encourage the assumption derived from findings in manual driving and CAD that fatigued drivers have slower take-over times than non-fatigued drivers. This finding is in line with experiment 3 and supports the hypothesis stated in chapter 7.3.4 that the RtI has a strongly stimulating and, therefore, awakening effect on the participants, leading to a quick and startled response. However, in contrast to experiment 3, the systematical braking maneuver showed by fatigued participants was not found in this study. Indeed, the initial response differed significantly between the fatigue conditions, however, in the sense that fatigued participants steered less frequently than expected and non-fatigued participants more frequently. Instead of steering, which was the initial response to the RtI for only 10 % of participants in the FC, they tended to rather accelerate (55 %) than brake (35 %). In contrast, in the NFC, almost half of the participants steered as an initial response. Operating the gas pedal and accelerating in such an already time-critical situation in the first place cannot be considered an adequate reaction, which would rather be braking or steering. This behavior may be attributed to the above-mentioned startle reaction. The strong stimulus induced by the RtI and being confronted by the difficult situation out of a state of great fatigue and probably low situation awareness resulted in a reflexive and thoughtless reaction, namely accelerating. Furthermore, participants in the FC collided with the broken-down vehicle twice as often as participants in the NFC. However, due to the small sample size of crashes in general and the non-significant result, a generalized conclusion cannot be drawn. It can only be assumed that fatigued participants could potentially not cope with their erroneous behavior within the given time, which might be the reason for more crashes. No significant difference in AccLong between fatigued and non-fatigued drivers was found when comparing mean values. However, the distribution of AccLong revealed that 75 % of fatigued participants fall within a range of medium to high AccLong (below $-5\,m/s^2$), whereas in the NFC, the AccLong of 75 % of the participants is greater than the drag torque ($-0.64\,m/s^2$), meaning that 25 % of participants did not brake at all, but resolved the take-over situation only by steering (evasive maneuver). This means that the majority of fatigued participants still showed the distinct, strong decelerating behavior; however, there were also more participants in the NFC in this study who showed strong braking behavior. This can probably be attributed to the smaller time budget in this study. According to the results of Gold (2016), a smaller time budget causes higher AccLong. The overload and the need to react very fast caused participants in the NFC to strongly decelerate in order to gain more time to cope with the highly demanding situation. Results of the criticality rating additionally support this assumption, since there was no significant difference in the perceived criticality between the fatigue conditions. Eventually, the final response was mainly the evasive maneuver for both conditions which is why there was no significant difference.

The above supposed assumption that the smaller time budget of this study caused higher AccLong for participants in the NFC compared to experiment 3 can be confirmed when taking the analysis of take-over performance over both experiments (3 and 4) into account. Here, results showed that a smaller time budget caused significantly higher AccLong. Furthermore, the difference between fatigued and non-fatigued drivers comes into effect again, since a significant difference of AccLong between the fatigue conditions was found. The effect of fatigue condition and time budget was found to be of the same small to medium size. Moreover, as expected before, a medium to large effect of time budget on the TOT was found, namely the less time to react, the faster the intervention. This is in line with many previous research findings (see chapter 3.3.3 or e.g., Gold, 2016, J. Schmidt et al., 2017).

The evaluation of PERCLOS revealed a high data availability of more than 90 % in all three fatigue conditions (FC, NFC and CAD90) which is comparable to the result of all three previous experiments of this thesis considering only conditions without NDRAs. The comparison with the fatigue rating showed a good conformity of PERCLOS with FL1 (assessed by low mean, median, dispersion and false positive rate of PERCLOS) in all three fatigue conditions. As in experiment 3, a good conformity of PERCLOS also with FL2 (assessed by low mean, median, dispersion and false positive rate of PERCLOS) in CAD90 was found. Unlike in experiment 3, the increase of mean and median PERCLOS in the FC from FL2 to FL3 was comparable to the one in experiment 1 and experiment 2 in the UC. The absolute numbers of PERCLOS mean and median at FL3 was higher than in experiment 2 and experiment 3, which is why they are rather comparable to experiment 1. However, the values cannot be allocated clearly to high fatigue levels when comparing to literature (see chapter 7.1.4 or chapter 4.8.2.1). The false positive rate and dispersion of PERCLOS at FL1 and FL2 was in the FC also in a range comparable to experiment 1 and experiment 2. Results of PERCLOS at FL4 are not evaluable, since only four data points are available in this study.

The results suggest a validity of PERCLOS comparable to the one found in experiment 1 and experiment 2 for FL1, FL2 and FL3 when comparing the groups without a NDRA. This does not confirm the findings of experiment 3 that the validity of PERCLOS is lower in terms of fatigue detection at FL3 for the real-time expert rating. A great effort was made to apply the exact same method in this study (for instance mostly the same experts performed the fatigue rating). One possible reason for the differences between experiment 3 and experiment 4 is that the expert rating was documented in a slight different way, not in terms of assigning participants to the respective fatigue level depending on the behavioral indicator, but in terms of setting the point in time on the protocol paper when a participant was at a stable fatigue level and when assigning her/him to the next higher fatigue level. The considerably smaller sample size of PERCLOS values at FL3 in experiment 4 than in experiment 3 also supports this assumption, indicating that raters potentially set the fatigue level to FL3 later on the protocol paper. As a result, the mean PERCLOS of FL3 is high when compared to experiment 1 and experiment 2. The very good consistency of PERCLOS and of the fatigue rating in CAD90 in both experiments may be due to the fact that participants were at a rather stable FL1 or FL2 during (almost) the entire drive. Consequently, the transition phases between FL2 and FL3 were not there, which in experiment 3 were identified as the main reason for uncertainties in PERCLOS (cf. chapter 7.3.4).

Limitations of the Study and Conclusion

A good validity of PERCLOS could be shown in this experiment, especially for FL1. Since great care (strict method and training) was taken to ensure that the exact same method was applied in both studies, the validity and reliability in terms of rating FL3 and FL4 is not questioned. The deviation of results of the PERCLOS evaluation from experiment 3 in terms of validity is rather attributable to a slightly different documentation style of the raters. To some extent, this must be accounted for by the real-time application of the method where a documentation accurate to a second is not always possible. This highlights the dependency of the outcome of such a minute-wise comparison on the ground truth and the relevance of an exact and detailed guideline for documenting the fatigue state.

This study confirms the findings of experiment 3 that TOT is not reduced due to fatigue, attributable probably to the strong stimulating effect of the RtI. Even though the differences between fatigued and non-fatigued participants appear to be less explicit in this more urgent take-over situation, there are still indicators for a lower take-over quality of fatigued participants, namely the high proportion participants who accelerated. Stronger decelerations for fatigued participants were found when analyzing data of both experiments together, supporting the hypothesis of missing situation awareness and lower take-over quality for fatigued drivers. Furthermore, results of the mutual data analysis revealed the strong influencing factor of time budget for the take-over performance. Taking effect sizes into account, time budget seems to be the predominating factor over fatigue condition. This result emphasizes the importance of situational characteristics.

Taking experiment 3 and experiment 4 together, results showed that driver fatigue did not have the expected effect on the TOT. However, an effect of fatigue on single aspects of the take-over quality was found.

8 General Discussion

8.1 The Relevance of Fatigue in Conditionally Automated Driving

The studies presented in this thesis have proven that passive TR fatigue does arise during monotonous phases of CAD. There are two main findings on the development of fatigue: a) it is highly individual whether and when a high level of fatigue is reached and b) even very short phases of CAD may be sufficient to reach a high level of fatigue. The minimum duration with CAD over all studies was six minutes to reach a high level of fatigue, while there was a considerable proportion of participants in all studies that did not reach higher fatigue levels within the given time of 60 to 90 minutes. These findings regarding individual differences in fatigue development are in line with many previous studies on CAD (e.g., J. Schmidt, 2018; Vogelpohl et al., 2018; Wehlack, 2019), but also with research findings on driver fatigue in manual driving (see chapter 4.6). According to the theories described in chapter 4.6 and as discussed in detail in chapter 7.1.4, fatigue vulnerability depends on personality (traits) and the transaction between a task/the environment and the personality. This means that there is probably an individual predisposition to become fatigued quickly and that this is also affected by the task or the environment an individual is confronted with. Relevant personality traits are mainly unknown and were not assessed in this thesis. However, it is hypothesized that age might be one. Theories on different coping strategies point out that motivation and interest in performing a task are one crucial influencing factor for countering fatigue (see chapter 4.6). It might play a large role whether a driver is willing—from a motivational perspective—as well as she/he is able—from a resource perspective—to apply compensatory effort and to cope with upcoming fatigue. If both is the case, a driver might resist underload conditions as they were applied in the studies of this thesis and will not become fatigued over a prolonged period of time. Motivation and stimulation can also be provided by NDRAs, which was to be addressed in this thesis by offering naturalistic NDRAs (RQ3) in order to display a realistic scenario for future CAD. Results showed that NDRAs were effective countermeasures for the development of passive TR fatigue; however, they could not prevent fatigue for all participants. Interestingly, voluntary NDRAs were not as effective against fatigue as a video game. Results suggest that individual motivation of participants to play the game was higher than engaging in the available free choice of NDRAs (see chapter 7.2.4). These results again emphasize the importance of the personal preferences for a task and also the strong individual differences in fatigue vulnerability. Therefore, it might be possible to counteract fatigue for a certain proportion of drivers, but there will always be drivers who cannot be stimulated by NDRAs, because the effective ones are not available in the vehicle or—which might be the more likely circumstance—because participants are not willing to counteract their fatigue. These drivers will always bear an increased risk to get into a critical fatigue state and to fall asleep while driving in a conditionally automated mode. Even though the effect of automation duration could not be verified explicitly in the studies of this thesis (RQ2), it is still assumed that there is a correlation between fatigue development and automation duration (as it was already proposed for the factor time-on-task, cf. chapter 4.4). The strength of the effect probably depends—among others—on the individual traits, for instance, how vulnerable the person is in general to fatigue, the daily constitution of a person (due to the circadian rhythm, hours of sleep, etc.), the possibility to apply their individual compensatory strategy (individual NDRA) and the willingness to fight fatigue. The latter might be a very important factor, especially with regard to passive TR fatigue, because for this fatigue type the chance to fight is much higher (the cause can just be removed in the best case) than for SR fatigue where, above all, sleep is the most effective countermeasure (Williamson, 2012; cf. chapter 4.4 and 4.7).

How does fatigue which could be proven to result from prolonged durations with CAD affect take-over performance (RQ1)?

Results of the studies of this thesis did not reveal the clear picture that had been expected beforehand due to research findings on manual driving. Slower reaction times due to slowed information processing and an increased decision-making time are probably the most prominent hypothesized effects of fatigue when driving and have been found in previous studies on manual driving (see chapter 4.5). However, slower TOT could not be confirmed in this thesis, which is in accordance with similar studies (as discussed in chapter 7.3.4 and 7.4.4). The salient and strongly stimulating RtI probably had an awakening effect and caused a quick and more of a startle reaction in the participants. Instead, a deterioration of take-over quality was found for fatigued drivers (undeliberated maneuvers such as strong braking and/or accelerating at first), which did not lead to a significant increase in accidents, though. Fatigue limits both visual attention due to prolonged eye closure durations and cognitive attention due to slowed information processing or mind wandering ("driving without awareness" according to Brown (1994, p. 312), or also known as *look but failed to see* problem). Both result in reduced or missing situation awareness. Additionally, a worse recall of knowledge-based behavior is assumed, which is necessary in such unfamiliar situations. Therefore, the most safety-relevant aspect of passive TR fatigue in CAD is probably the risk of decreased cognitive and visual attention prior to the RtI and corresponding undeliberated actions.

Considering the findings of previous research on manual driving (see chapter 4.5) and on CAD (see chapter 4.9) as well as the detailed effects found in this thesis, it can be concluded that high fatigue levels (when microsleep or prolonged eyelid closure already occur) are never beneficial for any driving performance. Therefore, high fatigue levels should be avoided in CAD, since the readiness to perform a safe take-over is not guaranteed and the step to falling asleep is only small. Sleeping is explicitly prohibited during CAD (SAE International, 2018). This recommendation is in line with findings of Radlmayr, Feldhütter, et al. (2019) who consolidated research results on fatigue in CAD of the KoHAF project.

8.2 Implications for System Design

From the findings of this thesis several implications can be derived for the design of systems with CA.

Since the system will not know what type of driver (vulnerable to fatigue) and in which state (after a night shift) she/he enters the vehicle, limiting CAD to a fixed duration appears to be unrewarding, which is in line with the recommendation of Wehlack (2019). Instead, a driver-monitoring system seems to be essential to cover those drivers who are at risk of exhibiting high levels of fatigue or even of falling asleep.

The PERCLOS metric was examined intensely in this thesis in terms of its suitability for assessing fatigue in CAD in a realistic setup (RQ4). Results show that PERCLOS is especially distinctive for stable and also rather extreme fatigue states. In this thesis, these states were represented by FL1—when participants were not fatigued at all—and FL4—when participants closed their eyes for many seconds and were close to falling asleep. False negatives in terms of fatigue detection mainly occurred due to participants who put a great effort into keeping their eyes open despite strong fatigue. False positives in terms of fatigue detection occurred due to the high sensitivity of PERCLOS, which cannot take into account slight fluctuation between fatigue levels or physiological variations, such as dry eyes (causing high blink rates and prolonged blink durations) or narrowing the eyes to slits due to glaring sunlight. Especially the physiological factor caused a lot of false positives when the participants were engaged in NDRAs because they were looking down at the item in their hands. It is recommended to include more metrics in the fatigue detection algorithm, for instance, head pose or body movements to make fatigue detection more stable and less vulnerable to such variations. A greater evaluation time frames for PERCLOS calculation could have a positive effect on robustness and stability, as well. The time frame needs to be a compromise between short enough, to be able to detect a real change of the fatigue state to a high level, and long enough, to make fatigue detection more robust. A further finding of this thesis is that PERCLOS strongly depends on the data quality and availability. The system and camera set-up used in the studies of this thesis produced very good results when drivers were not engaged in a NDRA and mainly looked at the road. This is logical, since the focus of this system is on gaze tracking towards the road in manual driving and, therefore, the camera is mounted above the instrument cluster. Data availability for drivers engaging in NDRAs, and thus PERCLOS validity, was very limited and was mainly caused by the limited field of vision of the cameras. A better result can probably be achieved through the implementation of more cameras with different viewing angles. This would also further improve data availability for drivers without a NDRA, since those drivers also adapted their position in the seat, which has an adverse effect on the camera system when driving in conditionally automated mode. Even though this finding was achieved by a specific eye tracking system, these results can be transferred to systems with a similar camera set-up.

Furthermore, results of this thesis suggest that a RtI designed in a sufficiently salient way has the potential to break driver's fatigue and to cause rather quick reactions, even though drivers are fatigued. However, such a startle reaction always poses the risk that it is not as desired and leads to a risky driving behavior, which was partly the case in the studies of this thesis (see chapter 7.3.4 and chapter 7.4.4). In alignment with the suggestions of Eskandarian et al. (2007), a RtI should be designed as smoothly and predictably as possible, but also as saliently and urgently as required by the situation. This applies to fatigued but also to non-fatigued drivers. It might be beneficial to adapt the RtI for fatigued drivers to make it even more salient. However, the trade-off between awakening and startling the driver will be challenging.

Low situation awareness regardless of the reason (being fatigued or distracted by a NDRA) can have an adverse effect on take-over performance, as was shown in the studies of this thesis. Since NDRAs may counteract fatigue and are the key advantage of CAD, reduced or missing situation awareness is not only a fatigue-specific issue but needs to be addressed in CAD in any case. A reasonable integration of NDRAs in the system should be enforced, aiming at guiding the user's attention, for instance, by using only onboard terminals or mobile terminals with a coupling to the CA. In case of certain traffic situations which are known to potentially lead to short-term RtI, a RtM may be beneficial (cf. chapter 7.2.4) to prepare the driver for a potential take-over. Even though this was not the primary goal of this thesis, the comparison of the two time budgets in experiment 3 and experiment 4 emphasizes the importance of situational factors for take-over performance (see chapter 3.3.3 and Gold, 2016). Small time budgets are one of the most critical factors for a safe take-over and should be avoided. For the cases in which they cannot be avoided, the system should provide maximum support to resolve the situation, for instance, by further providing steering or braking assistance. Furthermore, to compensate for missing situation awareness, a maximum support for the driver during the take-over phase is to be provided on the part of both the system and the HMI, to build up situation awareness and to prevent collisions. This can be done, for instance, by displaying situation-relevant information and giving guidance for a specific action or by increasing the time budget by conducting a system-initiated braking maneuver.

Moreover, it is supposed due to different theories (cf. chapter 7.1.4) that the willingness to put effort into fighting fatigue might play a role in preventing this state. Therefore, the risk of fatigued driving, especially also with CA, has to be emphasized to the driver to increase the willingness to fight against fatigue or to stop the drive in a timely manner. This can be addressed by properly educating the users of such systems and by sufficiently pointing out the risks of falling asleep. This also includes the marketing messages of CAD to raise users' awareness that a high level of fatigue may reduce their take-over ability and that sleep is prohibited. The more enlightened the user, the fewer errors will occur in the usage of the system.

8.3 Transferability of the Driving Simulator Studies to Real Road Traffic

A driving simulator is a great and, to some extent, the only opportunity to conduct tests that might result in dangerous situations. Examining the take-over performance of fatigued drivers when assuming a negative effect of fatigue in real road traffic and jeopardizing people is unethical and the risk would not be acceptable. Nevertheless, a certain degree of uncertainty remains regarding the transferability of the results found in the driving simulator to real road traffic because of several limitations of a simulator. For the simulator used for the experiments of this thesis, limitations include the absence of a kinesthetic feedback channel for the dynamics of the vehicle (static driving simulator) and a not completely realistic steering and braking behavior. Therefore, absolute values of accelerations cannot be transferred to reality, whereas the relative comparison between participants can, as was done for AccLong and AccLat in this thesis. It is assumed that the principle findings of fatigue effects on take-over performance, for instance, reactions types or TOT, can be transferred to real road situations, since they result from cognitive variations which are assumed to be mainly independent from driving simulator effects.

Findings on the absolute temporal development of fatigue may deviate in real road traffic due to increased trust and missing risk awareness in a driving simulator. However, it can be expected that after experiencing a very reliable CA for a while, the same trust will ensue. Therefore, the temporal development found in the studies of this thesis can give an indication of fatigue development in CAD after a certain habituation phase. Furthermore, first driving studies under real traffic conditions with a wizard-of-oz approach show that fatigue can occur there as fast as in driving simulator studies (Jarosch, Paradies, et al., 2019).

8.4 Limitations and Future Work

In the experiments of this thesis, participants' SR fatigue/sleepiness was not directly manipulated (e.g., by sleep deprivation), but they were focused on inducing passive TR fatigue by the task characteristics (monotonous drive with less stimulations). However, the sleep duration of the participants was not explicitly controlled and the circadian rhythm was utilized in the sense that for experiment 2, experiment 3 and experiment 4, the slots for the fatigued and underload condition, respectively, took place in the early morning or in the early afternoon after lunch in order to accelerate the fatigue onset. Therefore, it cannot be ruled out completely that SR fatigue sometimes played a more and sometimes a less significant role in the experiments depending on the individual circadian rhythm of the test person. However, as already discussed in chapter 4.2 and chapter 4.5, both constructs will be relevant in CAD and the effects might be similar regarding task performance. The results of the experiments in terms of take-over performance can probably not be allocated exclusively to passive TR fatigue, but that does not make the results any less meaningful. In the future scenario of CAD, it is more likely that users will decide to use their car even though they are not in a proper condition in terms of fatigue, since they do not have to drive on their own. A study of Vogelpohl et al. (2018) suggests that SR fatigue evolved significantly faster, indicating that it is even more difficult to resist. Furthermore, Vogelpohl et al. (2018) found indicators for slowed response times to a RtI, which some findings from manual driving also point towards (Phillips, 2014). Therefore, in future research, the effect of SR fatigue (sleep deprivation) should be examined in a similar way as it was done in this thesis for passive TR fatigue. This also applies to the effect of motivating NDRAs that proved to counteract the onset of passive TR fatigue. The validity for SR fatigue needs to be examined, as well, since no conclusion can be drawn as to whether motivating NDRAs would also counteract a high SR fatigue level in the same way. The prerequisite would be that drivers of a high SR fatigue level would engage in a self-chosen NDRA, and it is questionable whether such drivers would do so. Therefore, motivating NDRAs can probably not be considered a general countermeasure for any type of fatigue, which supports the hypothesis of May and Baldwin (2009) that the cause of fatigue is the key to preventing and counteracting it.

As mentioned in chapter 4.1, driver state is a very complex construct and so far, not all relevant factors for CAD have been identified. In this thesis, the focus was on the driver state factor of fatigue, and great effort was expended to isolate this state. However, there are other driver state components that are hypothesized by more and more authors to be important influencing factors on take-over performance, maybe even more relevant for take-over performance than passive TR fatigue. Such components are, for example, personal reaction times, the individual driving skills or individual take-over skills (Körber, 2018; J. Schmidt et al., 2017). In this thesis, it was not possible to examine these factors due to the between-subjects designs in which each participant only experienced one take-over. This was chosen on purpose to exclude training effects between take-overs completely. In future studies, a within-subjects design is recommended to be able to further examine individual characteristics assumed to be relevant for take-over performance. Personality traits were also hypothesized to be relevant in terms of fatigue vulnerability. It is recommended to address these traits in future research, also with respect to intercorrelation with NDRAs, to be able to adjust and optimize the system design.

In this thesis, some weaknesses of PERCLOS were revealed. As J. Schmidt (2018) stated, the detection rate of eyelid closure metrics including PERCLOS strongly depends on the recording method, the sampling rate and the algorithm for signal processing. Putting it in a more generalized way, a metric or algorithm to assess fatigue can only be as good as the sensors and data processing algorithm on which the metric is based (Karrer-Gauß, 2011). In the studies of this thesis, these parameters were not varied or controlled, since the same system was used at all times. Therefore, further research should work on an optimized combination of eye tracking, metrics used and processing algorithm to be able to detect high fatigue states. Ideally, pre-stages of fatigue should be detected already to improve safety, comfort and flexibility. In the light of the findings of this thesis, PERCLOS might have some weaknesses for some users with regard to detecting such rather early stages of fatigue. Therefore, further research needs to be done to find suitable additional metrics. Nevertheless, a certain correlation of fatigue development and driving duration is still supposed, even if this was not found in this thesis. It can probably not be attributed to a specific driving duration, but it rather also depends on personality and physiological form of the day. Therefore, an algorithm may include driving time with CA to some extent, in addition to driver-monitoring data. To which extent this is to be the case should be the object of future research.

9 Summary

Based on an extensive literature review on the effect of fatigue in manual driving, it was expected that a deterioration of take-over performance during CAD will occur due to fatigue, and thus, that fatigue is relevant for safety in CAD. Four main research questions were derived regarding the effect of fatigue on take-over performance, the development of fatigue during CAD, NDRAs as a countermeasure for fatigue and the PERCLOS metric as an assessment method for fatigue. These were addressed with the help of four experiments in this thesis.

Results of the studies showed that fatigue development is highly individual regarding whether and when a high level of fatigue is reached, and that very short phases of CAD are already sufficient to reach a high level of fatigue (close to falling asleep). Motivating NDRAs may counter the onset of high levels of fatigue; however, the success strongly depends on the individual, suggesting that personality (e.g., the personal motivation for an activity or personal fatigue vulnerability) has a great influence. The effect of fatigue on take-over performance was not as clear as expected from literature review. Fatigued drivers did not show slower TOT. Instead, the effect rather manifested in a reduced take-over quality: undeliberate maneuvers, such as accelerating in already time-critical situations or strong decelerations. Furthermore, fatigued participants frequently showed startle reactions and perceived the take-over as more critical, indicating a less confident behavior. Accidents occurred more frequently for fatigued participants, however, with a low occurrence in general, which is why no significant differences between fatigued and non-fatigued drivers occurred. Overall, the effect of fatigue does not give a clear indication in terms of the safety-relevance of fatigue in CAD. However, due to its numerous effects found in previous research and in the studies of this thesis, it is concluded that fatigue will never be beneficial for any driving performance. Therefore, it is recommended to avoid high levels of fatigue (close to falling asleep) during CAD, since a safe take-over cannot be guaranteed.

Based on the findings of this thesis, further implications for system design were derived: In order to avoid high fatigue levels, a fatigue monitoring system is recommended during CAD. PERCLOS was found to be distinctive for extreme fatigue states, being not fatigued at all on the one hand, and being close to falling asleep at the other hand. To achieve more robustness and reliability for the fatigue monitoring system, multiple metrics should be fused. Additionally, the importance of sufficient data availability was emphasized. The importance of a well-balanced RtI which stimulates the driver but does not startle her/him was highlighted. Furthermore, it is recommended to provide situation-specific maximum support for the driver in lateral and longitudinal guidance during the take-over to compensate missing situation awareness, which was found in this thesis to accompany fatigue, but which is also a general issue in CAD.

References

Ackerman, P. L., Calderwood, C., & Conklin, E. M. (2012). Task Characteristics and Fatigue. In G. Matthews, P. A. Desmond, C. Neubauer, & P. A. Hancock (Eds.), *The Handbook of Operator Fatigue* (pp. 91–101). Burlington, VT: Ashgate Pub. Company.

Aeberhard, M. (2017). *Object-Level Fusion for Surround Environment Perception in Automated Driving Applications* (Dissertation). Technischer Universität Dortmund, Dortmund.

Åkerstedt, T., & Gillberg, M. (1990). Subjective and Objective Sleepiness in the Active Individual. *International Journal of Neuroscience*, *52*(1-2), 29–37. https://doi.org/10.3109/00207459008994241

Anund, A., Fors, C., Hallvig, D., Åkerstedt, T., & Kecklund, G. (2013). Observer rated sleepiness and real road driving: An explorative study. *PloS One*, *8*(5), e64782. https://doi.org/10.1371/journal.pone.0064782

Aptiv, Audi, Baidu, BMW, Continental, Daimler, . . . Volkswagen (2019). *Saftey First For Automated Driving*. Retrieved from https://www.daimler.com/innovation/case/autonomous/safety-first-for-automated-driving.html

Atchley, P., Chan, M., & Gregersen, S. (2014). A Strategically Timed Verbal Task Improves Performance and Neurophysiological Alertness During Fatiguing Drives. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *56*(3), 453–462. https://doi.org/10.1177/0018720813500305

Bainbridge, L. (1983). Ironies of automation. *Automatica*, *19*(6), 775–779. https://doi.org/10.1016/0005-1098(83)90046-8

Baulk, S. D., Reyner, L. A., & Horne, J. A. (2001). Driver sleepiness-evaluation of reaction time measurement as a secondary task. *Sleep*, *24*(6), 695–698. https://doi.org/10.1093/sleep/24.6.695

Befelein, D., Boschet, J., & Neukum, A. (2018). Influence of non-driving-related tasks' motivational aspects and interruption effort on driver take-over performance in conditionally automated driving. In T. Victor, M. P. Bruyas, M. Regan, C. Brusque, A. Fort, & C. Jallais (Chairs), *6th Driver Distraction and Inattention conference,* Gothenburg, Sweden.

Belz, S. M. (2000). *An On-road Investigation of Self-rating of Alertness and Temporal Separation as Indicators of Driver Fatigue in Commercial Motor Vehicle Operators* (Dissertation). Virginia Polytechnic Institute and State University, Blacksburg, VA, USA.

Bergasa, L. M., Nuevo, J., Sotelo, M. A., Barea, R., & Lopez, E. (2008). Visual Monitoring of Driver Inattention. In J. Kacprzyk & D. Prokhorov (Eds.), *Studies in Computational Intelligence. Computational Intelligence in Automotive Applications* (Vol. 132, pp. 19–37). Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-79257-4_2

Bittner, R., Hána, K., Poušek, L., Smrka, P., Schreib, P., & Vysoký, P. (2000). Detecting of Fatigue States of a Car Driver. In R. W. Brause & E. Hanisch (Eds.), *Lecture Notes in Computer Science: Vol. 1933. Medical Data Analysis: First International Symposium, ISMDA 2000 Frankfurt, Germany, September 29-30, 2000 Proceedings* (1st ed., pp. 260–273). Berlin, Heidelberg: Springer Berlin Heidelberg; Imprint: Springer.

BMW Group (2013, May 6). Attention Assistant. Retrieved from https://www.press.bmwgroup.com/global/article/detail/T0141144EN/bmw-model-upgrade-measures-taking-effect-from-the-summer-of-2013

Bosch Group (2019, December 12). Driver Drowsiness Detection. Retrieved from https://www.bosch-mobility-solutions.com/en/products-and-services/passenger-cars-and-light-commercial-vehicles/driver-assistance-systems/driver-drowsiness-detection/

Bosch Group (2020, December 8). Interior Monitoring Systems. Retrieved from https://www.bosch-mobility-solutions.com/en/products-and-services/passenger-cars-and-light-commercial-vehicles/interior-and-body-systems/interior-monitoring-systems/

Bourrelly, A., Jacobé de Naurois, C., Zran, A., Rampillon, F., Vercher, J.-L., & Bourdin, C. (2019). Long automated driving phase affects take-over performance. *IET Intelligent Transport Systems*, *13*(8), 1249–1255. https://doi.org/10.1049/iet-its.2019.0018

Boverie, S., Rodriguez, N., & Bande, D. Saccagno, A. (2013). *General driver monitoring module definition SoA*. Deliverable D3.2.1.

Bowman, D., Hanowski, R. J., Alden, A., Gupta, S., Wiegand, D. M., Baker, S., . . . Wierwille, W. W. (December 2012). *Development and Assessment of a Driver Drowsiness Monitoring System*. Technical Report FMCSA-RRR-12-008. Blacksburg, VA, USA.

Brown, I. D. (1994). Driver Fatigue. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *36*(2), 298–314. https://doi.org/10.1177/001872089403600210

Bubb, H. (2015a). Einführung. In H. Bubb, K. Bengler, R. E. Grünen, & M. Vollrath (Eds.), *Automobilergonomie* (pp. 1–25). Wiesbaden: Springer Fachmedien Wiesbaden.

Bubb, H. (2015b). Der Mensch als Fahrer. In H. Bubb, K. Bengler, R. E. Grünen, & M. Vollrath (Eds.), *Automobilergonomie* (pp. 67–158). Wiesbaden: Springer Fachmedien Wiesbaden.

Bubb, H., Bengler, K., Lange, C., Aringer, C., & Trübswetter, N. (2015). Messmethoden. In H. Bubb, K. Bengler, R. E. Grünen, & M. Vollrath (Eds.), *Automobilergonomie* (pp. 617–659). Wiesbaden: Springer Fachmedien Wiesbaden.

Buehler, M., Iagnemma, K., & Singh, S. (Eds.) (2007). *Springer tracts in advanced robotics: Vol. 36*. *Darpa Grand Challenge: The Race of the Century*. Berlin, Heidelberg: Springer-Verlag Berlin Heidelberg. Retrieved from http://dx.doi.org/10.1007/978-3-540-73429-1 https://doi.org/10.1007/978-3-540-73429-1

Buehler, M., Iagnemma, K., & Singh, S. (Eds.) (2009). *Springer tracts in advanced robotics: Vol. 56*. *The DARPA urban challenge: Autonomous vehicles in city traffic*. Berlin: Springer.

Bühner, M., & Ziegler, M. (2017). *Statistik für Psychologen und Sozialwissenschaftler* (2., aktualisierte und erweiterte Auflage). Hallbergmoos: Pearson. Retrieved from http://lib.myilibrary.com?id=1003080

Cabrall, C., Janssen, N., Gonçalves, J., Morando, A., Sassman, M., & de Winter, J. C. F. (2016). Eye-Based Driver State Monitor of Distraction, Drowsiness, and Cognitive Load for Transitions of Control in Automated Driving. In *2016 IEEE International Conference on Systems, Man, and Cybernetics: Conference proceedings* (pp. 1981–1982). Piscataway, NJ: IEEE.

Cluydts, R., Valck, E. de, Verstraeten, E., & Theys, P. (2002). Daytime sleepiness and its evaluation. *Sleep Medicine Reviews*, *6*(2), 83–96. https://doi.org/10.1053/smrv.2002.0191

Cohen, R. A. (2011a). Arousal. In J. S. Kreutzer (Ed.), *Encyclopedia of clinical neuropsychology* (pp. 247–249). New York, NY: Springer. https://doi.org/10.1007/978-0-387-79948-3_1266

Cohen, R. A. (2011b). Yerkes–Dodson Law. In J. S. Kreutzer (Ed.), *Encyclopedia of clinical neuropsychology* (pp. 2737–2738). New York, NY: Springer. https://doi.org/10.1007/978-0-387-79948-3_1340

Connor, J., Norton, R., Ameratunga, S., Robinson, E., Civil, I., Dunn, R., . . . Jackson, R. (2002). Driver sleepiness and risk of serious injury to car occupants: population based case control study. *British Medical Journal*, *324*(7346).

Daimler AG (2010, June 1). ATTENTION ASSIST: Drowsiness-detection system warns drivers to prevent them falling asleep momentarily. Retrieved from https://media.daimler.com/marsMediaSite/de/instance/ko/Der-ATTENTION-ASSIST-und-die-Konditionssicherheit-Mach-mal-Pause.xhtml?oid=9361847

Damböck, D. (2013). *Automationseffekte im Fahrzeug – von der Reaktion zur Übernahme* (Dissertation). Technical University of Munich, Munich.

Damböck, D., Farid, M., Tönert, L., & Bengler, K. (2012). Übernahmezeiten beim hochautomatisierten Fahren. *5. Tagung Fahrerassistenz*. Retrieved from http://mediatum.ub.tum.de/doc/1142102/file.pdf

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. https://doi.org/10.1016/j.trf.2014.06.016

Dement, W. C. [William C.], & Carskadon, M. A. (1982). Current perspectives on daytime sleepiness: The issues. *Sleep*, *5 Suppl 2*, S56-66. https://doi.org/10.1093/sleep/5.s2.s56

Desmond, P. A., & Hancock, P. A. (2001). Active and passive fatigue states. In P. A. Hancock & P. A. Desmond (Eds.), *Human factors in transportation series. Stress, workload, and fatigue* (pp. 455–465). Mahwah NJ u.a.: Erlbaum Ass.

Desmond, P. A., & Matthews, G. (2009). Individual differences in stress and fatigue in two field studies of driving. *Transportation Research Part F: Traffic Psychology and Behaviour*, *12*(4), 265–276. https://doi.org/10.1016/j.trf.2008.12.006

Deutscher Bundestag (2017, March 29). *Regelungen zum Fahren von Autos mit hoch- und vollautomatiiserten Fahrfunktionen zur Änderung des Straßenverkehrsgesetzes: Drucksache 18/11776*.

Dinges, D. F. (1995). An overview of sleepiness and accidents. *Journal of Sleep Research*, *4*(S2), 4–14.

Donges, E. (2015). Fahrerverhaltensmodelle. In H. Winner, S. Hakuli, F. Lotz, & C. Singer (Eds.), *Handbuch Fahrerassistenzsysteme: Grundlagen, Komponenten und Systeme für aktive Sicherheit und Komfort* (3rd ed.). Wiesbaden: Springer Fachmedien Wiesbaden.

Ebrahim, P. (2016). *Driver drowsiness monitoring using eye movement features derived from electrooculography* (Dissertation). Universität Stuttgart, Stuttgart.

Endsley, M. R. (1995). Toward a Theory of Situation Awareness in Dynamic Systems. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *37*(1), 32–64. https://doi.org/10.1518/001872095779049543

Eskandarian, A., Sayed, R., Delaigue, P., Blum, J., & Mortazavi, A. (2007). *Advanced Driver Fatigue Research*. Washington, D.C.

Euro NCAP (September 2017). *Euro NCAP 2025 Roadmap: In Pursuit of Vision Zero*. Technical Paper. Brussels. Retrieved from https://cdn.euroncap.com/media/30700/euroncap-roadmap-2025-v4.pdf

European Commission (2019, April 4). *2018 road safety statistics: What is behind the figures?* Retrieved from https://europa.eu/rapid/press-release_MEMO-19-1990_en.htm

Feierle, A. (2017). *Konzeption und Implementierung einer Fahrermüdigkeitserkennung und zeitlich variabler Übernahmeszenarien für das hochautomatisierte Fahren* (Unpublished master's thesis). Technical University of Munich, Munich.

Feldhütter, A., Feierle, A., Kalb, L., & Bengler, K. (2018). A New Approach for a Real-Time Non-Invasive Fatigue Assessment System for Automated Driving. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *62*(1), 1669–1673. https://doi.org/10.1177/1541931218621379

Feldhütter, A., Gold, C., Schneider, S., & Bengler, K. (2017). How the Duration of Automated Driving Influences Take-Over Performance and Gaze Behavior. In C. Schlick, S. Duckwitz, F. Flemisch, M. Frenz, S. Kuz, A. Mertens, & S. Mütze-Niewöhner (Eds.), *Advances in Ergonomic Design of Systems, Products and Processes: Proceedings of the Annual Meeting of GfA 2016* (1st ed., pp. 309–318). Berlin: Springer. https://doi.org/10.1007/978-3-662-53305-5_22

Feldhütter, A., Härtwig, N., Kurpiers, C., Mejia Hernandez, J., & Bengler, K. (2019). Effect on Mode Awareness When Changing from Conditionally to Partially Automated Driving. In S. Bagnara, R. Tartaglia, S. Albolino, T. Alexander, & Y. Fujita (Eds.), *Proceedings of the 20th Congress of the International Ergonomics Association (IEA 2018)* (pp. 314–324). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-96074-6_34

Feldhütter, A., Hecht, T., Kalb, L., & Bengler, K. (2019). Effect of prolonged periods of conditionally automated driving on the development of fatigue: With and without non-driving-related activities. *Cognition, Technology & Work*, *21*(1), 33–40. https://doi.org/10.1007/s10111-018-0524-9

Feldhütter, A., Kroll, D., & Bengler, K. (2018). Wake Up and Take Over! The Effect of Fatigue on the Take-over Performance in Conditionally Automated Driving. In *2018 21st International Conference on Intelligent Transportation Systems (ITSC)* (pp. 2080–2085). IEEE. https://doi.org/10.1109/ITSC.2018.8569545

Feldhütter, A., Ruhl, A., Feierle, A., & Bengler, K. (2019). The Effect of Fatigue on Take-over Performance in Urgent Situations in Conditionally Automated Driving. In *2019 IEEE Intelligent Transportation Systems Conference (ITSC)* (pp. 1889–1894). IEEE. https://doi.org/10.1109/ITSC.2019.8917183

Feldhütter, A., Segler, C., & Bengler, K. (2018). Does Shifting Between Conditionally and Partially Automated Driving Lead to a Loss of Mode Awareness? In N. A. Stanton, S. Landry, G. Di Bucchianico, & A. Vallicelli (Eds.), *Advances in Intelligent Systems and Computing: Vol. 597, Advances in Human Aspects of Transportation: Proceedings of the AHFE 2017 International Conference on Human Factors in Transportation, July 17−21, 2017, The Westin Bonaventure Hotel, Los Angeles, California, USA* (pp. 730–741). Cham: Springer. https://doi.org/10.1007/978-3-319-60441-1_70

Ford Motor Company. Driver Alert System. Retrieved from https://www.haynesford.co.uk/Tech-Ford-driver-alert

Friedrichs, F., & Yang, B. (2010). Camera-based drowsiness reference for driver state classification under real driving conditions. In Institute of Electrical and Electronics Engineers (Ed.), *2010 IEEE Intelligent Vehicles Symposium: Iv; 21 - 24 June 2010, University of California, San Diego, CA, USA* (pp. 101–106). Piscataway, NJ: IEEE. https://doi.org/10.1109/IVS.2010.5548039

Funke, G. J., Matthews, G., Warm, J. S., Emo, A., & Fellner, A. N. (2016). The Influence of Driver Stress, Partial-Vehicle Automation, and Subjective State on Driver Performance. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *49*(10), 936–940. https://doi.org/10.1177/154193120504901014

Galley, N. [Niels], Schleicher, R. [Robert], & Galley, L. A. (2003). Blink Parameter as Indicators of Driver's Sleepiness - Possibilities and Limitations. Retrieved from https://pdfs.semanticscholar.org/020e/c5b82c024d61e4abbae0935cc40123ec3fb6.pdf

Garcia, I., Bronte, S., Bergasa, L. M., Hernandez, N., Delgado, B., & Sevillano, M. (2010). Vision-based drowsiness detector for a realistic driving simulator. In *13th International IEEE Conference on Intelligent Transportation Systems (ITSC), 2010: 19 - 22 Sept. 2010, Funchal, Madeira Island, Portugal* (pp. 887–894). Piscataway, NJ: IEEE. https://doi.org/10.1109/ITSC.2010.5625097

Geiser, G. (1985). Mensch-Maschine-Kommunikation im Kraftfahrzeug. *Automobiltechnische Zeitschrift (ATZ)*, *87*(2), 77–84.

German Aerospace Center. Pegasus Research Project: Securing Automated Driving Effectively. Retrieved from https://www.pegasusprojekt.de/en/about-PEGASUS

Gershon, P., Ronen, A., Oron-Gilad, T., & Shinar, D. (2009). The effects of an interactive cognitive task (ICT) in suppressing fatigue symptoms in driving. *Transportation Research Part F: Traffic Psychology and Behaviour*, *12*(1), 21–28. https://doi.org/10.1016/j.trf.2008.06.004

Gimeno, P. T., Cerezuela, G. P., & Montanes, M. C. (2006). On the concept and measurement of driver drowsiness, fatigue and inattention: Implications for countermeasures. *International Journal of Vehicle Design*, *42*(1/2), 67. https://doi.org/10.1504/IJVD.2006.010178

Glass, G. V., Peckham, P. D., & Sanders, J. R. (2016). Consequences of Failure to Meet Assumptions Underlying the Fixed Effects Analyses of Variance and Covariance. *Review of Educational Research*, *42*(3), 237–288. https://doi.org/10.3102/00346543042003237

Gold, C. (2016). *Modeling of Take-Over Performance in Highly Automated Vehicle Guidance* (Dissertation). Technische Universität München, Garching. Retrieved from https://mediatum.ub.tum.de/1296132

Gold, C., & Bengler, K. (2014). Taking Over Control from Highly Automated Vehicles. In Ahram T., W. Karwowski, & T. Marek (Eds.), *Proceedings of the 5th International Conference on Applied Human Factors and Ergonomics AHFE 2014. 19-23 July 2014* (pp. 3662–3667). Kraków, Poland.

Gold, C., Berisha, I., & Bengler, K. (2015). Utilization of Drivetime - Performing Non-Driving Related Tasks While Driving Highly Automated. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *59*(1), 1666–1670. https://doi.org/10.1177/1541931215591360

Gold, C., Damböck, D., Lorenz, L., & Bengler, K. (2013). "Take over!": How long does it take to get the driver back into the loop? *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *57*(1), 1938–1942. https://doi.org/10.1177/1541931213571433

Gold, C., Happee, R., & Bengler, K. (2017). Modeling take-over performance in level 3 conditionally automated vehicles. *Accident; Analysis and Prevention*. Advance online publication. https://doi.org/10.1016/j.aap.2017.11.009

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. https://doi.org/10.1177/0018720816634226

Gold, C., Lorenz, L., & Bengler, K. (2014). Influence of automated brake application on take-over situations in highly automated driving scenarios. *Proceedings of FISITA World Automotive Congress. Maastricht, Netherlands.*

Gold, C., Naujoks, F., Radlmayr, J., Bellem, H., & Jarosch, O. (2018). Testing Scenarios for Human Factors Research in Level 3 Automated Vehicles. In N. A. Stanton, S. Landry, G. Di Bucchianico, & A. Vallicelli (Eds.), *Advances in Intelligent Systems and Computing: Vol. 597, Advances in Human*

Aspects of Transportation: Proceedings of the AHFE 2017 International Conference on Human Factors in Transportation, July 17−21, 2017, The Westin Bonaventure Hotel, Los Angeles, California, USA (pp. 551–559). Cham: Springer.

Gonçalves, J., Happee, R., & Bengler, K. (2016). Drowsiness in Conditional Automation: Proneness, diagnosis and driving performance effects. In *2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC): Windsor Oceanico Hotel, Rio de Janeiro, Brazil, November 1-4, 2016* (pp. 873–878). Piscataway, NJ: IEEE. https://doi.org/10.1109/ITSC.2016.7795658

Hancock, P. A., & Warm, J. S. (1989). A dynamic model of stress and sustained attention. *Human Factors*, *31*(5), 519–537. https://doi.org/10.1177/001872088903100503

Hanowski, R. J., Bowman, D., Alden, A., Wierwille, W. W., & Carroll, R. (2008). PERCLOS+: Moving Beyond Single-Metric Drowsiness Monitors. In SAE International (Chair), *Commercial Vehicle Engineering Congress & Exhibition*.

Hargutt, V. (2003). *Das Lidschlussverhalten als Indikator für Aufmerksamkeits- und Müdigkeitsprozesse bei Arbeitshandlungen* (Dissertation). Julius-Maximilians-Universität Würzburg, Würzburg. Retrieved from http://www.psychologie.uni-wuerzburg.de/izvw/texte/2003_hargutt_Das_Lidschlussverhalten.pdf

Hargutt, V., & Tietze, H. (2001, June). *Erfassung von Ermüdung und Müdigkeit via EEG und Lidschlagverhalten unter besonderer Berücksichtigung des PERCLOS-Maßes.* Vortrag auf der Tagung „Müdigkeit im Verkehr. Ursachen, Erkennung und Gegenmaßnahmen", Essen, Essen, 20. – 21. Juni 2001.

Harwell, M. R., Rubinstein, E. N., Hayes, W. S., & Olds, C. C. (2016). Summarizing Monte Carlo Results in Methodological Research: The One- and Two-Factor Fixed Effects ANOVA Cases. *Journal of Educational Statistics*, *17*(4), 315–339. https://doi.org/10.3102/10769986017004315

Hecht, T. (2016). *Die Auswirkung einer längeren hochautomatisierten Fahrt und freier Beschäftigung auf den Fahrerzustand und die Übernahmeleistung* (Unpublished master's thesis). Technical University of Munich, Munich.

Hecht, T., Feldhütter, A., Draeger, K., & Bengler, K. (2020). What do you do? An Analysis of Non-driving Related Activities During a 60 Minutes Conditionally Automated Highway Drive. In T. Ahram, R. Taiar, S. Colson, & A. Choplin (Eds.), *Advances in Intelligent Systems and Computing: Vol. 1018. Human Interaction and Emerging Technologies: Proceedings of the 1st International Conference on Human Interaction and Emerging Technologies (IHIET 2019), August 22-24, 2019, Nice, France* (1st ed., pp. 28–34). Cham, Cham: Springer International Publishing; Springer.

Hecht, T., Feldhütter, A., Radlmayr, J., Nakano, Y., Miki, Y., Henle, C., & Bengler, K. (2019). A Review of Driver State Monitoring Systems in the Context of Automated Driving. In S. Bagnara, R. Tartaglia, S. Albolino, T. Alexander, & Y. Fujita (Eds.), *Proceedings of the 20th Congress of the International Ergonomics Association (IEA 2018)* (pp. 398–408). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-96074-6_43

Heikoop, D. D., de Winter, J. C. F., van Arem, B., & Stanton, N. A. (2015). Psychological constructs in driving automation: A consensus model and critical comment on construct proliferation. *Theoretical Issues in Ergonomics Science*, *17*(3), 284–303. https://doi.org/10.1080/1463922X.2015.1101507

Hergeth, S., Lorenz, L., & Krems, J. F. (2017). Prior Familiarization With Takeover Requests Affects Drivers' Takeover Performance and Automation Trust. *Human Factors*, *59*(3), 457–470. https://doi.org/10.1177/0018720816678714

Hirose, T., Kitabayashi, D., & Kubota, H. (2015). Driving Characteristics of Drivers in a State of Low Alertness when an Autonomous System Changes from Autonomous Driving to Manual Driving. In *SAE Technical Paper Series*. SAE International 400 Commonwealth Drive, Warrendale, PA, United States. https://doi.org/10.4271/2015-01-1407

Hockey, G. R. J. (1997). Compensatory control in the regulation of human performance under stress and high workload: A cognitive-energetical framework. *Biological Psychology*, *45*(1-3), 73–93. https://doi.org/10.1016/S0301-0511(96)05223-4

Hoddes, E., Zarcone, V., Smythe, H., Phillips, R., & Dement, W. C. [W. C.] (1973). Quantification of sleepiness: A new approach. *Psychophysiology*, *10*(4), 431–436. https://doi.org/10.1111/j.1469-8986.1973.tb00801.x

Hoeger, R., Amditis, A., Kunert, M., Hoess, A., Flemisch, F., Krueger, H.-P., . . . Beutner, A. (2008). Highly Automated Vehicles for Intelligent Transport: HAVEit Approach. *ITS World Congress*.

Horne, J. A., & Baulk, S. D. (2004). Awareness of sleepiness when driving. *Psychophysiology*, *41*(1), 161–165. https://doi.org/10.1046/j.1469-8986.2003.00130.x

Horne, J. A., & Reyner, L. A. (1995). Sleep Related Vehicle Accidents. *British Medical Journal*, *310*(6979), 565–567.

Huang, K. (2018). *Effect of Automation Duration and Non-Driving Related Activities on the Level of Drowsiness in Conditionally Automated Driving* (Unpublished master's thesis). Technical University of Munich, Munich.

Ingre, M., Akerstedt, T., Peters, B., Anund, A., & Kecklund, G. (2006). Subjective sleepiness, simulated driving performance and blink duration: Examining individual differences. *Journal of Sleep Research*, *15*(1), 47–53. https://doi.org/10.1111/j.1365-2869.2006.00504.x

International Organization for Standardization (2012-11). *Road Vehicles -- Ergonomic aspects of transport information and control systems -- Calibration tasks for methods which assess driver demand due to the use of in-vehicle systems*. (Norm, ISO/TS 14198:2012).

International Organization for Standardization (2014-09). *Road vehicles — Measurement of driver visual behaviour with respect to transport information and control systems — Part 2: Equipment and procedures*. (Prestandard, ISO/TS 15007-2:2014).

Jan, T. von, Karnahl, T., Seifert, K., Hilgenstock, J., & Zobel, R. (2005, June). Don't sleep and drive – VW's fatigue detection technology. In National Highway Traffic Safety Administration (Chair), *19th International Technical Conference on the Enhanced Safety of Vehicles (ESV),* Washington DC, United States.

Jarosch, O., Bellem, H., & Bengler, K. (2019). Effects of Task-Induced Fatigue in Prolonged Conditional Automated Driving. *Human Factors*, *61*(7), 1186–1199. https://doi.org/10.1177/0018720818816226

Jarosch, O., & Bengler, K. (2019a). Is It the Duration of the Ride or the Non-driving Related Task? What Affects Take-Over Performance in Conditional Automated Driving? In S. Bagnara, R. Tartaglia, S. Albolino, T. Alexander, & Y. Fujita (Eds.), *Proceedings of the 20th Congress of the International Ergonomics Association (IEA 2018)* (pp. 512–523). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-96074-6_54

Jarosch, O., & Bengler, K. (2019b). Rating of Take-Over Performance in Conditionally Automated Driving Using an Expert-Rating System. In N. A. Stanton (Ed.), *Advances in Intelligent Systems and Computing: Vol. 786. Advances in human aspects of transportation: Proceedings of the AHFE 2018*

International Conference on Human Factors in Transportation, July 21-25, 2018, Loews Sapphire Falls Resort at Universal Studios, Orlando, Florida, USA (pp. 283–294). Cham: Springer. Retrieved from https://doi.org/10.1007/978-3-319-93885-1_26

Jarosch, O., Gold, C., Naujoks, F., Wandtner, B., Marberger, C., Weidl, G., & Schrauf, M. (2019). The Impact of Non-Driving Related Tasks on Take-over Performance in Conditionally Automated Driving – A Review of the Empirical Evidence. In Lehrstuhl für Fahrzeugtechnik, Technische Universität München & TÜV Süd Akademie GmbH (Chairs), *9. Tagung Automatisiertes Fahren,* München.

Jarosch, O., Kuhnt, M., Paradies, S., & Bengler, K. (2017). It's Out of Our Hands Now! Effects of Non-Driving Related Tasks During Highly Automated Driving on Drivers' Fatigue. In *Proceedings of the 9th International Driving Symposium on Human Factors in Driver Assessment, Training, and Vehicle Design: driving assessment 2017* (pp. 319–325). Iowa City, Iowa: University of Iowa. https://doi.org/10.17077/drivingassessment.1653

Jarosch, O., Paradies, S., Feiner, D., & Bengler, K. (2019). Effects of non-driving related tasks in prolonged conditional automated driving – A Wizard of Oz on-road approach in real traffic environment. *Transportation Research Part F: Traffic Psychology and Behaviour*, *65*, 292–305. https://doi.org/10.1016/j.trf.2019.07.023

Job, S. R. F., & Dalziel, J. (2001). Defining Fatigue as a Condition of the Organism and Distinguishing It From Habituation, Adaption and Boredom. In P. A. Hancock & P. A. Desmond (Eds.), *Human factors in transportation series. Stress, workload, and fatigue* (pp. 466–475). Mahwah NJ u.a.: Erlbaum Ass.

Johns, M. W. (2000). A sleep physiologist's view of the drowsy driver. *Transportation Research Part F: Traffic Psychology and Behaviour*, *3*(4), 241–249. https://doi.org/10.1016/S1369-8478(01)00008-0

Johns, M. W. (2007). *Drowsy Driving and the Law: A Submission to the Tasmania Law Reform Institute re: Issues Paper No 12*. Melbourne, Australia.

Kaida, K., Akerstedt, T., Kecklund, G., Nilsson, J. P., & Axelsson, J. (2007). Use of subjective and physiological indicators of sleepiness to predict performance during a vigilance task. *Industrial Health*, *45*(4), 520–526. https://doi.org/10.2486/indhealth.45.520

Kalb, L. (2017). *Übernahmezeitpunkt in Abhängigkeit des Fahrerzustandes: Erhebung von PERCLOS in Echtzeit und Gestaltung eines zeitlich variablen Übernahmeszenarios* (Unpublished master's thesis). Technical University of Munich, Munich.

Kämpchen, N., Aeberhard, M., Ardelt, M., & Rauch, S. (2012). Technologies for highly automated driving on highways. *ATZ Worldwide*, *114*(6), 34–38. https://doi.org/10.1007/s38311-012-0176-y

Karrer-Gauß, K. (2011). *Prospektive Bewertung von Systemen zur Müdigkeitserkennung: Ableitung von Gestaltungsempfehlungen zur Vermeidung von Risikokompensation aus empirischen Untersuchungen* (Dissertation). Technische Universität Berlin, Berlin.

Kerschbaum, P. (2017). *Design for Automation: The Steering Wheel in Highly Automated Cars* (Dissertation). Technical University of Munich, Munich.

Klauer, S. G., Dingus, T. A., Neale, V. L., Sudweek, J. D., & Ramsey, D. J. (2006). *The Impact of Driver Inattention on Near-Crash/Crash Risk: An Analysis Using the 100-Car Naturalistic Driving Study Data*. Technical Report DOT HS 810 594. Blacksburg, Virginia, USA.

Knipling, R., & Wang, J.-S. (1994). *Crashes and Fatalities Related to Driver Drowsiness/Fatigue*. Research Note. Washington, DC.

Kompass, K., Gruber, C., & Domsch, C. (2010). Der Beitrag von Fahrerassistenzsystemen zur Aktiven und Passiven Sicherheit - die Integrale Sicherheit als Antwort auf die wachsende Anforderungen an die Fahrzeugsicherheit. In Lehrstuhl für Fahrzeugtechnik, Technische Universität München & TÜV Süd Akademie GmbH (Chairs), *4. Tagung Sicherheit durch Fahrerassistenz*, München.

Körber, M. (2018). *Individual Differences in Human-Automation Interaction: A Driver-Centered Perspective on the Introduction of Automated Vehicles* (Dissertation). Technical University of Munich, Munich.

Körber, M., Gold, C., Lechner, D., & Bengler, K. (2016). The influence of age on the take-over of vehicle control in highly automated driving. *Transportation Research Part F: Traffic Psychology and Behaviour*, *39*, 19–32. https://doi.org/10.1016/j.trf.2016.03.002

Krajewski, J., Batliner, A., & Golz, M. (2009). Acoustic sleepiness detection: Framework and validation of a speech-adapted pattern recognition approach. *Behavior Research Methods*, *41*(3), 795–804. https://doi.org/10.3758/BRM.41.3.795

Krajewski, J., Trutschel, U., Golz, M., Sommer, D., & Edwards, D. (2017). Estimating Fatigue from Predetermined Speech Samples Transmitted by Operator Communication Systems. In *Proceedings of the 5th International Driving Symposium on Human Factors in Driver Assessment, Training, and Vehicle Design : Driving Assessment 2009* (pp. 468–474). Iowa City, Iowa: University of Iowa. https://doi.org/10.17077/drivingassessment.1359

Kröger, F. (2016). Automated Driving in Its Social, Historical and Cultural Contexts. In M. Maurer, J. C. Gerdes, B. Lenz, & H. Winner (Eds.), *Autonomous Driving* (pp. 41–68). Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-662-48847-8_3

Kroll, D. (2018). *Effect of Fatigue on the Take-Over Performance in Conditionally Automated Driving* (Unpublished master's thesis). Technical University of Munich, Munich.

L3Pilot consortium. L3Pilot. Piloting Automated Driving on European Roads. Retrieved from https://www.l3pilot.eu/

Lal, S. K., & Craig, A. (2001). A critical review of the psychophysiology of driver fatigue. *Biological Psychology*, *55*(3), 173–194. https://doi.org/10.1016/S0301-0511(00)00085-5

Lal, S. K., & Craig, A. (2002). Driver fatigue: Electroencephalography and psychological assessment. *Psychophysiology*, *39*(3), 313–321. https://doi.org/10.1017/S0048577201393095

Langenberg, J., Bartels, A., & Etemand, A. (2014). Eu-Projekt "AdaptIVe". Ansätze für hochautomatisches Fahren. In *VDI-Berichte: Vol. 2223, Fahrerassistenz und Integrierte Sicherheit: 30. Vdi/vw-Gemeinschaftstagung ; Wolfsburg, 14. Und 15. Oktober 2014*. Düsseldorf: VDI-Verl.

Langer, B., Abendroth, B., & Bruder, R. (2015). Fahrerzustandserkennung. In H. Winner, S. Hakuli, F. Lotz, & C. Singer (Eds.), *Handbuch Fahrerassistenzsysteme: Grundlagen, Komponenten und Systeme für aktive Sicherheit und Komfort* (3rd ed., pp. 687–699). Wiesbaden: Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-05734-3_38

Large, D., Burnett, G., Morris, A., Muthumani, A., & Matthias, R. (2017). Design Implications of Drivers' Engagement with Secondary Activities During Highly-Automated Driving – A Longitudinal Simulator Study. *Road Safety and Simulation International Conference.*

Larue, G., Rakotonirainy, A., & Pettitt, A. (2010). Predicting Driver's Hypovigilance on Monotonous Roads: Literature Review. *1st International Conference, Gotheburg, Sweden.*

Lazarus, R. S. (1999). *Stress and Emotion: A New Synthesis*. New York: Springer.

Lin, L., Huang, C., Ni, X., Wang, J., Zhang, H. [Hao], Li, X., & Qian, Z. (2015). Driver fatigue detection based on eye state. *Technology and Health Care : Official Journal of the European Society for Engineering and Medicine*, *23 Suppl 2*, S453-63. https://doi.org/10.3233/THC-150982

Liu, C. C., Hosking, S. G., & Lenné, M. G. (2009). Predicting driver drowsiness using vehicle measures: Recent insights and future challenges. *Journal of Safety Research*, *40*(4), 239–245. https://doi.org/10.1016/j.jsr.2009.04.005

Loftis, C. (2011). Alertness. In J. S. Kreutzer (Ed.), *Encyclopedia of clinical neuropsychology* (p. 77). New York, NY: Springer. https://doi.org/10.1007/978-0-387-79948-3_2122

Lu, Z., Zhang, B., Feldhütter, A., Happee, R., Martens, M., & de Winter, J. C. F. (2019). Beyond mere take-over requests: The effects of monitoring requests on driver attention, take-over performance, and acceptance. *Transportation Research Part F: Traffic Psychology and Behaviour*, *63*, 22–37. https://doi.org/10.1016/j.trf.2019.03.018

Mackworth, N. H. (1957). Vigilance. *The Advancement of Science*, *53*, 389–393.

Marberger, C., Mielenz, H., Naujoks, F., Radlmayr, J., Bengler, K., & Wandtner, B. (2018). Understanding and Applying the Concept of "Driver Availability" in Automated Driving. In N. A. Stanton, S. Landry, G. Di Bucchianico, & A. Vallicelli (Eds.), *Advances in Intelligent Systems and Computing: Vol. 597, Advances in Human Aspects of Transportation: Proceedings of the AHFE 2017 International Conference on Human Factors in Transportation, July 17−21, 2017, The Westin Bonaventure Hotel, Los Angeles, California, USA* (pp. 595–605). Cham: Springer. https://doi.org/10.1007/978-3-319-60441-1_58

Matthews, G., Campbell, S. E., Falconer, S., Joyner, L. A., Huggins, J., Gilliland, K., . . . Warm, J. S. (2002). Fundamental dimensions of subjective state in performance settings: Task engagement, distress, and worry. *Emotion (Washington, D.C.)*, *2*(4), 315–340. https://doi.org/10.1037//1528-3542.2.4.315

Matthews, G., & Desmond, P. A. (2002). Task-induced fatigue states and simulated driving performance. *The Quarterly Journal of Experimental Psychology. A, Human Experimental Psychology*, *55*(2), 659–686. https://doi.org/10.1080/02724980143000505

Matthews, G., Desmond, P. A., & Hitchcock, E. M. (2012). Dimensional Models of Fatigue. In G. Matthews, P. A. Desmond, C. Neubauer, & P. A. Hancock (Eds.), *The Handbook of Operator Fatigue* (pp. 139–154). Burlington, VT: Ashgate Pub. Company.

Matthews, G., Hancock, P. A., & Desmond, P. A. (2012). Models of Individual Differences in Fatigue for Performance Research. In G. Matthews, P. A. Desmond, C. Neubauer, & P. A. Hancock (Eds.), *The Handbook of Operator Fatigue* (pp. 155–170). Burlington, VT: Ashgate Pub. Company.

Maurer, M., Gerdes, J. C., Lenz, B., & Winner, H. (Eds.) (2016). *Autonomous Driving*. Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-662-48847-8

May, J. F., & Baldwin, C. L. (2009). Driver fatigue: The importance of identifying causal factors of fatigue when considering detection and countermeasure technologies. *Transportation Research Part F: Traffic Psychology and Behaviour*, *12*(3), 218–224. https://doi.org/10.1016/j.trf.2008.11.005

Miley, A. Å., Kecklund, G., & Åkerstedt, T. (2016). Comparing two versions of the Karolinska Sleepiness Scale (KSS). *Sleep and Biological Rhythms*, *14*(3), 257–260. https://doi.org/10.1007/s41105-016-0048-8

Mullins, H. M., Cortina, J. M., Drake, C. L., & Dalal, R. S. (2014). Sleepiness at work: A review and framework of how the physiology of sleepiness impacts the workplace. *The Journal of Applied Psychology*, *99*(6), 1096–1112. https://doi.org/10.1037/a0037885

Naujoks, F., Befelein, D., Wiedemann, K., & Neukum, A. (2018). A review of non-driving-related tasks used in studies on automated driving. In N. A. Stanton, S. Landry, G. Di Bucchianico, & A. Vallicelli (Eds.), *Advances in Intelligent Systems and Computing: Vol. 597, Advances in Human Aspects of Transportation: Proceedings of the AHFE 2017 International Conference on Human Factors in Transportation, July 17−21, 2017, The Westin Bonaventure Hotel, Los Angeles, California, USA* (pp. 525–537). Cham: Springer. https://doi.org/10.1007/978-3-319-60441-1_52

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. https://doi.org/10.1016/j.mex.2018.05.007

Neubauer, C. [C.], Matthews, G., & Saxby, D. (2012). The Effects of Cell Phone Use and Automation on Driver Performance and Subjective State in Simulated Driving. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *56*(1), 1987–1991. https://doi.org/10.1177/1071181312561415

Neubauer, C. [C.], Matthews, G., & Saxby, D. (2014). Fatigue in the Automated Vehicle: Do Games and Conversation Distract or Energize the Driver? *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *58*(1), 2053–2057. https://doi.org/10.1177/1541931214581432

Niederl, T. (2007). *Untersuchungen zu kumulativen psychischen und physiologischen Effekten des fliegenden Personals auf der Kurzstrecke* (Dissertation). Universität Kassel, Kassel.

Omae, M., Fujioka, T., Hashimoto, N., & Shimizu, H. (2006). The Application of RTK-GPS and Steer-by-wire Technology to the Automatic Driving of Vehicles and an Evaluation of Driver Behavior. *IATSS Research*, *30*(2), 29–38. https://doi.org/10.1016/S0386-1112(14)60167-9

Otmani, S., Pebayle, T., Roge, J., & Muzet, A. (2005). Effect of driving duration and partial sleep deprivation on subsequent alertness and performance of car drivers. *Physiology & Behavior*, *84*(5), 715–724. https://doi.org/10.1016/j.physbeh.2005.02.021

Papadelis, C., Lithari, C., Kourtidou-Papadeli, C., Bamidis, P. D., Portouli, E., & Bekiaris, E. (2009). Monitoring driver's sleepiness on-board for preventing road accidents. *Studies in Health Technology and Informatics*, *150*, 485–489.

Parasuraman, R., & Riley, V. (1997). Humans and Automation: Use, Misuse, Disuse, Abuse. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *39*(2), 230–253. https://doi.org/10.1518/001872097778543886

Parasuraman, R., Warm, J. S., & See, J. S. (1998). Brain Systems of Vigilance. In R. Parasuraman (Ed.), *A Bradford book. The Attentive Brain* (pp. 221–256). London: MIT Press.

Petermann-Stock, I. [I.], Hackenberg, L., Muhr, T. [T.], & Mergl, C. (2013). Wie lange braucht der Fahrer?: Eine Analyse zu Übernahmezeiten aus verschiedenen Nebentätigkeiten während einer hochautomatisierten Staufahrt. In Lehrstuhl für Fahrzeugtechnik, Technische Universität München & TÜV Süd Akademie GmbH (Chairs), *6. Tagung Fahrerassistenz - Auf dem Weg zum automatischen Fahren,* München.

Petermeijer, S. M. (2017). *A vibrotactile interface to support the driver during the take-over process* (Dissertation). Technical University of Munich, Munich.

Petermeijer, S. M., de Winter, J. C. F., & Bengler, K. (2016). Vibrotactile Displays: A Survey With a View on Highly Automated Driving. *IEEE Transactions on Intelligent Transportation Systems*, *17*(4), 897–907. https://doi.org/10.1109/TITS.2015.2494873

Phillips, R. O. (09.2014). *What is fatigue and how does it affect the safety performance of human transport operators?: Fatigue in Transport Report*. Oslo, Norway.

Phillips, R. O. (2015). A review of definitions of fatigue – And a step towards a whole definition. *Transportation Research Part F: Traffic Psychology and Behaviour*, *29*, 48–56. https://doi.org/10.1016/j.trf.2015.01.003

Platho, C., Pietrek, A., & Kolrep-Rometsch, H. (2013). *Erfassung der Fahrermüdigkeit: Bericht zum Forschungsprojekt FE 82.513/2010. Berichte der Bundesanstalt für Strassenwesen F, Fahrzeugtechnik: Vol. 89*. Bremen: Fachverl. NW in der Carl Schünemann Verl. GmbH. Retrieved from http://bast.opus.hbz-nrw.de/volltexte/2013/659/pdf/BASt_F_89_barrierefrei.pdf

Popieul, J. C., Simon, P., & Loslever, P. (2002). Using driver's head movements evolution as a drowsiness indicator. In *Ieee IV2003 Intelligent Vehicles Symposium: Proceedings : Hilton Hotel, Columbus, Ohio, USA, June 9-11, 2003* (pp. 616–621). Piscataway, N.J.: IEEE. https://doi.org/10.1109/IVS.2003.1212983

Radlmayr, J., Feldhütter, A., Frey, A., Jarosch, O., Marberger, C., Naujoks, F., . . . Bengler, K. (2019). Drowsiness and fatigue in conditionally automated driving – Towards an integrative framework. In D. de Waard, K. Brookhuis, D. Coelho, S. Fairclough, D. Manzey, A. Naumann, . . . R. Wiczorek (Chairs), *Human Factors and Ergonomics Society Europe Chapter Annual Meeting,* Berlin, Germany. Retrieved from https://www.hfes-europe.org/technology-ageing-society/

Radlmayr, J., Gold, C., Lorenz, L., Farid, M., & Bengler, K. (2014). How Traffic Situations and Non-Driving Related Tasks Affect the Take-Over Quality in Highly Automated Driving. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *58*(1), 2063–2067. https://doi.org/10.1177/1541931214581434

Radlmayr, J., Ratter, M., Feldhütter, A., Körber, M., Prasch, L., Schmidtler, J., . . . Bengler, K. (2019). Take-Overs in Level 3 Automated Driving – Proposal of the Take-Over Performance Score (TOPS). In S. Bagnara, R. Tartaglia, S. Albolino, T. Alexander, & Y. Fujita (Eds.), *Proceedings of the 20th Congress of the International Ergonomics Association (IEA 2018)* (pp. 436–446). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-96074-6_46

Rasch, D., Kubinger, K. D., & Moder, K. (2011). The two-sample t test: Pre-testing its assumptions does not pay off. *Statistical Papers*, *52*(1), 219–231. https://doi.org/10.1007/s00362-009-0224-x

Rasmussen, J. (1983). Skills, rules, and knowledge; signals, signs, and symbols, and other distinctions in human performance models. *IEEE Transactions on Systems, Man, and Cybernetics*, *SMC-13*(3), 257–266. https://doi.org/10.1109/TSMC.1983.6313160

Rasmussen, J. (1986). *Information processing and human-machine interaction: An approach to cognitive engineering. North-Holland series in system science and engineering: Vol. 12*. New York, N.Y.: North-Holland.

Rauch, N., Kaussner, A. [Armin], Boverie, S., & Giralt, A. (2009, April 30). *HAVEit -Highly automated vehicles for intelligent transport: Model of driver behaviour for the assessment of driver's state.* Deliverable D32.1 Report on driver assessment methodology. Retrieved from http://haveit-eu.org/LH2Uploads/ItemsContent/24/HAVEIt_212154_D32.1_public_version.pdf

Rauch, N., Kaussner, A. [A.], Krüger, H.-P., Boverie, S., & Flemisch, F. O. (2009). The importance of driver state assessment within highly automated vehicles. *16th ITS World Congress and Exhibition on Intelligent Transport Systems and Services*.

Reyner, L. A., & Horne, J. A. (1998). Falling asleep whilst driving: Are drivers aware of prior sleepiness? *International Journal of Legal Medicine*, *111*(3), 120–123. https://doi.org/10.1007/s004140050131

Richter, P., & Hacker, W. (1998). *Belastung und Beanspruchung: Stress, Ermüdung und Burnout im Arbeitsleben*. Heidelberg: Asanger.

Rosario, H. de, Solaz, J. S., Rodriguez, N., & Bergasa, L. M. (2010). Controlled inducement and measurement of drowsiness in a driving simulator. *IET Intelligent Transport Systems*, *4*(4), 280. https://doi.org/10.1049/iet-its.2009.0110

Ruhl, A. (2018). *Einfluss von Müdigkeit auf das Übernahmeverhalten in komplexen Situationen beim hochautomatisierten Fahren* (Unpublished master's thesis). Technical University of Munich, Munich.

Ruxton, G. D. (2006). The unequal variance t-test is an underused alternative to Student's t-test and the Mann–Whitney U test. *Behavioral Ecology*, *17*(4), 688–690. https://doi.org/10.1093/beheco/ark016

SAE International (2018, June 15). *Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles*. (SAE Standard, J3016_201806).

Salkind, N. J. J. (2010). *Encyclopedia of Research Design*. Thousand Oaks: SAGE Publications. Retrieved from http://gbv.eblib.com/patron/FullRecord.aspx?p=996805

Sarter, N. B., & Woods, D. D. (1995). How in the World Did We Ever Get into That Mode?: Mode Error and Awareness in Supervisory Control. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *37*(1), 5–19. https://doi.org/10.1518/001872095779049516

Saxby, D. J., Matthews, G., Hitchcock, E. M., & Warm, J. S. (2007). Development of Active and Passive Fatigue Manipulations Using a Driving Simulator. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *51*(18), 1237–1241. https://doi.org/10.1177/154193120705101839

Saxby, D. J., Matthews, G., Hitchcock, E. M., Warm, J. S., Funke, G. J., & Gantzer, T. (2008). Effect of Active and Passive Fatigue on Performance Using a Driving Simulator. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *52*(21), 1751–1755. https://doi.org/10.1177/154193120805202113

Saxby, D. J., Matthews, G., Warm, J. S., Hitchcock, E. M., & Neubauer, C. [Catherine] (2013). Active and passive fatigue in simulated driving: Discriminating styles of workload regulation and their safety impacts. *Journal of Experimental Psychology. Applied*, *19*(4), 287–300. https://doi.org/10.1037/a0034386

Schlake, C. (2019a). *Analyse von Übernahmekonzepten für das hochautomatisierte Fahren im Kontext der Gebrauchs- und Funktionssicherheit bei partiellem Systemversagen: Analysis of Take-over Request concepts for conditionally automated driving regarding safety of use and functional safety in terms of partial system failure* (Unpublished Semester's thesis). Technical University of Munich, Munich.

Schlake, C. (2019b). *Realfahrzeugstudie: Wirksamkeitsnachweis einer unimodalen, vestibulären Übernahmeaufforderung zum Erzeugen einer Mode Awareness im hochautomatisierten Fahren: Proofing the Efficacy of a Unimodal Vestibular Take-Over Request to Create a Mode Awareness in Conditionally Automated Driving* (Unpublished Master's thesis). Technical University of Munich, Munich.

Schleicher, R. [R.], Galley, N. [N.], Briest, S., & Galley, L. (2008). Blinks and saccades as indicators of fatigue in sleepiness warnings: Looking tired? *Ergonomics*, *51*(7), 982–1010. https://doi.org/10.1080/00140130701817062

Schmidt, E. A., Schrauf, M., Simon, M., Buchner, A., & Kincses, W. E. (2011). The short-term effect of verbally assessing drivers' state on vigilance indices during monotonous daytime driving. *Transportation Research Part F: Traffic Psychology and Behaviour*, *14*(3), 251–260. https://doi.org/10.1016/j.trf.2011.01.005

Schmidt, E. A., Schrauf, M., Simon, M., Fritzsche, M., Buchner, A., & Kincses, W. E. (2009). Drivers' misjudgement of vigilance state during prolonged monotonous daytime driving. *Accident; Analysis and Prevention*, *41*(5), 1087–1093. https://doi.org/10.1016/j.aap.2009.06.007

Schmidt, J. (2018). *Detektion der Reaktionsbereitschaft beim hochautomatisierten Fahren* (Dissertation). Technischen Universität Berlin, Berlin.

Schmidt, J., Braunagel, C., Stolzmann, W., & Karrer-Gauss, K. (2016). Driver drowsiness and behavior detection in prolonged conditionally automated drives. In *2016 IEEE Intelligent Vehicles Symposium (IV): 19-22 June 2016* (pp. 400–405). Piscataway, NJ: IEEE. https://doi.org/10.1109/IVS.2016.7535417

Schmidt, J., Dreißig, M., Stolzmann, W., & Rötting, M. (2017). The Influence of Prolonged Conditionally Automated Driving on the Take-Over Ability of the Driver. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *61*(1), 1974–1978. https://doi.org/10.1177/1541931213601972

Schmidt, J., Stolzmann, W., & Karrer-Gauss, K. (2016). Experimental evaluation of different request intervals for a driver alertness device for conditionally automated driving with induced drowsiness. In Intelligente Transport- und Verkehrssysteme und -dienste Niedersachsen e.V. (Ed.), *17. Braunschweiger Symposium AAET 2016 - Automatisierungssysteme, Assistenzsysteme und eingebettete Systeme für Transportmittel: 10. bis 11. Februar 2016, Stadthalle Braunschweig*. Braunschweig: ITS automotive nord.

Schömig, N., Hargutt, V., Neukum, A., Petermann-Stock, I. [Ina], & Othersen, I. (2015). The Interaction Between Highly Automated Driving and the Development of Drowsiness. *Procedia Manufacturing*, *3*, 6652–6659. https://doi.org/10.1016/j.promfg.2015.11.005

Schultz, G., & Young, H. (2007). *A Safety Analysis of Fatigue and Drowsy Driving in the state of Utah*. Provo, Utah, USA.

Selvakumar, K., Jerome, J., Rajamani, K., & Shankar, N. (2016). Real-Time Vision Based Driver Drowsiness Detection Using Partial Least Squares Analysis. *Journal of Signal Processing Systems*, *85*(2), 263–274. https://doi.org/10.1007/s11265-015-1075-4

Senaratne, R., Hardy, D., Vanderaa, B., & Halgamuge, S. (2007). Driver Fatigue Detection by Fusing Multiple Cues. In D. Liu, S. Fei, Z. Hou, C. Sun, & H. Zhang (Eds.), *Lecture Notes in Computer Science: Vol. 4492. Advances in Neural Networks - ISNN 2007: 4th International Symposium on Neural Networks, ISNN 2007, Nanjing, China, June 3-7, 2007, Proceedings, Part II*. Berlin, Heidelberg: Springer-Verlag Berlin Heidelberg. https://doi.org/10.1007/978-3-540-72393-6_96

Shen, J., Barbera, J., & Shapiro, C. M. (2006). Distinguishing sleepiness and fatigue: Focus on definition and measurement. *Sleep Medicine Reviews*, *10*(1), 63–76. https://doi.org/10.1016/j.smrv.2005.05.004

Stanton, N. A., & Young, M. S. (2000). A proposed psychological model of driving automation. *Theoretical Issues in Ergonomics Science*, *1*(4), 315–331. https://doi.org/10.1080/14639220052399131

Statistisches Bundesamt (2019). *Traffic accidents: Persons killed in traffic accidents*. Retrieved from https://www.destatis.de/DE/Themen/Gesellschaft-Umwelt/Verkehrsunfaelle/Tabellen/getoetete-alter.html

Stern, J. A., Boyer, D., & Schroeder, D. (1994). Blink rate: A possible measure of fatigue. *Human Factors*, *36*(2), 285–297. https://doi.org/10.1177/001872089403600209

Szalma, J. L. (2012). Individual Differences in Stress, Fatigue and Performance. In G. Matthews, P. A. Desmond, C. Neubauer, & P. A. Hancock (Eds.), *The Handbook of Operator Fatigue* (pp. 75–90). Burlington, VT: Ashgate Pub. Company.

Thiffault, P., & Bergeron, J. (2003a). Fatigue and individual differences in monotonous simulated driving. *Personality and Individual Differences*, *34*(1), 159–176. https://doi.org/10.1016/S0191-8869(02)00119-8

Thiffault, P., & Bergeron, J. (2003b). Monotony of road environment and driver fatigue: A simulator study. *Accident Analysis & Prevention*, *35*(3), 381–391. https://doi.org/10.1016/S0001-4575(02)00014-3

Tijerina, L., Gleckler, M., Stoltzfus, D., Johnston, S., Goodman, M. J., & Wierwille, W. W. (1999). *A Preliminary Assessment of Algorithms for Drowsy and Inattentive Driver Detection on the Road*. Technical Report (DOT HS 808). Washington, D.C.

Toyota Motor Company. Driver Attention Monitor. Retrieved from https://global.toyota/en/detail/248128

Tsuchida, A., Bhuiyan, M., & Oguri, K. (2009). Estimation of drowsiness level based on eyelid closure and heart rate variability. *Conference Proceedings : ... Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual Conference*, *2009*, 2543–2546. https://doi.org/10.1109/IEMBS.2009.5334766

Van den Beukel, A. P. (2016). *Driving Automation Interface Design: supporting drivers' changing role* (PhD Thesis). Unisversity of Twente, Enschede, the Netherlands.

Verwey, W. B., & Zaidel, D. M. (1999). Preventing drowsiness accidents by an alertness maintenance device. *Accident Analysis & Prevention*, *31*(3), 199–211. https://doi.org/10.1016/S0001-4575(98)00062-1

Verwey, W. B., & Zaidel, D. M. (2000). Predicting drowsiness accidents from personal attributes, eye blinks and ongoing driving behaviour. *Personality and Individual Differences*, *28*(1), 123–142. https://doi.org/10.1016/S0191-8869(99)00089-6

Vogelpohl, T., Kühn, M., Hummel, T., & Vollrath, M. (2018). Asleep at the automated wheel-Sleepiness and fatigue during highly automated driving. *Accident; Analysis and Prevention.* Advance online publication. https://doi.org/10.1016/j.aap.2018.03.013

Volkswagen AG. Driver Alert System. Retrieved from https://www.volkswagen-newsroom.com/de/muedigkeitserkennung-3932

Volvo Group (2020, January 31). Driver Alert Control (DAC). Retrieved from https://www.volvocars.com/de/support/manuals/s60/2018w46/fahrerunterstutzung/driver-alert-control/driver-alert-control

Wehlack, V. (2019). *Automated Driving: Development of a Drowsiness Managament Concept and Evaluation of Related Key Elements* (Dissertation). Technical University of Munich, Munich.

Weinbeer, V., Baur, C., Radlmayr, J., Bill, J.-S., Muhr, T. [Tobias], & Bengler, K. (2018). Highly automated driving: How to get the driver drowsy and how does drowsiness influence various take-over aspects? *8. Tagung Fahrerassistenz, Einführung Hochautomatisiertes Fahren, 22. - 23. November 2017, München.*

Weinbeer, V., Bill, J.-S., Baur, C., & Bengler, K. (2018). Automated Driving: Subjective Assessment of Different Strategies to Manage Drowsiness. In D. de Waard, F. Di Nocera, D. Coelho, J. Edworthy, K. Brookhuis, F. Ferlazzo, . . . A. Toffetti (Chairs), *Human Factors and Ergonomics Society Europe Chapter.* Symposium conducted at the meeting of Human Factors and Ergonomics Society (HFES), Rome, Italy. Retrieved from https://www.hfes-europe.org/wp-content/uploads/2017/10/Weinbeer2017.pdf

Weinbeer, V., Muhr, T. [Tobias], & Bengler, K. (2019). Automated Driving: The Potential of Non-driving-Related Tasks to Manage Driver Drowsiness. In S. Bagnara, R. Tartaglia, S. Albolino, T. Alexander, & Y. Fujita (Eds.), *Proceedings of the 20th Congress of the International Ergonomics Association (IEA 2018)* (pp. 179–188). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-96074-6_19

Wickens, C. D., Hollands, J. G., Banbury, S., & Parasuraman, R. (2013). *Engineering psychology and human performance* (Fourth edition). New Jersey, United States: Pearson Education, Inc.

Wiegand, D. M., McClafferty, J., McDonald, S. E., & Hanowski, R. J. (2009, February 25). *Development and Evaluation of a Naturalistic Observer Rating of Drowsiness Protocol: Final Report.* Blacksburg, Virginia, USA.

Wierwille, W. W., & Ellsworth, L. A. (1994). Evaluation of driver drowsiness by trained raters. *Accident; Analysis and Prevention*, *26*(5), 571–581.

Wierwille, W. W., Wreggit, S., Kirn, C., Ellsworth, L., & Fairbanks, R. (1994). *Research on Vehicle-Based DriverStatus/Performance Monitoring: Development, Validation, and Refinementof Algorithms for Detection of Driver Drowsiness.* DOT HS 808 247.

Wilkinson, V. E., Jackson, M. L., Westlake, J., Stevens, B., Barnes, M., Swann, P., . . . Howard, M. E. (2013). The accuracy of eyelid movement parameters for drowsiness detection. *Journal of Clinical Sleep Medicine : JCSM : Official Publication of the American Academy of Sleep Medicine*, *9*(12), 1315–1324. https://doi.org/10.5664/jcsm.3278

Williamson, A. (2012). Countermeasures for Driver Fatigue. In G. Matthews, P. A. Desmond, C. Neubauer, & P. A. Hancock (Eds.), *The Handbook of Operator Fatigue* (pp. 441–455). Burlington, VT: Ashgate Pub. Company.

Williamson, A., Lombardi, D. A., Folkard, S., Stutts, J., Courtney, T. K., & Connor, J. L. (2011). The link between fatigue and safety. *Accident; Analysis and Prevention*, *43*(2), 498–515. https://doi.org/10.1016/j.aap.2009.11.011

Winkle, T. (2016). Safety Benefits of Automated Vehicles: Extended Findings from Accident Research for Development, Validation and Testing. In M. Maurer, J. C. Gerdes, B. Lenz, & H. Winner (Eds.), *Autonomous Driving* (pp. 335–364). Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-662-48847-8_17

World Health Organization (2018). *Global status report on road safety 2018*. Geneva, Switzerland: World Health Organization.

Wu, G., Liu, Z., Pan, X., Chen, F., Xu, M., Feng, D., & Xia, Z. (2018). Fatigue Driving Influence Research and Assessment. In N. A. Stanton, S. Landry, G. Di Bucchianico, & A. Vallicelli (Eds.), *Advances in Intelligent Systems and Computing: Vol. 597, Advances in Human Aspects of Transportation:*

Proceedings of the AHFE 2017 International Conference on Human Factors in Transportation, July 17−21, 2017, The Westin Bonaventure Hotel, Los Angeles, California, USA (pp. 677–688). Cham: Springer. https://doi.org/10.1007/978-3-319-41682-3_57

Wu, Y., Kihara, K., Hasegawa, K., Takeda, Y., Sato, T., Akamatsu, M., & Kitazaki, S. (2020). Age-related differences in effects of non-driving related tasks on takeover performance in automated driving. *Journal of Safety Research*, *72*, 231–238. https://doi.org/10.1016/j.jsr.2019.12.019

Wylie, C. D., Shultz, T., Miller, J. C., Mitler, M. M., & Mackie, R. R. (November 1996). *Commercial Motor Vehicle: Driver Fatigue and Alertness Study*. Technical Summary (FHWA-MC-97-00 1). Washington, D.C.

Yang, Y., Karakaya, B., Dominioni, G. C., Kawabe, K., & Bengler, K. (2018). An HMI Concept to Improve Driver's Visual Behavior and Situation Awareness in Automated Vehicle. In *2018 21st International Conference on Intelligent Transportation Systems (ITSC)* (pp. 650–655). IEEE. https://doi.org/10.1109/ITSC.2018.8569986

Zeeb, K., Buchner, A., & Schrauf, M. (2015). What determines the take-over time? An integrated model approach of driver take-over after automated driving. *Accident; Analysis and Prevention*, *78*, 212–221. https://doi.org/10.1016/j.aap.2015.02.023

ZENTEC Zentrum für Technologie, Existenzgründung und Cooperation GmbH. Projekt Ko-HAF. Retrieved from https://www.ko-haf.de/startseite/

Zhao, X., & Rong, J. (2013). The Relationship between Driver Fatigue and Monotonous Road Environment. In W. Wang & G. Wets (Eds.), *Atlantis Computational Intelligence Systems. Computational Intelligence for Traffic and Mobility* (Vol. 8, pp. 19–36). Paris: Atlantis Press. https://doi.org/10.2991/978-94-91216-80-0_2

Zilberg, E., Xu, Z. M., Burton, D., Karrar, M., & Lal, S. (2009). Statistical validation of physiological indicators for non-invasive and hybrid driver drowsiness detection system. *African Journal of Information & Communication Technology*, *5*(2). https://doi.org/10.5130/ajict.v5i2.1128

Appendix

Generic Annotation Guideline for Post-Hoc Assessment

Level of fatigue	Behavior	Cue Type	Source
Level 1: Not fatigued	No behavior annotated	Duration	
	- Normal facial tone - Normal fast eye blinks (<0.5 seconds) - Short ordinary glances - Occasional body movements or gestures		(Karrer-Gauß, 2011; Wierwille & Ellsworth, 1994)
Level 2: Fatigued		Duration	
	Extended eye closure (0.5 – 1 seconds)	Onetime	(Karrer-Gauß, 2011; Wierwille & Ellsworth, 1994)
	Glazed look or rare saccades (“glassy-eyed appearance”)	Onetime	(Karrer-Gauß, 2011; Wierwille & Ellsworth, 1994)
	Frequent eye blinks	Onetime	(Karrer-Gauß, 2011)
	Low facial tone: “floppy” (slack) face	Onetime	(Karrer-Gauß, 2011; Wierwille & Ellsworth, 1994)
	Yawning	Onetime	(Belz, 2000; Eskandarian et al., 2007; Wierwille & Ellsworth, 1994)
	Self-activating behaviors (mannerisms) - Eye rubbing (Karrer-Gauß, 2011) - Face rubbing (Karrer-Gauß, 2011) - Grimace (Karrer-Gauß, 2011) - Scratching (Karrer-Gauß, 2011) - Keeping hands busy (drumming, using Infotainment system, crossing arms, other) based on Eskandarian et al. (2007) - Moving restlessly in the seat (Karrer-Gauß, 2011) - Stretching (Anund et al., 2013; Senaratne et al., 2007) - Frequent normal posture adjustment (Anund et al., 2013; Senaratne et al., 2007) - Other behavior (documentation)	Onetime (Each)	(Anund et al., 2013; Eskandarian et al., 2007; Karrer-Gauß, 2011; Senaratne et al., 2007; Wierwille & Ellsworth, 1994)

Level 3: Very fatigued		Duration	
	Long eye closure (1.0 – 2.0 seconds)	Onetime	(Karrer-Gauß, 2011)
	Eye rolling	Onetime	(Karrer-Gauß, 2011; Wierwille & Ellsworth, 1994)
	Fixed gaze (no fixation, rare blinks) or cross-eyed look ("*lack of proper vergence"*)	Onetime	(Karrer-Gauß, 2011; Wierwille & Ellsworth, 1994)
	Comfortable sitting position or lack of activity - Resting head (Wierwille & Ellsworth, 1994) - Leaning head back (Tsuchida et al., 2009) - Rare head movement (Eskandarian et al., 2007; J. Schmidt, Braunagel, et al., 2016) and rare body movement (Karrer-Gauß, 2011; Wierwille & Ellsworth, 1994) - Others (slouching: slide down the seat, etc.) (Eskandarian et al., 2007; Senaratne et al., 2007)	Onetime (Each)	(Eskandarian et al., 2007; Karrer-Gauß, 2011; J. Schmidt, Braunagel, et al., 2016; Senaratne et al., 2007; Tsuchida et al., 2009; Wierwille & Ellsworth, 1994)
Level 4: Extremely fatigued		Duration	
	Very long eye closure (>= 2.0 seconds)	Onetime	(Karrer-Gauß, 2011; Wierwille & Ellsworth, 1994)
	Microsleep event (e.g., head nodding, sudden body movement)	Onetime	(Karrer-Gauß, 2011; Wierwille & Ellsworth, 1994)
	No body movement	Onetime	(Karrer-Gauß, 2011; Wierwille & Ellsworth, 1994)
Other Cues			
	Automation activated	Onetime	
	Monitoring Request	Onetime	
	End of automation (TOR2)	Onetime	
	Eye closure (Main category, Level 2)	Duration	

Generic Guideline and Documentary Material for Real-Time Assessment (Exemplary)

Fatigue level 1: non-fatigued		
Behavior indicator	**Occurrence of indicator/ frequency**	**Notes/characteristics (e.g., time of occurrence, prevalence, etc.)**
- Normal facial tone	☐	
- Normal fast eye blinks (<0.5 seconds)	☐	
- Short ordinary glances	☐	
- Occasional body movements or gestures	☐	
Time stamp of fatigue level change (after alignment with other rater):		
Miscellaneous:		

Fatigue level 2: fatigued		
Behavior indicator	**Occurrence of indicator/ frequency**	**Notes/characteristics (e.g., time of occurrence, prevalence, etc.)**
- Extended eye closure (0.5 – 1 seconds)	☐	
- Glazed look or rare saccades ("glassy-eyed appearance")	☐	
- Frequent eye blinks	☐	
- Low facial tone: "floppy" (slack) face	☐	
- Yawning	☐	
- Mannerisms	☐	
Time stamp of fatigue level change (after alignment with other rater):		
Miscellaneous:		

Fatigue level 3: very fatigued		
Behavior indicator	**Occurrence of indicator/ frequency**	**Notes/characteristics (e.g., time of occurrence, prevalence, etc.)**
- Long eye closure (1.0 – 2.0 seconds)	◻	
- Eye rolling	◻	
- Fixed gaze (no fixation, rare blinks) *and / or* - Cross-eyed look ("*lack of proper vergence*")	◻ ◻	
- Comfortable sitting position *and/ or* - Lack of activity	◻ ◻	
Time stamp of fatigue level change (after alignment with other rater):		
Miscellaneous:		

Fatigue level 4: extremely fatigued		
Behavior indicator	**Occurrence of indicator/ frequency**	**Notes/characteristics (e.g., time of occurrence, prevalence, etc.)**
- Very long eye closure (>= 2.0 seconds)	◻	
- Microsleep event (e.g., head nodding, sudden body movement)	◻	
- No body movement	◻	
Time stamp of fatigue level change (after alignment with other rater):		
Miscellaneous:		

Additional Statistics

Experiment 1

Table A.1. Test results on normal distribution and homogeneity of variance for mean fatigue level averaged over the entire drive of 60 minutes depending on the activity condition.

		Shapiro-Wilk	Levene
Activity condition	UC	W=0.917, p=0.099	F(1, 38)=3.371, p=0.074
	NLC	W=0.868, p=0.009	

Table A.2. Test results on normal distribution and homogeneity of variance for the point in time when reaching FL3 and FL4.

		Shapiro-Wilk	Levene
Activity condition	UC	W=0.938, p=0.536	F(1, 38)=2.448, p=0.140
	NLC	W=0.878, p=0.261	

Table A.3. Contingency tables for the count of participants who reached L3 or L4 during 60 minutes of CAD depending on the activity condition.

Activity condition		**Count of participants who reached L3 or L4**		
		UC	**NLC**	**Total**
UC	Count	9.0	10.0	19.0
	Expected count	11.205	7.795	19.0
NLC	Count	14.0	6.0	20.0
	Expected count	11.795	8.205	20.0
Total	Count	23.0	16.0	39.0
	Expected count	23.0	16.0	39.0

Table A.4. Test results on normal distribution and homogeneity of variance for mean duration in FL3 and FL4 depending on the activity condition.

		Shapiro-Wilk	Levene
Activity condition	UC	W=0.850, p=0.057	F(1, 38)=4.020, p=0.065
	NLC	W=0.774, p=0.034	

Table A.5. Contingency tables for the count of participants who stayed at FL1 during 60 minutes of CAD depending on the activity condition.

Activity condition		**Count of participants who stayed at FL1**		
		UC	**NLC**	**Total**
UC	Count	19.0	1.0	20.0
	Expected count	16.190	3.810	20.0
NLC	Count	15.0	7.0	22.0
	Expected count	17.810	4.190	22.0
Total	Count	34.0	8.0	42.0
	Expected count	34.0	8.0	42.0

Table A.6. Test results on normal distribution and homogeneity of variance for all metrical dependent variables of take-over performance depending on the activity condition.

Variable	Activity condition	Shapiro-Wilk	Levene
TOT	UC	W=0.957, p=0.489	F(1, 40)=0.276, p=0.602
	NLC	W=0.972, p=0.747	
TTC	UC	W=0.987, p=0.990	F(1, 40)=1.150, p=0.290
	NLC	W=0.972, p=0.784	
AccLong	UC	W=0.644, p<0.001	F(1, 40)=14.128, p<0.001
	NLC	W=0.762, p<0.001	
AccLat	UC	W=0.861, p=0.008	F(1, 40)=0.363, p=0.550
	NLC	W=0.933, p=0.139	

Table A.7. Contingency tables for the count of initial response types depending on the activity condition.

Activity condition		InRe			
		Steer	**Brake**	**Accelerate**	**Total**
UC	Count	13.0	3.0	4.0	20.0
	Expected count	12.381	5.238	2.381	20.0
NLC	Count	13.0	8.000	1.0	22.0
	Expected count	13.619	5.762	2.619	22.0
Total	Count	26.0	11.0	5.0	42.0
	Expected count	26.0	11.0	5.0	42.0

Table A.8. Contingency tables for the count of final response types depending on the activity condition.

Activity condition		FinRe		
		Lane change	**Full braking**	**Total**
UC	Count	20.0	0.0	20.0
	Expected count	18.095	1.905	20.0
NLC	Count	18.0	4.0	22.0
	Expected count	19.905	2.095	22.0
Total	Count	38.0	4.0	42.0
	Expected count	38.0	4.0	42.0

Table A.9. Contingency tables for the count of crashes depending on the activity condition.

Activity condition		Crash		
		No	**Yes**	**Total**
UC	Count	20.0	0.0	20.0
	Expected count	19.524	0.476	20.0
NLC	Count	21.0	1.0	22.0
	Expected count	21.476	0.524	22.0
Total	Count	41.0	1.0	42.0
	Expected count	41.0	1.0	42.0

Table A.10. Contingency tables for the count of mirror checks depending on the activity condition.

Activity condition		MC		
		No	Yes	Total
UC	Count	10.0	10.0	20.0
	Expected count	10.952	9.048	20.0
NLC	Count	13.0	9.0	22.0
	Expected count	12.048	9.952	22.0
Total	Count	23.0	19.0	42.0
	Expected count	23.0	19.0	42.0

Table A.11. Test results on normal distribution and homogeneity of variance for all metrical dependent variables of take-over performance.

Variable	Fatigue level	Shapiro-Wilk	Levene
TOT	FL1	W=0.904, p=0.353	F(3, 34)=2.162, p=0.111
	FL2	W=0.869, p=0.027	
	FL3	W=0.904, p=0.401	
	FL4	W=0.982, p=0.971	
TTC	FL1	W=0.880, p=0.227	F(3, 33)=0.529, p=0.666
	FL2	W=0.928, p=0.230	
	FL3	W=0.950, p=0.743	
	FL4	W=0.984, p=0.981	
AccLong	FL1	W=0.651, p=0.001	F(3, 34)=0.316, p=0.813
	FL2	W=0.729, p<0.001	
	FL3	W=0.774, p=0.034	
	FL4	W=0.755, p=0.009	
AccLat	FL1	W=0.827, p=0.075	F(3, 34)=1.656, p=0.195
	FL2	W=0.901, p=0.083	
	FL3	W=0.873, p=0.238	
	FL4	W=0.941, p=0.623	

Table A.12. Result of Kruskal-Wallis test on mean TOT depending on the fatigue level during the RtI.

Factor	Statistic	*df*	*p*
Fatigue level	H=8.015	3	0.046

Table A.13. Contingency tables for the count of final response types depending on the fatigue level.

Fatigue level		FinRe: Lane change	Full braking	Total
FL1	Count	7.0	1.0	8.0
	Expected count	7.4	0.6	8.0
FL2	Count	14.0	2.0	16.0
	Expected count	14.7	1.3	16.0
FL3	Count	6.0	0.0	6.0
	Expected count	5.5	0.5	6.0
FL4	Count	8.0	0.0	8.0
	Expected count	7.4	0.6	8.0
Total	Count	35.0	3.0	38.0
	Expected count	35.0	3.0	38.0

Table A.14. Contingency tables for the count of initial response types depending on the fatigue level.

Fatigue level		InRe: Steer	Brake	Accelerate	Total
FL1	Count	6.0	2.0	0.0	8.0
	Expected count	4.8	2.1	1.1	8.0
FL2	Count	9.0	4.0	3.0	16.0
	Expected count	9.7	4.2	2.1	16.0
FL3	Count	2	3.0	1.0	6.0
	Expected count	3.6	1.6	0.8	6.0
FL4	Count	6	1.0	1.0	8.0
	Expected count	4.8	2.1	1.1	8.0
Total	Count	23.0	10.0	5.0	38.0
	Expected count	23.0	10.0	5.0	38.0

Table A.15. Contingency tables for the count of crashes depending on the fatigue level.

Activity Condition		Crash: No	Yes	Total
FL1	Count	7.0	1.0	8.0
	Expected count	7.8	0.2	8.0
FL2	Count	16.0	0.0	16.0
	Expected count	15.6	0.4	16.0
FL3	Count	6.0	0.0	6.0
	Expected count	5.8	0.2	6.0
FL4	Count	8.0	0.0	8.0
	Expected count	7.8	0.2	8.0
Total	Count	37.0	1.0	38.0
	Expected count	37.0	1.0	38.0

Table A.16. Contingency tables for the count of mirror checks depending on the fatigue level.

Activity Condition		MC		
		No	Yes	Total
FL1	Count	2.0	6.0	8.0
	Expected count	4.2	3.8	8.0
FL2	Count	10.0	6.0	16.0
	Expected count	8.4	7.6	16.0
FL3	Count	2.0	4.0	6.0
	Expected count	3.2	2.8	6.0
FL4	Count	6.0	2.0	8.0
	Expected count	4.2	3.8	8.0
Total	Count	20.0	18.0	38.0
	Expected count	20.0	18.0	38.0

Table A.17. Test results on normal distribution and homogeneity of variance for mean PERCLOS averaged over the entire drive of 60 minutes depending on the activity condition.

		Shapiro-Wilk	Levene
Activity condition	UC	W=0.672, p<0.001	F(1, 39)=6.838, p=0.013
	NLC	W=0.722, p<0.001	

Experiment 2

Table A.18. Test results on normal distribution and homogeneity of variance for mean fatigue level averaged over the entire drive of 35 minutes depending on the activity condition.

		Shapiro-Wilk	Levene
Activity condition	UC	W=0.961, p=0.558	F(1, 37)=45.468, p<0.001
	NC	W=0.244, p<0.001	

Table A.19. Contingency tables for the count of participants who reached L3 or L4 during 35 minutes of CAD depending on the activity condition.

Activity condition		Count of participants who reached L3 or L4		
		UC	NC	Total
UC	Count	9.0	11.0	20.0
	Expected count	14.4	5.6	20.0
NC	Count	19.0	0.0	19.0
	Expected count	13.6	5.4	19.0
Total	Count	28.0	11.0	39.0
	Expected count	28.0	11.0	39.0

Table A.20. Contingency tables for the count of participants who stayed at FL1 during 35 minutes of CAD depending on the activity condition.

Activity condition		Count of participants who stayed at FL1		
		UC	**NC**	**Total**
UC	Count	18.0	2.0	20.0
	Expected count	9.7	10.3	20.0
NC	Count	1.0	18.0	19.0
	Expected count	9.3	9.7	19.0
Total	Count	19.0	20.0	39.0
	Expected count	19.0	20.0	39.0

Table A.21. Test results on normal distribution and homogeneity of variance for all metrical dependent variables of take-over performance after 35 minutes of CAD depending on the activity condition.

Variable	Activity condition	Shapiro-Wilk	Levene
TOT_CAD35	UC	W=0.983, p=0.963	F(1, 38)=2.334, p=0.135
	NC	W=0.869, p=0.011	
TTC_CAD35	UC	W=0.947, p=0.324	F(1, 38)=0.194, p=0.662
	NC	W=0.916, p=0.085	
AccLong_CAD35	UC	W=0.584, p<0.001	F(1, 38)=5.952, p=0.019
	NC	W=0.454, p<0.001	
AccLat_CAD35	UC	W=0.905, p=0.051	F(1, 38)=10.423, p=0.003
	NC	W=0.796, p<0.001	

Table A.22. Contingency tables for the count of initial response types after 35 of CAD depending on the activity condition.

Activity condition		InRe_CAD35			
		Steer	**Brake**	**Accelerate**	**Total**
UC	Count	13.0	5.0	2.0	20.0
	Expected count	12.5	4.0	3.5	20.0
NC	Count	12.0	3.0	5.0	20.0
	Expected count	12.5	4.0	3.5	20.0
Total	Count	25.0	8.0	7.0	40.0
	Expected count	25.0	8.0	7.0	40.0

Table A.23. Contingency tables for the count of final response types after 35 of CAD depending on the activity condition.

Activity condition		FinRe_CAD35		
		Lane change	**Full braking**	**Total**
UC	Count	18.0	2.0	20.0
	Expected count	18.5	1.5	20.0
NC	Count	19.0	1.0	20.0
	Expected count	18.5	1.5	20.0
Total	Count	37.0	3.0	40.0
	Expected count	37.0	3.0	40.0

Table A.24. Contingency tables for the count of mirror checks after 35 of CAD depending on the activity condition.

Activity condition		MC_CAD35 No	MC_CAD35 Yes	Total
UC	Count	4.0	16.0	20.0
	Expected count	3.0	17.0	20.0
NC	Count	3.0	18.0	20.0
	Expected count	6.0	17.0	20.0
Total	Count	6.0	34.0	40.0
	Expected count	6.0	34.0	40.0

Table A.25. Test results on normal distribution and homogeneity of variance for mean PERCLOS averaged over the entire drive of 35 minutes depending on the activity condition.

		Shapiro-Wilk	Levene
Activity condition	UC	W=0.558, p<0.001	F(1, 37)=8.818, p=0.005
	NC	W=0.691, p<0.001	

Table A.26. Test results on normal distribution and homogeneity of variance for all metrical dependent variables depending on the activity condition and the automation duration.

Variable	Activity condition	N	Shapiro-Wilk	Levene	Hartley's F_{max} (variance ratio)
TOT_3	UC	20	W=0.979, p=0.927	F(1, 38)=0.878, p=0.335	-
	NC	20	W=0.926, p=0.128		
TOT_35	UC	20	W=0.983, p=0.963	F(1, 38)=2.334, p=0.135	-
	NC	20	W=0.869, p=0.011		
TTC_3	UC	20	W=0.963, p=0.604	F(1, 38)=2.386, p=0.131	-
	NC	20	W=0.941, p=0.249		
TTC_35	UC	20	W=0.947, p=0.324	F(1, 38)=0.194, p=0.662	-
	NC	20	W=0.916, p=0.085		
AccLong_3	UC	20	W=0.449, p<0.001	F(1, 38)=2.453, p=0.126	-
	NC	20	W=0.382, p<0.001		
AccLong_35	UC	20	W=0.584, p<0.001	F(1, 38)=5.952, p=0.019	2.2
	NC	20	W=0.454, p<0.001		
AccLat_3	UC	20	W=0.775, p<0.001	F(1, 38)=0.276, p=0.603	-
	NC	20	W=0.846, p=0.005		
AccLat_35	UC	20	W=0.905, p=0.051	F(1, 38)=10.423, p=0.003	3.9
	NC	20	W=0.796, p=0.001		

Experiment 3

Table A.27. Test results on normal distribution and homogeneity of variance for all metrical dependent variables of take-over performance depending on the fatigue condition.

Variable	Fatigue condition	Shapiro-Wilk	Levene
TOT	FC	W=0.965, p=0.587	F(1, 45)=0.075, p=0.785
	NFC	W=0.973, p=0.728	
TTC	FC	W=0.927, p=0.121	F(1, 43)=1.973, p=0.167
	NFC	W=0.953, p=0.322	
AccLong	FC	W=0.746, p<0.001	F(1, 45)=7.916, p=0.007
	NFC	W=0.625, p<0.001	
AccLat	FC	W=0.958, p=0.455	F(1, 45)=0.139, p=0.711
	NFC	W=0.934, p=0.109	

Table A.28. Contingency tables for the count of initial response types depending on the fatigue condition.

Fatigue condition		**InRe**			
		Steer	**Brake**	**Accelerate**	**Total**
FC	Count	5.0	7.0	10.0	22.0
	Expected count	7.5	4.2	10.3	22.0
NFC	Count	11.0	2.0	12.0	25.0
	Expected count	8.5	4.8	11.7	25.0
Total	Count	16.0	9.0	22.0	47.0
	Expected count	16.0	9.0	22.0	47.0

Table A.29. Contingency tables for the count of final response types depending on the fatigue condition.

Fatigue condition		**FinRe**		
		Lane change	**Full braking**	**Total**
FC	Count	14.0	8.0	22.0
	Expected count	16.6	5.1	22.0
NFC	Count	22.0	3.0	25.0
	Expected count	19.1	5.9	25.0
Total	Count	36.0	11.0	47.0
	Expected count	36.0	11.0	47.0

Table A.30. Contingency tables for the count of crashes depending on the fatigue condition.

Fatigue condition		**Crash**		
		No	**Yes**	**Total**
FC	Count	21.0	1.0	22.0
	Expected count	21.1	0.9	22.0
NFC	Count	24.0	1.0	25.0
	Expected count	23.9	1.1	25.0
Total	Count	45.0	2.0	47.0
	Expected count	45.0	2.0	47.0

Table A.31. Contingency tables for the count of mirror checks depending on the fatigue condition.

Fatigue condition		**MC**		
		No	**Yes**	**Total**
FC	Count	9.0	13.0	22.0
	Expected count	6.6	15.4	22.0
NFC	Count	5.0	20.0	25.0
	Expected count	7.4	17.6	25.0
Total	Count	14.0	33.0	47.0
	Expected count	14.0	33.0	47.0

Experiment 4

Table A.32. Test results on normal distribution and homogeneity of variance for all metrical dependent variables of take-over performance depending on the fatigue condition.

Variable	Fatigue condition	Shapiro-Wilk	Levene
TOT	FC	W=0.967, p=0.686	F(1, 38)=1.363, p=0.250
	NFC	W=0.978, p=0.913	
TTC	FC	W=0.972, p=0.869	F(1, 32)=8.519, p=0.006
	NFC	W=0.962, p=0.643	
AccLong	FC	W=0.682, p<0.001	F(1, 38)=3.069, p=0.088
	NFC	W=0.786, p<0.001	
AccLat	FC	W=0.937, p=0.213	F(1, 38)=1.478, p=0.232
	NFC	W=0.986, p=0.987	

Table A.33. Contingency tables for the count of initial response types depending on the fatigue condition.

Fatigue condition		**InRe**			
		Steer	**Brake**	**Accelerate**	**Total**
FC	Count	2.0	7.0	11.0	20.0
	Expected count	5.5	6.0	8.5	20.0
NFC	Count	9.0	5.0	6.0	20.0
	Expected count	5.5	6.0	8.5	20.0
Total	Count	11.0	12.0	17.0	40.0
	Expected count	11.0	12.0	17.0	40.0

Table A.34. Contingency tables for the count of final response types depending on the fatigue condition.

Fatigue condition		**FinRe**		
		Lane change	**Full braking**	**Total**
FC	Count	18.0	2.0	20.0
	Expected count	16.5	3.5	20.0
NFC	Count	15.0	5.0	20.0
	Expected count	16.5	3.5	20.0
Total	Count	33.0	7.0	40.0
	Expected count	33.0	7.0	40.0

Table A.35. Contingency tables for the count of crashes depending on the fatigue condition.

Fatigue condition		**Crash**		
		No	**Yes**	**Total**
FC	Count	16.0	4.0	20.0
	Expected count	17.0	3.0	20.0
NFC	Count	18.0	2.0	20.0
	Expected count	17.0	3.0	20.0
Total	Count	34.0	6.0	40.0
	Expected count	34.0	6.0	40.0

Table A.36. Contingency tables for the count of mirror checks depending on the fatigue condition.

Fatigue condition		**MC**		
		No	**Yes**	**Total**
FC	Count	10.0	10.0	20.0
	Expected count	9.0	11.0	20.0
NFC	Count	8.0	12.0	20.0
	Expected count	9.0	11.0	20.0
Total	Count	18.0	22.0	40.0
	Expected count	18.0	22.0	40.0

Table A.37. Test results on normal distribution and homogeneity of variance for all metrical dependent variables of take-over performance of experiment 3 and experiment 4 depending on the fatigue condition.

Variable	Fatigue condition / time budget	Shapiro-Wilk	Levene	Hartley's F_{max} (variance ratio)
TOT	FC	W=0.983, p=0.845	F(3, 83)=1.194, p=0.317	-
	NFC	W=0.977, p=0.552		
	5s	W=0.983, p=0.855		
	6s	W=0.972, p=0.355		
TTC	FC	W=0.963, p=0.250	F(3, 75)=3.138, p=0.030	4.0
	NFC	W=0.979, p=0.640		
	5s	W=0.975, p=0.603		
	6s	W=0.949, p=0.048		
AccLong	FC	W=0.741, p<0.001	F(3, 83)=3.627, p=0.016	1.5
	NFC	W=0.707, p<0.001		
	5s	W=0.762, p<0.001		
	6s	W=0.697, p<0.001		
AccLat	FC	W=0.965, p=0.294	F(3, 83)=0.793, p=0.501	-
	NFC	W=0.964, p=0.212		
	5s	W=0.956, p=0.182		
	6s	W=0.971, p=0.302		

Table A.38. Result of Kruskal-Wallis test on mean TTC depending on the fatigue condition and time budget.

Factor	***H*-Statistic**	***df***	***p***
Fatigue condition	1.126	1	0.289
Time budget	0.047	1	0.828

Table A.39. Result of Kruskal-Wallis test on mean AccLong depending on the fatigue condition and time budget.

Factor	***H*-Statistic**	***df***	***p***
Fatigue condition	6.074	1	0.014
Time budget	3.727	1	0.054